Mack C. Shelley II
Larry D. Yore
Brian Hand
Editors

Quality Research in Literacy and Science Education

International Perspectives and Gold Standards

Springer

Quality Research in Literacy and Science Education

Mack C. Shelley II
Larry D. Yore
Brian Hand
Editors

Quality Research in Literacy and Science Education

International Perspectives and Gold Standards

Springer

Editors
Dr. Mack C. Shelley II
Iowa State University
USA
mshelley@iastate.edu

Dr. Larry D. Yore
University of Victoria
Canada
lyore@uvic.ca

Dr. Brian Hand
University of Iowa
USA
brian-hand@ uiowa.edu

ISBN 978-1-4020-8426-3 e-ISBN 978-1-4020-8427-0

Library of Congress Control Number: 2008937476

Printed on acid-free paper

springer.com

Contents

Part III Curriculum and Pedagogy

Part IV Statistics, Research Methods, and Science Literacy

Part V Public Policy and "Gold Standard(s)" Research

Part VI Epilogue: New Standards, New Directions, and New Realities

About the Coeditors

Mack C. Shelley II
Email: mshelley@iastate.edu

Mack Shelley is University Professor in the Department of Statistics and the Department of Political Science (and Director of the Public Policy and Administration program) at Iowa State University, Ames, IA, USA. From 2003 to 2007 he served as Director of the Research Institute for Studies in Education (and was Coordinator of Research from 1999–2003), and from 1999 to 2007 he was a Professor in the Department of Educational Leadership and Policy Studies. He was coeditor of the *Policy Studies Journal* (1993–2002) and a member of the Editorial Advisory Board for *TESOL Quarterly* (2003–2005), and currently is Associate Editor of the *Journal of Information Technology & Politics*. His research and teaching focus on statistical methods and their applications to public policy and program evaluation. His funding sources include the National Science Foundation, US Department of Education, Center for Substance Abuse Prevention, US Department of Health and Human Services, Pew Foundation, American Judicature Society, Iowa Department of Education, Iowa Department of Public Health, Des Moines Independent Community School District, Iowa Department of Public Health, and Iowa Board of Regents. His publications include 10 books, 19 book chapters, 85 journal articles and refereed proceedings papers, and over 200 other publications. He serves regularly as a statistical consultant.

Larry D. Yore
Email: lyore@uvic.ca

Larry Yore is a University Distinguished Professor in the Faculty of Education at the University of Victoria, Victoria, British Columbia, Canada. He has teaching experience at both secondary and tertiary levels, serving as a junior-senior secondary school science teacher, K-12 science coordinator, secondary science department head, and instructor and supervisor of student teachers. In nearly four decades in teaching and research, he has been engaged in developing provincial science curricula, national science frameworks, and national K-12 assessment projects. In addition, he has been involved in many administrative positions and is currently codirecting the Pacific CRYSTAL Centre for Science and Technology Literacy in

Western Canada. He has served on or is currently a member of the editorial boards of the *Journal of Research in Science Teaching, School Science and Mathematics, Science Education, Journal of Science Teacher Education, Journal of Elementary Science Education, International Journal of Science Education, L1—Educational Studies in Language and Literacy, Science and Technology Education*, and *International Journal of Science and Mathematics Education*. His recent research focuses on the role of language uses in science and science education and how language arts affect science inquiry. He has published numerous journal articles; coauthored elementary science textbooks; edited special issues related to applications of language arts in science education; consulted on various research, curriculum policy, and professional development projects provincially, nationally, and internationally; and presented numerous lectures, workshops, etc.

Brian Hand
Email: brian-hand@uiowa.edu

Brian Hand is a Professor of Science Education at the University of Iowa, Iowa City, IA, USA. Prior to moving to the University of Iowa, he was the Director of the Research Center for Excellence in Science and Mathematics Education at Iowa State University. His research focuses on two areas. The first area focuses on language as a learning tool to improve students' understanding of science and the use of multimodal representation within science classrooms. This research extends the use of writing as a learning tool to include different modes of representation. The second area of research is the development of scientific argument through the use of the Science Writing Heuristic. This research is aimed at helping students learn to use science argument to construct science knowledge. He has received external funding from the National Science Foundation and the Iowa Department of Education. He has served on or is currently a member of the editorial boards of the *Journal for Research in Science Teaching, International Journal of Science Education, Research in Science Education, Science Education*, and *Elementary Science Education Journal*. He has published 2 books, with 2 in press, 16 book chapters, and 60 referred journal articles. He was a secondary school chemistry/physics teacher for 11 years and has extensive experience working with educators from K–13 in professional development settings.

Acknowledgments and Disclaimer

The editors wish to thank the National Science Foundation (Grant number 0437198), Iowa State University, the University of Iowa, and the University of Victoria for the support and related opportunities provided. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, Iowa State University, the University of Iowa, and the University of Victoria.

The editors want to thank the 2nd Island Conference planning team (Donna E. Alvermann, University of Georgia; Stephen P. Norris, University of Alberta; Linda M. Phillips, University of Alberta; and E. Wendy Saul, University of Missouri, St. Louis), the conference participants, and the authors and reviewers for the chapters in this book, especially the contributions made by Victor Battistich, University of Missouri, St. Louis (deceased), and Christine D. Tippett, University of Victoria. Finally, the editors and authors want to recognize and thank Sharyl A. Yore for her administrative support, manuscript management, and technical editing. Her guidance and encouragement were instrumental in the completion of this book.

About the Authors

Recai Akkus
College of Education
Abant Izzet Baysal University
Email: akkus_r@ibu.edu.tr
Recai Akkus is an Assistant Professor of Mathematics Education at Abant Izzet Baysal University, Bolu, Turkey. His current research focuses on the issues of problem solving, students' reasoning during problem solving, and writing-to-learn in mathematics.

Donna E. Alvermann
Department of Language & Literacy Education
University of Georgia
Email: dalverma@uga.edu
Donna Alvermann is a Distinguished Research Professor at the University of Georgia, Athens, GA, USA. A former classroom teacher in Texas and New York, she codirected the National Reading Research Center from 1992–1997 and is immediate past editor of *Reading Research* Q*uarterly*. Her current research focuses on adolescents' online literate practices.

John O. Anderson
Department of Educational Psychology
University of Victoria
Email: anderson@uvic.ca
John Anderson is a Professor of Educational Psychology at the University of Victoria, British Columbia, Canada. He was Director of Research for the Educational Research Institute of British Columbia before coming to the University of Victoria to teach and research in the area of educational measurement. His current research focuses on assessment issues and secondary analysis of large evaluation datasets.

Robert J. Anthony
Department of Curriculum & Instruction
University of Victoria
Email: ranthony@uvic.ca
Robert Anthony is an Associate Professor and currently Chair of the Department of Curriculum & Instruction at the University of Victoria, British Columbia, Canada. He is a Co-Principal Investigator for the Natural Sciences and Engineering Research Council of Canada Pacific Centre for Research in Youth, Science Teaching and Learning (CRYSTAL) Project. His current research focuses on the enrichment of science learning through the explicit teaching of science literacy.

Huub van den Bergh
Utrecht Institute of Linguistics & Graduate School of Teaching & Learning
Utrecht University & University of Amsterdam
Email: Huub.vandenBergh@let.uu.nl
Huub van den Bergh is a Professor at Utrecht University and the University of Amsterdam, The Netherlands. He has been involved in many nationwide studies on educational effectiveness as well as in many small-scale studies on writing and reading processes.

Pietro Boscolo
Department of Educational Psychology
University of Padova
Email: pietro.boscolo@unipd.it
Pietro Boscolo is a Professor and Head of the Department of Developmental & Socialization Psychology at the University of Padova, Italy. He is President of the European Association for Research on Learning & Instruction. His current research focuses on literacy and the relationship between cognition and motivation in learning science, history, and literature.

Martine A. H. Braaksma
Graduate School of Teaching and Learning
University of Amsterdam
Email: Braaksma@uva.nl
Martine Braaksma is a Postdoctoral Fellow at the University of Amsterdam, The Netherlands. She did her Ph.D. on observational learning in argumentative writing. Her current research focuses on the issues of hypertext writing in an educational context; this research project is funded by the Netherlands Organisation for Scientific Research (NWO).

Andy Cavagnetto
School of Education
State University of New York
Email: acavagne@binghamton.edu
Andy Cavagnetto is an Assistant Professor of Science Education at State University of New York, Binghamton, NY, USA. His current research focuses on the issues of

science literacy, specifically student dialogue during inquiry investigations and student interpretation of science in popular media sources.

Wen-Hua Chang
Department of Life Science
National Taiwan Normal University
Email: sujudy@ntnu.edu.tw
Wen-Hua Chang is an Associate Professor of Science Education at National Taiwan Normal University, Taipei. Her current research focuses on the issues of science curriculum studies, developing nature-of-science-explicit science curricular materials, and professional development.

Mei-Hung Chiu
Graduate Institute of Science Education
National Taiwan Normal University
Email: mhchiu@ntnu.edu.tw
Mei-Hung Chiu is a Professor of Science Education at National Taiwan Normal University, Taipei. She is Chair, International Committee of the National Association for Research in Science Teaching; Chair of the Subcommittee on Chemistry Education for Development, and Member of the Committee on Chemistry Education, IUPAC; and Project Director of the Asian Chemical Education Network of the Federation of Asian Chemical Societies. Her current research focuses on the issues of mental models, conceptual change, and modeling ability in learning and instruction of science.

Richard K. Coll
Centre for Science & Technology Education Research
University of Waikato
Email: rcoll@waikato.ac.nz
Richard Coll is an Associate Professor of Science Education at the University of Waikato, Hamilton, New Zealand, and Deputy Dean in the School of Science & Engineering. His current research focuses on the issues of scientific literacy and the use of models including analogies in the teaching and learning of science.

Dianne Cook
Department of Statistics
Iowa State University
Email: dicook@iastate.edu
Dianne Cook is a Professor of Statistics at Iowa State University, Ames, IA, USA. She is a coauthor of the book *Interactive and Dynamic Graphics for Data Analysis*, codeveloper of the open source software GGobi, principal investigator on a National Science Foundation grant for developing interactive graphics methods, and a past chair of the graphics section of the American Statistical Association. Her current research focuses on dynamic graphics, exploratory data analysis, multivariate statistics, and statistical computing.

Ida J. Cook
Department of Sociology
University of Central Florida
Email: cook@mail.ucf.edu
Ida Cook is an Associate Professor of Sociology at the University of Central Florida, Orlando, FL, USA. She served as Interim Director of the Karen Smith Faculty Center for Teaching and Learning and as Faculty Senate Chair during the research project. Her current research focuses on the issues of retirement expectations and experiences among university professionals, and evaluation of social programs such as substance abuse, domestic violence, and political behavior.

Justin Dillon
Science & Technology Education Group
King’s College London
Email: justin.dillon@kcl.ac.uk
Justin Dillon is a Senior Lecturer in Science & Environmental Education at King’s College London, UK. He is head of the Science & Technology Education Group, President of the European Science Education Research Association, and coeditor of the *International Journal of Science Education*. His current research focuses on environmental education and learning science in informal contexts.

David J. Dude
Iowa Testing Programs
University of Iowa
Email: david-dude@uiowa.edu
David Dude is an Application Developer for Iowa Testing Programs at the University of Iowa, Iowa City, IA, USA. Before joining Iowa Testing Programs, he was a secondary school mathematics teacher for 11 years. His current research focuses on educational accountability, particularly in the area of school administration.

Charles D. Dziuban
Research Initiative for Teaching Effectiveness
University of Central Florida
Email: dziuban@mail.ucf.edu
Charles Dziuban is Professor Emeritus in the College of Education and Director of the Research Initiative for Teaching Effectiveness at the University of Central Florida, Orlando, FL, USA. His research focuses on the impact evaluation of UCF’s distributed learning initiative examining student and faculty outcomes as well as gauging the impact of online courses on the university. He received the 2005 Sloan Consortium award for Most Outstanding Achievement in Online Learning by an Individual. In 2007, he was appointed to the National Information and Communication Technology (ICT) Literacy Policy Council.

Sibel Erduran
Science Education
University of Bristol
Email: Sibel.Erduran@bristol.ac.uk
Sibel Erduran is a Reader in Science Education at the University of Bristol, UK. She is the coeditor for the Science Studies Section of *Science Education* and serves on the editorial boards of several other academic journals. She has received research and development funding from the Fulbright Program; European Union; Economic and Social Research Council; Spencer, Gatsby, and Nuffield Foundations; and Training and Development Agency for Schools. Her research interests include the cognitive and epistemological aspects of science in science education with a particular emphasis on argumentation and chemistry learning.

Cynthia Fakudze
Schools Development Unit
University of Cape Town
Email: Cynthia.Fakudze@uct.ac.za
Cynthia Fakudze is a Science Education Specialist at the University of Cape Town, South Africa. She is the Project Manager of the Department of Testing, Evaluation, Assessment, & Measurement. She has led teams in large-scale projects on the performance of learners in numeracy and literacy skills. Her current research focuses on the sociocultural factors on the teaching and learning of science.

Amy G. Froelich
Department of Statistics
Iowa State University
Email: amyf@iastate.edu
Amy Froelich is an Associate Professor of Statistics at Iowa State University, Ames, IA, USA. She has conducted workshops on multidimensional item response theory at the National Council of Measurement in Education annual meetings and presented an invited paper on the Measurement of Food Insecurity and Hunger at the National Academies of Sciences. She was recognized for her teaching at ISU, receiving the 2004 ISU Foundation Award for Early Achievement in Teaching. Her current research focuses on statistics education and the application of psychometric modeling to measurement problems.

Wolfgang Gräber
Leibniz-Institute for Science Education
University of Kiel
Email: wgraeber@ipn.uni-kiel.de
Wolfgang Gräber is a Senior Researcher at the Leibniz-Institute for Science Education (IPN) at the University of Kiel, Germany. He taught in secondary schools prior to university teaching and chemistry education research at the University of Essen and, since 1987, at IPN and the University of Kiel. His current research focuses on bridging the two cultures of science and general education,

aiming for scientifically literate citizens, and promoting self-directed learning (with new technologies) as a prerequisite for lifelong learning.

Irene Grimberg
Science & Mathematics Resource Center
Montana State University
Email: grimberg@montana.edu
Irene Grimberg is an Associate Research Professor of the Science & Mathematics Resource Center at Montana State University, Bozeman, MT, USA. She was involved in numerous programs related to science teaching and learning in rural settings throughout the USA. Her current research focuses on science professional development of teachers on Native American reservations.

Murat Gunel
Science Education
Ataturk University
Email: mgunel@atauni.edu.tr
Murat Gunel is an Assistant Professor of Science Education at Ataturk University, Erzurum, Turkey. His current research focuses on the issues of multimodel representations in science learning, implementation of inquiry teaching in physics, and professional development for science teachers.

Denyse V. Hayward
Canadian Centre for Research on Literacy
University of Alberta
Email: dhayward@ualberta.ca
Denyse Hayward is a Research Associate at the Canadian Centre for Research on Literacy at the University of Alberta, Edmonton, AB, Canada, and a Researcher with Capital Health in the area of speech language pathology. Her current research focuses on the use of assessment of language and literacy abilities, linking assessment to intervention for children with language learning difficulties, and dynamic assessment.

Tanja Janssen
Graduate School of Teaching & Learning
University of Amsterdam
Email: T.M.Janssen@uva.nl
Tanja Janssen is a Senior Researcher at the University of Amsterdam, The Netherlands. Her current research focuses on literature reading and learning in secondary education.

Alister Jones
School of Education
University of Waikato
Email: ajones@waikato.ac.nz
Alister Jones is a Professor and currently Dean of the School of Education at the University of Waikato, Hamilton, New Zealand. He is the former Director of the

Wilf Malcolm Institute of Educational Research. He has been extensively involved in science and technology education research since 1980 and has been director of a number of science and technology education contracts, including policy adviser to the Ministry of Education, science and technology curriculum development, and research into student and teacher learning and assessment.

Rosária Justi
Chemistry Department and Faculty of Education
Universidade Federal de Minas Gerais
Email: rjusti@ufmg.br
Rosária Justi is an Associate Professor of Chemistry Education at Universidade Federal de Minas Gerais, Brazil. She is an editor of the *International Journal of Science Education* and member of the editorial board of several other science education journals. Her current research focuses on the issues of modeling-based teaching, analogies in science education, and science teachers' development.

Johannes Klumpers
Directorate-General Research
European Commission
Email: Johannes.Klumpers@ec.europa.eu
Johannes Klumpers is Head of Unit, Scientific Culture & Gender Issues, at the European Commission, Brussels, Belgium. He was a researcher with Swedish forest-based industries. He joined the EC in 1998 where he first worked as a project officer for research in renewable raw materials. The unit he is currently responsible for focuses on formal and informal science education, gender equality in science, scientific culture in general, and research funding to the Humanities.

Tamar Levin
School of Education
Tel Aviv University
Email: tami1@post.tau.ac.il
Tamar Levin is a Professor of Education at Tel Aviv University, Israel. She has served in numerous research, evaluation, and curriculum capacities within the university and with the Ministry of Education and private corporations. Her research and expertise are in instructional science, evaluation and research methodology, teachers' and students' beliefs, and a variety of topics related to characteristics of school achievements with emphasis on science and mathematics among immigrants, and the use of communication technologies in the schools.

John S. Macnab
Jasper Place Secondary School
Edmonton Public School Board
Email: jmacnab@amedia.ca
John Macnab is a Senior Secondary School Mathematics Teacher at Jasper Place School in Edmonton and a Project Associate with both the Canadian Centre for Research on Literacy (CCRL) and the Centre for Research in Youth, Science Teaching and Learning (CRYSTAL)—Alberta, Edmonton, AB, Canada.

He was pursuing a doctorate in the Department of Educational Policy Studies, University of Alberta, when he worked on this chapter. His current research focuses on the philosophy of educational testing and the philosophy of mathematics education.

Christine A. Mallozzi
Department of Language & Literacy Education
University of Georgia
Email: mallozzi@uga.edu
Christine Mallozzi is a doctoral candidate of reading education at the University of Georgia, Athens, GA, USA. She was a classroom teacher in Ohio and was recently awarded the UG Elmer Jackson Carson Memorial Scholarship and the 2007 Outstanding Student Research Paper from the Georgia Educational Research Association. Her current research focuses on the issues of personal and professional lives of teachers.

Elizabeth H. McEneaney
Department of Sociology
California State University, Long Beach
Email: emcenean@csulb.edu
Elizabeth McEneaney is an Associate Professor of Sociology at California State University, Long Beach, CA, USA. She is a former secondary school mathematics and science teacher, and has taught research methods at undergraduate and graduate levels in both sociology and education. Her current research focuses on the effectiveness of charter school reforms in the USA and family numeracy practices.

Michelle A. Mengeling
Iowa Testing Programs
University of Iowa
Email: michelle-mengeling@uiowa.edu
Michelle Mengeling is an Associate Research Scientist for Iowa Testing Programs at the University of Iowa, Iowa City, IA, USA. She is an author of the Iowa Tests technical publications including the *Guide to Research and Development*. Her current research focuses on the use of longitudinal data for data-driven decisions.

Mary C. Meyer
Department of Statistics
Colorado State University
Email: meyer@stat.colostate.edu
Mary Meyer is an Associate Professor of Statistics at Colorado State University, Fort Collins, CO, USA. Her current research focuses on inference methods using nonparametric function estimation with shape restrictions.

Todd Milford
Department of Educational Psychology
University of Victoria
Email: tmilford@uvic.ca
Todd Milford is a doctoral candidate in educational measurement at the University of Victoria, British Columbia, Canada. He has taught secondary science at both schools and online schools. His current research focuses on the analysis and modeling of achievement data from large-scale assessment programs.

Robin Millar
Department of Educational Studies
University of York
Email: rhm1@york.ac.uk
Robin Millar is Salters' Professor of Science Education at the University of York, UK. He was Coordinator of the Evidence-based Practice in Science Education Research Network, which was part of the UK Economic & Social Research Council Teaching and Learning Research Programme. He has directed a number of major curriculum development projects in science, including AS-level Science for Public Understanding, Twenty First Century Science, and A-level Science in Society. He was a member of the Science Expert Group for the OECD PISA 2006 survey. His current research focuses on issues concerning student learning in the sciences and the implications of the goal of scientific literacy for curriculum design and student assessment.

Shereeza F. Mohammed
Department of Instructional Technology & Research
Florida Atlantic University
Email: shereezam@yahoo.com
Shereeza Mohammed is an Instructor with Florida Atlantic University, Boca Raton, FL, USA. She is on the editorial board of the *Education Policy Analysis Archives* and reviews for various organizations, such as the National Society for the Study of Education and the American Educational Research Association. Her current research focuses on program implementation at the state level, decision making at the district level, and scale-up issues for programs being implemented in schools.

Eduardo Mortimer
Faculty of Education
Universidade Federal de Minas Gerais
Email: mortimer@netuno.lcc.ufmg.br
Eduardo Mortimer is an Associate Professor of Chemistry Education at Universidade Federal de Minas Gerais, Brazil. He is the President of the Brazilian Science Education Research Association and editorial board member of several science education journals. His current research focuses on the issues of students' construction of scientific knowledge and the use of language in science classes.

Patsy D. Moskal
Research Initiative for Teaching Effectiveness
University of Central Florida
Email: pdmoskal@mail.ucf.edu
Patsy Moskal is a Faculty Research Associate and the Associate Director for the Research Initiative for Teaching Effectiveness at the University of Central Florida, Orlando, FL, USA. Since 1996, she has served as the liaison for faculty research of distributed learning and teaching effectiveness. She has coauthored a number of book chapters and journal articles on research in online and blended courses.

Daniel J. Mundfrom
Department of Applied Statistics & Research Methods
University of Northern Colorado
Email: Daniel.Mundfrom@unco.edu
Daniel Mundfrom is a Professor of Applied Statistics & Research Methods at the University of Northern Colorado, Greeley, CO, USA. He was the 2003 recipient of the M. Lucile Harrison Award for professional excellence in teaching, scholarship, and service (UNC's highest faculty honor) and has received three awards for teaching excellence from two different universities. His current research focuses on issues of statistical methodology and their properties.

Martina Nieswandt
Department of Mathematics & Science Education
Illinois Institute of Technology
Email: mnieswan@iit.edu
Martina Nieswandt is an Associate Professor of Science Education at the Illinois Institute of Technology, Chicago, IL, USA. She has received various national and international research grants for her classroom-based research utilizing mixed methods. Her current research focuses on the relationships among motivation, affect, and cognition of learning science in secondary school as well as science teachers' beliefs about science and science teaching.

Deborah Nolan
Department of Statistics
University of California, Berkeley
Email: nolan@stat.Berkeley.edu
Deborah Nolan is a Professor of Statistics at the University of California, Berkeley, CA, USA. She is a Fellow of the Institute of Mathematical Statistics; coauthor and editor of four books on teaching statistics and mathematics through case studies, activities, and projects; and the recipient of several awards for teaching and mentoring. Her current research interests are in the area of cyber-infrastructure for statistics education.

Stephen P. Norris
Department of Educational Policy Studies
University of Alberta
Email: stephen.norris@ualberta.ca
Stephen Norris is a Professor of Educational Policy, holds the Canada Research Chair in Scientific Literacy and the Public Understanding of Science, and is the Director of the Centre for Research in Youth, Science Teaching and Learning at the University of Alberta (CRYSTAL—Alberta), Edmonton, AB, Canada. In addition to his research on scientific literacy, he has published on the philosophy of educational research and the philosophy of reading.

Lori A. Norton-Meier
Department of Curriculum & Instruction
Iowa State University
Email: nortonme@iastate.edu
Lori Norton-Meier is an Assistant Professor of Curriculum & Instruction at Iowa State University, Ames, IA, USA. She is codirector of several University of Iowa, Iowa State University, and National Science Foundation-sponsored projects and coeditor of a book on the Science Writing Heuristic. Her current research focuses on the intersection of science and literacy practices for children in grades K-6.

Jonathan Osborne
School of Education
Stanford University
Email: jonathan.osborne@kcl.ac.uk
Jonathan Osborne holds an Endowed Chair of Science Education at Stanford University, Palo Alto, CA, USA. Prior to this he held the Chair of Science Education and was the Head of the Department of Education & Professional Studies at King’s College London, UK. Previously he taught physics in Inner London for 9 years. He has an extensive record of publications and research grants in science education in the field of primary science, science education policy, the teaching of the history of science, argumentation, and informal science education. His current research focuses on developing teachers’ practice with the use of argumentation in schools and on students’ attitudes to school science.

Stephen Parker
Directorate-General Research
European Commission
Email: stephen.parker@ec.europa.eu
Stephen Parker is Deputy Head of Unit, Scientific Culture & Gender Issues, at the European Commission, Brussels, Belgium. Prior to this, he was Head of Sector for Science Education and before that Head of Sector for Young People and Science, all within Directorate-General Research. He has been very closely involved with the evolution of policy in this area since the beginning of the Fifth Framework Programme.

Linda M. Phillips
Canadian Centre for Research on Literacy
University of Alberta
Email: linda.phillips@ualberta.ca
Linda M. Phillips is a Professor of Reading and Director of the Canadian Centre for Research on Literacy (CCRL) at the University of Alberta, Edmonton, AB, Canada. Her research focuses on language and literacy assessment, family literacy development, scientific literacy, and the use of magnetic resonance imaging to study the underlying causes of reading difficulties.

Vaughan Prain
Faculty of Education
La Trobe University
Email: v.prain@latrobe.edu.au
Vaughan Prain is a Professor of Education at La Trobe University, Bendigo, Australia. He was the literacy consultant on Primary Connections, the national professional learning program that linked learning science and literacy in elementary schools. His current research focuses on the use of multimodal representations in learning science in schools.

Gert C. W. Rijlaarsdam
Graduate School of Teaching & Learning
University of Amsterdam
Email: G.C.W.Rijlaarsdam@uva.nl
Gert Rijlaarsdam is a Professor of Education at the University of Amsterdam, The Netherlands. He is series editor of the international book series *Studies in Writing* and the journals *L1-Educational Studies in Language and Literature* and *Journal of Writing Research*. His current research focuses on the issues of writing processes and processes in reading literary texts, and effective interventions in learning to write and writing to learn.

Marissa Rollnick
Marang Centre for Science & Mathematics Education
University of the Witwatersand
Email: marissa.rollnick@wits.ac.za
Marissa Rollnick is a Professor and Chair of Science Education at the University of the Witwatersand, Johannesburg, South Africa. She taught at teachers' colleges and universities in Swaziland for 15 years prior to returning to South Africa in 1990. She has 16 years of doctoral and masters supervision and has published over 35 refereed articles. Her research interests have covered the areas of language in science and learning of chemistry at the foundation level, but she is now engaged in research into subject matter for teaching or pedagogical content knowledge.

Shelley P. Ross
Department of Family Medicine
University of Alberta
Email: shelleyross@med.ualberta.ca
Shelley Ross is an Assistant Professor in the Department of Family Medicine at the University of Alberta, Edmonton, AB, Canada, where she works as a Medical Education researcher. She was pursuing a doctorate in the Department of Educational Psychology, University of Victoria, British Columbia, when she worked on this chapter. Her current research examines procedures, outcomes, and policy recommendations related to how physicians are trained and assessed.

Gretchen B. Rossman
Center for International Education
University of Massachusetts
Email: Gretchen@educ.umass.edu
Gretchen B. Rossman is a Professor of International Education at the University of Massachusetts, Amherst, MA, USA. She served as Co-Principal Investigator on the University of Massachusetts–University of Malawi Linkages Project and is currently serving in that role for the Multigrade Demonstration Schools Project in Senegal and The Gambia. Her current research focuses on the issues of ethics in qualitative research practice.

E. Wendy Saul
Division of Teaching & Learning
University of Missouri, St. Louis
Email: saulw@umsl.edu
Wendy Saul serves as the Allan B. and Helen S. Shopmaker Professor of Education and International Studies at the University of Missouri-St. Louis, MO, USA. She has written extensively about science and literacy connections for both the practitioner and research communities and served as principal investigator for six National Science Foundation grants. Her current research focuses on promoting the science literacy of secondary school students through science journalism.

Hsiao-Ching She
Institute of Education
National Chiao Tung University
Email: hcshe@mail.nctu.edu.tw
Hsiao-Ching She is a Professor in the Institute of Education at National Chiao Tung University, Hsinchu, Taiwan. She has received the outstanding research award from the Taiwan National Science Council and the citation classic award from the Information Science Institute. She is serving as academic field coordinator for the NSC’s Science Education Department, Division of Science Education II: Science Teaching. Her current research focuses on the issues of conceptual change, scientific reasoning, web-based argumentation, multiple representations and science learning, and eye movement and scientific concept construction.

Robert D. Sherwood
Department of Curriculum & Instruction
Indiana University
Email: rdsherwo@indiana.edu
Robert Sherwood is a Professor of Science Education and Associate Dean for Research at the School of Education, Indiana University, Bloomington, IN, USA. He held faculty and administrative positions at New York University and Vanderbilt University before coming to Indiana in 2006. From 2004 to 2006, he was a Program Director in the Division of Elementary, Secondary, and Informal Education at the National Science Foundation. His current research focuses on the issues of situated cognition in science learning and educational policy issues related to science education.

Kim Chwee Daniel Tan
Natural Sciences & Science Education
Singapore Institute of Education
Email: daniel.tan@nie.edu.sg
Daniel Tan is an Associate Professor of Science Education at the Singapore National Institute of Education. He taught chemistry in schools for 8 years before completing a Ph.D. in science education from Curtin University of Technology. His current research focuses on the issues of the development of diagnostic instruments to determine students' understanding of science conceptions, multimodality in chemistry, science curriculum, practical work, and information communication technologies in science education.

Duncan Temple Lang
Department of Statistics
University of California, Davis
Email: duncan@wald.ucdavis.edu
Duncan Temple Lang is an Associate Professor of Statistics at the University of California, Davis, CA, USA. Before joining UC Davis in 2004, he was a researcher for 7 years in the Statistics & Data Mining Group at Bell Labs, Lucent Technologies. His current research focuses on developing new languages and environments for statistical computing, integrating information technologies and computer science research into the process of scientific and statistical research in general, automated ways.

Marion Tillema
Utrecht Institute of Linguistics
Utrecht University
Email: kortman-tillema@home.nl
Marion Tillema is a Ph.D. student of linguistics at Utrecht University, The Netherlands. Her current research focuses on cognitive processing during L1 Dutch and L2 English writing.

David F. Treagust
Science & Mathematics Education Centre
Curtin University of Technology
Email: d.f.treagust@curtin.edu.au
David Treagust is a Professor of Science Education at Curtin University of Technology, Perth, Australia. He holds a Ph.D. from the University of Iowa and is a past president of the National Association for Research in Science Teaching. His current research focuses on identification, design, and implementation of intervention strategies to challenge students' conceptions, the design of tests and other assessment instruments to diagnose student understanding of content in specific science areas, and students' use of analogies and models as an aid to their understanding of science concepts.

Russell Tytler
Faculty of Arts & Education
Deakin University
Email: tytler@deakin.edu.au
Russell Tytler is a Professor of Science Education at Deakin University, Geelong, Victoria, Australia. He has been involved as principal researcher in a number of large, funded projects dealing with teacher and school change and student learning in science. He is a significant voice for science education reform in Australia. His current research focuses on representation and learning in science, reasoning in science, school-community links, and professional learning of teachers of science.

Tili Wagner
School of Education
Beit Berl College
Email: tiliw@beitberl.ac.il
Tili Wagner is a Lecturer and the elected Director of the School of Education at Beit Berl College, Kfar Saba, Israel. Her major fields of research and expertise are science teaching, with special emphasis on the development of scientific literacy in the schools, and teacher professional development, with emphasis on developing teacher education models.

Bruce Waldrip
Faculty of Education
University of Southern Queensland
Email: waldrip@usq.edu.au
Bruce Waldrip is a Professor of Science Education at the University of Southern Queensland, Toowoomba, Australia. He was the recipient of seven Australian Research Council grants. His current research focuses on utilizing student-generated representations to enhance science learning.

David A. Walker
College of Education
Northern Illinois University
Email: dawalker@niu.edu
David Walker is an Associate Professor of Educational Research & Assessment at Northern Illinois University, DeKalb, IL, USA. He has 135 publications and software programs and 85 presentations throughout his career. His current research focuses on issues of research design, statistical methodology, and school evaluation.

Bing-Jyun Wang
Department of Information Management
Yuan Ze University
Email: imbjw@saturn.yzu.edu.tw
Bing-Jyun Wang is an Associate Professor of Information Management at Yuan Ze University, Chung-Li, Taiwan. He is a former Associate Editor of *Management Review* and served as Councilor to the National Economic Development Meeting and representative to the National Development Council Meeting. His current research focuses on the issues of organization theory, organizational culture and value, business ethics, and organizational learning and strategy.

Morgan C. Wang
Department of Statistics & Actuarial Science
University of Central Florida
Email: cwang@mail.ucf.edu
Morgan Wang is a Professor of Statistics & Actuarial Science at the University of Central Florida, Orlando, FL, USA. He is an elected member of the International Statistical Institute and was the winner in the 2004 Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining Competition and the 2000 Data Visualization Contest of SAS Users Group International Conference. His current research focuses on the issues of data mining and competing analytics, with focus on business intelligence.

Paul Webb
Faculty of Education
Nelson Mandela Metropolitan University
Email: paul.webb@nmmu.ac.za
Paul Webb is a Professor of Science Education and Director of the Research, Technology, & Innovation Unit at the Nelson Mandela Metropolitan University, Port Elizabeth, South Africa. His research interests have been in the fields of the nature of science, classroom interactions, and teaching science in disadvantaged schools. His current interest is in developing new insights into the notion of scientific literacy and promoting scientific literacy at school level, particularly amongst second language learners.

Daphne van Weijen
Utrecht Institute of Linguistics
Utrecht University
Email: Daphne.vanWeijen@let.uu.nl
Daphne van Weijen is a Junior Researcher at Utrecht University, The Netherlands. Her current doctoral research focuses on the effect of the orchestration of cognitive activities during the writing process on text quality in L1 Dutch and L2 English.

Catherine J. Welch
Iowa Testing Programs
University of Iowa
Email: catherine-welch@uiowa.edu
Catherine Welch is a Professor at the University of Iowa. Iowa City, IA, USA. She is the Director of the statewide testing program for Iowa Testing Programs. Her current research focuses on issues of large-scale assessment and performance assessment scoring.

Part I
General Introduction

Chapter 1
Education Research Meets the "Gold Standard": Evaluation, Research Methods, and Statistics after No Child Left Behind

Mack C. Shelley II, Larry D. Yore, and Brian Hand

The fields of education, health sciences, and social sciences internationally have faced calls for better understanding of available datasets and research results from a variety of political, professional, and academic communities. Politicians, bureaucrats, administrators, and other managers desire compelling, evidence-based results and generalizations that they can use as foundations for public policy actions, to make decisions about public spending on research, effective practices, and available services, and to outline future policy directions, strategic plans, and funding demands. Sadly, some handbooks on research used in education do not mention these pressures or the need to craft research reports to inform and persuade a variety of audiences other than like-thinking academics—and they may only briefly consider the issue of generalizability. This is unfortunate in that much of the impact of high-quality, rigorous inquiries are lost or their results are having very limited effect to inform and persuade the various stakeholders because (a) the language used does not make access easy and (b) the findings may be viewed as isolated *info-bits* anchored strictly to unique problems, contexts, or settings not applicable widely or to a particular target audience's concerns or constituents. The call for evidence-based curricula, instruction, and professional education needs to be taken seriously; effort needs to be asserted on policy and decision makers to ensure their definition of evidence means quality, valid, and trustworthy evidence—not simply quantitative evidence of any level of validity and reliability.

Nowhere are these pressures more clearly defined than in elementary, secondary, and postsecondary education and in literacy, mathematics, and science education. Jonathan Osborne (2007), in his remarks as Past President of the National Association for Research in Science Teaching (NARST), called for more armchair

M.C. Shelley II
Iowa State University

L.D. Yore
University of Victoria

B. Hand
University of Iowa

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

science education; he claimed that 50 years of research, curriculum development, and implementation have not presented consistent and compelling patterns of outcomes. This realization by others and the pressures have provided much of the momentum behind several national task force reports on education research, reforms for science education, language education, and national or provincial/state laws or policies regarding education practice, research ethics, instructional materials, and school funding.

The focus, comments, and applications in this book are positioned within this international and interdisciplinary context, which reflects the US National Research Council (NRC) reports (Michaels, Shouse, & Schweingruber, 2008; US NRC, 2000; 2002, 2004, 2005a, 2005b, 2005c, 2007), several international reports on science education, language, and literacy education, and science literacy for all as well as articles, symposia, and workshops about education research and science literacy (Hand et al., 2003; Yore & Hand, 2006; Yore et al., 2004). The Gold Standard policy—which was developed in the United States through the 2001 H.R. 1 legislative extension of the Elementary and Secondary Education Act of 1965 (that established federal support for K-12 education) entitled the No Child Left Behind Act of 2001 (NCLB, 2002) and the 2002 H.R. 3801 legislation entitled the Education Sciences Reform Act of 2002 (ESRA, 2002)—occurred within the temporal context of several attempts to better define education outcomes (American Association for the Advancement of Science, 1990; National Council of Teachers of English & International Reading Association, 1996; National Council of Teachers of Mathematics, 2000; US NRC, 1996), improve the effects of schooling (NCLB), and enhance educational research. We do not intend to bash these policies, reforms, and practices from a post hoc privileged position; rather, we wish to explore an era beyond the initial interpretations of the NCLB and the early attempts to implement Gold Standard education research.

The Gold Standard policy has logical underpinnings and public support in trying to provide evidence-based knowledge claims about curriculum and instruction to decision makers—a process commonly referred to as *speaking truth to power*—that will justify and improve the expenditure of large amounts of taxpayer money on education and schooling. The lack of meaningful effects from massive expenditures in the past has placed current funding for education and research at risk. The government's cost–benefit analyses of research on enhanced practice, effective schools, and improved outcomes have not justified the continuation of these expenses. Unfortunately, it appears as if the Gold Standard for research practice (randomized controlled trials [RCTs]) is based on the stage 3 drug trial, or medical model, without duly recognizing the stage 1 and stage 2 trials necessitated by rarity of disease, risks, development of the problem space, availability of related technologies or innovations, and costs. Health sciences researchers frequently use single subjects and very small case-study approaches over an extended period because of the rarity of subjects with the specific disease, undefined side effects, unreasonable risk, and unjustified costs. These health scientists gradually move their research agenda and approach toward more focused questions and larger case studies before moving to full-scale RCTs or randomized placebo studies. Any interpretation of quality research needs to focus on the research agenda—not a single inquiry—and must consider the theoretical and canonical knowledge about the problem space, the associated technology and methodology appropriate to the specific inquiry, as well as ethical considerations arising from the nature of the inquiry and the involvement of human or animal subjects.

Some initial and current interpretations of the Gold Standard have privileged a single research approach and type of evidence regardless of the development of the problem space, specific research question, available technologies and instrumentation, and cost or ethical considerations. If such interpretations of this policy exclusively privilege RCT and quantitative evidence, it would disregard high-quality, qualitative research approaches and other contemporary approaches and, thus, the evidence flowing from such inquiries. Such an oversight would not fully recognize education as a social science that utilizes (a) epistemologies and methods that involve both hypothetico-deductive inquiry or normal hierarchical development and (b) inductive, nonexperimental inquiries that insert new theoretical discourses alongside existing ones (Yore & Lerman, 2008). Literacy and science education have benefited from melding quantitative and qualitative methods and approaches to knowledge building. The question that should be addressed—and the one we address here—is not an either/or issue of the methodological debates of the 1970s and 1980s; rather, it requires rigorous and appropriate consideration of qualitative approaches, quantitative approaches, mixed-methods research, or historical–philosophical inquiries that reflects the research question, current development of the problem space, and availability of associated research techniques, procedures, and technologies.

Recent moves to enhance the quality of research in the United States have influenced the organization, expectations, and operations of federal and state funding agencies, such as the National Science Foundation (NSF) and departments of education. Federal policies have stimulated reorganization of the US Office of Educational Research and Improvement into the Institute of Education Sciences and some of the science, technology, engineering, and mathematics education areas within NSF into Discovery Research K-12. These changes were in part to ensure research as a critical function of all funded projects. Importantly, there is recognition of the need to have both exploratory and ongoing research lines developed within the funding envelopes. Furthermore, scaling is seen as critical in building knowledge—from initial, small-scale projects to expanded explorations in which fidelity and validity of potential results from smaller projects are applied to much larger numbers of participants and diverse contexts. Scaling is viewed as important in moving education research forward into generalized results and applications. Therefore, scaling involves moving from exploratory research designs to clustered, randomized designs prior to moving to nationwide projects or implementation.

Evaluation of proposals and projects has shifted from straight reporting of exposure of or impact on teachers to determining impact on student performance. For example, the Mathematics and Science Partnerships grants, which flow through to the states, all required student impact data. Likewise, the recent Local Systemic Change projects funded by NSF transformed the evaluation requirements beyond the original quality and quantity of professional development for science and mathematics teachers, teacher and administrator surveys, teacher interviews, and classroom observations, to include data on student science and mathematics achievement. Further, the US Department of Education (US ED) research grants are clearly focused on student performance data.

The Gold Standard may be setting the international agenda for research and for funding research. This book does not bash the policy but provides international perspectives and alternatives that are meant to improve the quality of research and the impact of research on policy and decisions. Increasing public awareness of literacy, language, and science education research results and increasing the influence of public policies and decisions about funding, professional education, and practice are amongst the highest, but unrecognized, priorities facing academic communities. Few academics are prepared and experienced to address this lobbying effort. Effective research agendas and programs of study should point toward generalizable claims, powerful explanations, and new theories about learning, teaching, curriculum, and assessment. In addition, individual inquiries should be viewed in the overall context or evolutionary progression of the agenda or program and its potential to provide insights into the pattern of events and explanations under investigation.

1.1 Background

The *Gold Standard(s) of Quality Research in Science and Literacy Education Conference* (NSF Award #0437198) was held October 26–30, 2005, at Dunsmuir Lodge, on Vancouver Island, British Columbia, Canada. Program planning involved researchers and scholars from Iowa State University, the University of Alberta, the University of Georgia, the University of Iowa, the University of Missouri-St. Louis, and the University of Victoria as well as infrastructural and staff support from Iowa State University, the University of Iowa, and the University of Victoria. The 2nd Island Conference built on the successes and design of the 1st Island Conference held at the same site on September 12–15, 2002, entitled *Ontological, Epistemological, Linguistic and Pedagogical Considerations of Language and Science Literacy: Empowering Research and Informing Instruction* (NSF Award #0210002).

The interdisciplinary and collaborative nature of the conference's program design dictated that international research teams were invited from Asia, Australasia, Europe, and North America consisting of senior faculty, junior faculty, and graduate students from language and literacy, measurement, psychology, science education, and statistics. Participants included internationally recognized faculty (together with their graduate students and support staff) specializing in science education, reading and literacy research, statistics, and educational research methods from universities. A large proportion of participants were female and relatively junior researchers in their respective areas of specialization (see the Appendix for a listing of participants, planning team, and support staff). As planning for the conference unfolded, it became increasingly clear that the issues addressed within the US context played out differently in other countries although many of the same issues arise throughout the international arena. Accordingly, the breadth of the conference and its deliberations were expanded to encompass multinational perspectives.

The purpose of the 2nd Island Conference was to address the implications for research in education—especially in science education and literacy—of the expectations for Gold Standard research. This standard is encouraged, supported, and enforced by the US ED Institute of Education Sciences to provide evidence-based interventions on educational outcomes that have been found to be effective in “randomized controlled trials – research’s ‘gold standard’ for establishing what works” (US ED, 2003, p. iii) in education research, following patterns of evidence used in medicine and welfare policy. With appropriate national variants, similar expectations have emerged in other countries from education ministries, for researchers in other countries and in particular for those collaborating with colleagues in the United States, and for researchers working with support from multinational and nonprofit sources.

The US ED specifies standards for the quality of evidence required using different research designs. The quality needed to establish *strong* evidence includes the application of specific designs, methods, procedures, and practice. *Possible* evidence may include some of these expectations but fall short of providing strong evidence, and/or comparison-group studies in which the intervention and comparison groups are very closely matched on academic achievement, demographics, and other characteristics that may confound interpretations of program effectiveness. The official government criteria for evaluating whether an intervention is backed by strong evidence of effectiveness hinge on several key qualities of the research design: (a) randomized controlled trials that are well designed and implemented, (b) demonstrating that there are no systematic differences between intervention and control groups before the intervention, (c) using measures and instruments of proven validity, (d) real-world objective measures of the outcomes that the intervention is designed to affect, (e) attrition over time of no more than 25% of the original sample, (f) reasonable effect size combined with statistical significance, (g) a sample size sufficiently large to achieve statistical significance, and (h) controlled trials implemented in more than one site in schools that represent a cross section of all schools. In general, these expectations are difficult to realize in most concrete applications and frequently have not been well understood by many education researchers.

These directives regarding quality research in education are a consequence of two recent tsunamis of federal legislation that have cascaded upon the shores of education research: (1) the No Child Left Behind Act of 2001 and (2) the enactment of federal Gold Standard guidelines for fundable federal research that was at least implicit in NCLB and made much more explicit in the Education Sciences Reform Act of 2002. This confluence of legislative initiatives has been the source of considerable soul-searching—and angst—among many education researchers. NCLB established standards for academic assessments in mathematics, reading or language arts, and science that involved multiple measures of student academic achievement, including measures that assess higher-order thinking skills and understanding. These requirements for program assessment lead to many opportunities and circumstances for the application of statistical research methods and qualitative methods of equivalent rigor.

NCLB (2002) also established standards for indicators of program quality, as a key part of the evaluation process, to monitor, evaluate, and improve those programs. For adult program participants, the indicators include:

> (A) achievement in the areas of reading, writing, English-language acquisition, problem solving, and numeracy; (B) receipt of a secondary school diploma or a general equivalency diploma (GED); (C) entry into a postsecondary school, job retraining program, or employment or career advancement, including the military; and (D) such other indicators as the State may develop. (115 STAT, 20 USC 6381i., § 1240(1))

Acceptable indicators for children were delineated as: "(A) improvement in ability to read on grade level or reading readiness; (B) school attendance; (C) grade retention and promotion; and (D) such other indicators as the State may develop" (115 STAT. 1566, 20 USC 6381i., § 1240(2)).

The research program under NCLB was designed to examine the effect of the assessment and accountability systems on students, teachers, parents, families, schools, school districts, and states, including correlations between such systems and:

- Student academic achievement, progress to the state-defined level of proficiency, and progress toward closing achievement gaps, based on independent measures.
- Changes in course offerings, teaching practices, course content, and instructional materials.
- Changes in turnover rates among teachers, principals, and pupil-services personnel.
- Changes in dropout, grade-retention, and graduation rates for students.
- The effect of the academic assessments on students with disabilities.
- The effect of the academic assessments on low, middle, and high socioeconomic status students, limited and nonlimited.
- English proficient students, racial and ethnic minority students, and nonracial or non-ethnic minority students.
- Guidelines for assessing the validity, reliability, and consistency of those systems using nationally recognized professional and technical standards.
- The relationship between accountability systems and the inclusion or exclusion of students from the assessment system.

Scientific research in education is essential for sustainable progress to be made in such critical areas as science education, language arts, literacy, writing, and reading research. With emphasis on these essential aspects of student outcomes, the 2nd Island Conference was intended to provide a comprehensive examination of how Gold Standard research can be used to ascertain which programs and interventions are most effective in enhancing student achievement in science knowledge and reading proficiency. The specific purposes of the conference were to attempt the following:

- Clarify the established and emerging roles of quantitative methods in contemporary education research, focused on K-12 science education and literacy.

- Explore the influences of new software and enhancements of information technologies on the mathematical and statistical investigation of language and science literacy.
- Establish a contemporary theoretical framework involving language, information technology, science, and science literacy, anchored by fundamental ideas of applied statistical and mathematical methods, to provide concrete guidance for future work on theoretical and pedagogical issues of language and science learning.
- Extend, deepen, and begin to institutionalize the dialogue that should be taking place among quantitative research experts and experts in education research, involving applied cognitive science, language, and science education researchers, together with graduate students and teacher educators.
- Come to grips with the extent to which alternative research methods (qualitative or mixed methods) may be integrated most effectively with quantitative research modalities.

Questions addressed by presenters at the conference included:

- How should *hard* and *soft* modeling methods be used to study science education and literacy?
- How can qualitative methods be linked most productively to quantitative procedures in education research?
- How can education research benefit from applications of longitudinal, panel, and long-term analytical methods?
- To what extent do the provisions of H.R. 3801 intersect with those of NCLB?
- What new methods and what forms of advanced education and professional development will be required to conduct future education research?
- What role can meta-analysis and metasynthesis play in helping to stimulate a broader, scientific-based effort for comprehensive education reform?

A broader purpose of the conference was to address the extent to which it is possible to make education research compatible with the requirements and guidelines of H.R. 3801 and of subsequent interpretations of the meaning of that legislation by the US ED and other education research funders around the world. In the minds of many education researchers, a direct link exists between the H.R. 3801 language and the language of NCLB. In particular, the tools provided by contemporary statistical methods are of increased importance as means by which to satisfy the requirements and expectations of H.R. 3801 and NCLB.

The planning team has remained active in subsequent efforts to disseminate the conference findings, increase awareness, and build understanding. In addition, the principal investigators served as organizers, presenters, and panel members for presentations at national and international meetings of statistical and science education associations entitled "Education Research Meets the Gold Standard: Statistics, Education, and Research Methods after *No Child Left Behind*" (Joint Statistical Meetings, Minneapolis, MN, August 7–11, 2005), "Research Committee Symposium: Research Ethics Boards and Gold Standard(s) in Education Research"

(National Association for Research in Science Teaching, San Francisco, CA, April 13, 2006), "Gold Standard(s) for Research in Science Literacy: Results of an NSF-sponsored Conference" (Australasian Science Education Research Association, Canberra, NT, July 5, 2006), "Research Committee Symposium: Gold Standards for Research" and "President's Symposium: Critical Look at Science Education Research" (National Association for Research in Science Teaching, New Orleans, LA, April 16, 2007). Large audiences attended these sessions, and many of the association members participated actively in a robust discussion about the use of statistics in achieving science-based standards in education research and suggested local, national, and international actions arising from the presentations and discussion.

1.2 Organization of the Book

Part II of this book focuses on Setting the Agenda: Science Education and Science-Based Research in an era of political pressures, reduced funding for research, great potentials outlined by numerous reforms and task force reports, and sincere need for reconsideration of positions and advocacy for specific research methods in terms of public mandates of the No Child Left Behind and Education Sciences Reform acts. The insightful resolution of these issues will bring common good to the public and academic communities and provide a strong foundation from which to lobby for informed policies and decisions about education, literacy and language, and science education research.

The 2nd Island Conference built on the collective outcomes of the 1st Island Conference and the scholarship, actions, and research flowing from it. Now, we attempt to share the collective concerns, deliberations, and recommendations from the 2nd Island Conference by addressing the following issues:

- Why "Gold Standard" Needs Another "s": Results from the Gold Standard(s) in Science and Literacy Education Research Conference
- Research and Practice: A Complex Relationship?
- Moving Beyond the Gold Standard: Epistemological and Ontological Considerations of Research in Science Literacy
- Longitudinal Studies in Science Learning—Methodological Issues
- An International Perspective of Monitoring Educational Research Quality: Commonalities and Differences

The closing chapter of Part II addresses Considering Research Quality and Applicability through the Eyes of Stakeholders and moving the agenda forward into a post-Gold Standard era while being aware of the unproductive history of the research debates.

Part III of this book focuses on Curriculum and Pedagogy informed by quality research and progressive research agendas with literacy and language and science education inquiries. This part considers:

- Researching Effective Pedagogies for Developing the Literacies of Science: Some Theoretical and Practical Considerations
- Pedagogy, Implementation, and Professional Development for Teaching Science Literacy: How Students and Teachers Know and Learn
- Approaching Classroom Realities: The Use of Mixed Methods and Structural Equation Models in Science Education Research
- Mixed-methodology Research in Science Education: Opportunities and Challenges in Exploring and Enhancing Thinking Dispositions

The final chapter of Part III outlines New Directions in Science Literacy Education.

Part IV of this book focuses on Statistics, Research Methods, and Science Literacy that are not fully recognized or anticipated by the Gold Standard, but are if a second ‘s’ is added and the focus and guiding principles are Gold Standards. This part considers:

- Multilevel Modeling with HLM: Taking a Second Look at PISA
- Methods from Item Response Theory: Going Beyond Traditional Validity and Reliability in Standardizing Assessments
- Confounding in Observational Studies using Standardized Test Data: Careful Disentanglement of Statistical Interpretations and Explanations
- Predicting Group Membership using National Assessment of Educational Progress (NAEP) Mathematics Data
- Incorporating Exploratory Methods using Dynamic Graphics into Multivariate Statistics Classes: Curriculum Development
- Approaches to Broadening the Statistics Curricula
- Dr. Fox Rocks: Using Data-mining Techniques to Examine Student Ratings of Instruction
- Process Execution of Writing and Reading: Considering Text Quality, Learner and Task Characteristics

The last chapter of Part IV addresses a critical and perplexing issue: Can We Make a Silk Purse from a Sow’s Ear?

Part V of this book focuses on Public Policy and Gold Standard(s) Research and addresses:

- Speaking Truth to Power with Powerful Results: Impacting Public Awareness and Public Policy
- Funding Patterns and Priorities: An International Perspective
- Research Ethics Boards and the Gold Standard(s) in Literacy and Science Education Research
- Data Sharing: Disclosure, Confidentiality, and Security
- Stitching the Pieces Together to Reveal the Generalized Patterns: Systematic Research Reviews, Secondary Reanalyses, Case-to-case Comparisons, and Metasyntheses of Qualitative Research Studies

This part closes by addressing The Gold Standard and Knowing What to Do.

Part VI of this book, Epilogue: New Standards, New Directions, and New Realities, focuses on providing collective insights based on the four substantive sections.

1.3 Final Remarks

As you progress through the following parts and chapters, keep in mind that it is not the intent to bash or second-guess, but rather to build on the past experiences and move toward an informed position about literacy and science education. This involves open, honest, public debate about *Science Literacy for All*, the value of different research approaches, and making a difference for learners. The design of, and invitation to, the 2nd Island Conference emphasized diversity of backgrounds, personalities, problems, approaches, and solutions. The diversity manifested itself in passionate debate and in voices in these chapters that provide basic ingredients and insights on which you can form your own position and solutions.

References

American Association for the Advancement of Science. (1990). *Science for all Americans: Project 2061*. New York: Oxford University Press. Available from http://www.project2061.org/publications/sfaa/online/sfaatoc.htm

Education Sciences Reform Act of 2002. Pub. L. No. 107–279, 116 Stat. 1940. (2002).

Hand, B., Alvermann, D. E., Gee, J. P., Guzzetti, B. J., Norris, S. P., Phillips, L. M., et al. (2003). Message from the “Island group”: What is literacy in science literacy? [Guest editorial]. *Journal of Research in Science Teaching, 40*(7), 607–615.

Michaels, S., Shouse, A. W., & Schweingruber, H. A. (2008). *Ready, set, science! Putting research to work in K-8 science classrooms*. Board on Science Education, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Council of Teachers of English & International Reading Association. (1996). *Standards for English language arts*. Urbana, IL: National Council of Teachers of English. Available from http://www.ncte.org/about/over/standards?source = gs

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author. Available from http://standards.nctm.org/

No Child Left Behind Act of 2001. Pub. L. No. 107–110, 115 Stat. 1425. (2002).

Osborne, J. (2007). In praise of armchair science education. *E-NARST News, 50*(2). Retrieved from http://www.narst.org/news/e-narstnews_july2007.pdf

United States Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Washington, DC: Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Available from http://ies.ed.gov/ncee/pubs/evidence_based/evidence_based.asp

United States National Research Council. (1996). *The national science education standards*. Washington, DC: The National Academies Press. Available from http://www.nap.edu/catalog.php?record_id = 4962

United States National Research Council. (2000). *How people learn: Brain, mind, experience, and school—Expanded edition*. Committee on Developments in the Science of Learning.

J. D. Bransford, A. L. Brown, & R. R. Cocking (Eds.). *Commission on Behavioral and Social Sciences and Education*. Washington, DC: The National Academies Press.

United States National Research Council. (2002). *Scientific research in education*. Committee on Scientific Principles for Education Research. R. J. Shavelson & L. Towne (Eds.). *Center for Education, Division of Behavioral and Social Sciences and Education*. Washington, DC: The National Academies Press.

United States National Research Council. (2004). *Advancing scientific research in education*. Committee on Research in Education. L. Towne, L. L. Wise, & T. M. Winters (Eds.). *Center for Education, Division of Behavioral and Social Sciences and Education*. Washington, DC: The National Academies Press.

United States National Research Council. (2005a). *How students learn: History, mathematics, and science in the classroom*. Committee on *How people learn,* A Targeted Report for Teachers. M. S. Donovan & J. D. Bransford (Eds.). *Division of Behavioral and Social Sciences and Education*, DC: The National Academies Press.

United States National Research Council. (2005b). *How students learn: Mathematics in the classroom*. Committee on *How people learn,* A Targeted Report for Teachers. M. S. Donovan & J. D. Bransford (Eds.). *Division of Behavioral and Social Sciences and Education*. Washington, DC: The National Academies Press.

United States National Research Council. (2005c). *How students learn: Science in the classroom*. Committee on *How people learn,* A Targeted Report for Teachers. M. S. Donovan & J. D. Bransford (Eds.). *Division of Behavioral and Social Sciences and Education.* Washington, DC: The National Academies Press.

United States National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Committee on Science Learning, Kindergarten through Eighth Grade. R. A. Duschl, H. A. Schweingruber, & A. W. Shouse (Eds.). *Board on Science Education, Center for Education, Division of Behavioral and Social Sciences and Education*. Washington, DC: The National Academies Press.

Yore, L. D., & Hand, B. (2006). Gold standard(s) for research in science literacy: Results of an NSF-sponsored conference at the University of Victoria’s Dunsmuir Lodge, October 24–30, 2005. Paper presented at the annual meeting of the Australasian Science Education Research Association, Canberra, Australia.

Yore, L. D., Hand, B., Goldman, S. R., Hildebrand, G. M., Osborne, J., Treagust, D. F., et al. (2004). New directions in language and science education research. *Reading Research Quarterly*, *39*(3), 347–352.

Yore, L. D., & Lerman, S. (2008). Metasyntheses of qualitative research studies in mathematics and science education [Editorial]. *International Journal of Science and Mathematics Education*, *6*(2), 217–223.

Appendix

Roster of participants in the October 2005 Vancouver Island Conference, with Institutional and National Affiliations at that time

Surname	First name	Institutional affiliation	Location
Akkus	Recai	Graduate Student, Iowa State University	Ames, IA, USA
Alvermann	Donna	Distinguished Research Professor, University of Georgia	Athens, GA, USA
Anderson	John	Professor, University of Victoria	Victoria, BC, Canada
Anthony	Robert	Associate Professor, University of Victoria	Victoria, BC, Canada
Battistich	Victor	Professor, University of Missouri	St. Louis, MO, USA
van den Bergh	Huub	Professor, University of Amsterdam	Amsterdam, The Netherlands
Bordeaux	Janice	Professor, Rice University	Houston, TX, USA
Boscolo	Pietro	Professor, DPSS-University of Padova	Padova, Italy
Brickhouse	Nancy	Professor, University of Delaware	Newark, DE, USA
Cavagnetto	Andy	Graduate Student, University of Iowa	Iowa City, IA, USA
Chang	Wen-Hua (Judy)	Associate Professor, National Taiwan Normal University	Taipei, Taiwan
Coll	Richard	Associate Professor, University of Waikato	Hamilton, New Zealand
Cook	Dianne	Professor, Iowa State University	Ames, IA, USA
Donaldson	Allison	Conference Coordinator, Iowa State University	Ames, IA, USA
Ford	Danielle	Associate Professor, University of Delaware	Newark, DE, USA
Glynn	Shawn	Josiah Meigs Distinguished Teaching Professor, University of Georgia	Athens, GA, USA
Grimberg	Irene	Associate Research Professor, Montana State University	Bozeman, MT, USA
Gunel	Murat	Graduate Student, Iowa State University	Ames, IA, USA
Hand	Brian	Professor, Iowa State University	Ames, IA, USA
Hohenshell	Liesl	Postdoctoral Fellow, Western Washington University	Bellingham, WA, USA
Hsiung	Chao-Ti	Associate Professor, National Taipei University	Taipei, Taiwan
Johnson	Steve	Professor, University of Central Florida	Orlando, FL, USA
Levin	Tamar	Professor, Tel Aviv University	Tel Aviv, Israel
Macnab	John	Graduate Student, University of Alberta	Edmonton, AB, Canada
McEneaney	Elizabeth	Associate Professor, California State University, Long Beach	Long Beach, CA, USA
Meyer	Mary	Professor, University of Georgia	Athens, GA, USA
Mundfrom	Daniel	Professor, University of Northern Colorado	Greeley, CO, USA

(continued)

Appendix (continued)

Surname	First name	Institutional affiliation	Location
Narayan	Ratna	Graduate Student, University of Georgia	Athens, GA, USA
Nieswandt	Martina	Assistant Professor, University of Toronto	Toronto, ON, Canada
Nolan	Deborah	Professor, University of California, Berkeley	Berkeley, CA, USA
Norris	Stephen	Professor and Canada Research Chair, University of Alberta	Edmonton, AB, Canada
Phillips	Linda	Professor, and Director of the Center for Research on Literacy, University of Alberta	Edmonton, AB, Canada
Prain	Vaughan	Professor, LaTrobe University – Bendigo	Bendigo, Victoria, Australia
Rijlaarsdam	Gert	Professor, University of Amsterdam	Amsterdam, The Netherlands
Ross	Shelly	Graduate Student, University of Victoria	Victoria, BC, Canada
Romance	Nancy	Professor, Florida Atlantic University	Boca Raton, FL, USA
Rumann	Corey	Graduate Student, Iowa State University	Ames, IA, USA
Saul	Wendy	Allen B. & Helen S. Shopmaker Endowed Professor, University of Missouri, St. Louis	St. Louis, MO, US
She	Hsiao-Ching	Professor, National Chiao Tung University	Hsinchu, Taiwan
Shelley	Kathy	Department of Statistics, Iowa State University	Ames, IA, US
Shelley	Mack	Director of Research Institute for Studies in Education, and Professor, Iowa State University	Ames, IA, USA
Tolbert	Sara	Graduate Student, University of Georgia	Athens, GA, USA
Turner	Gwen	Associate Professor, University of Missouri, St. Louis	St. Louis, MO, USA
Tytler	Russell	Professor, Deakin University	Burwood, Victoria, Australia
Vitale	Michael	Professor, East Carolina University	Greenville, NC, USA
Wagner	Tili	Professor, Weizmann Institute	Rehovot, Israel
Walker	David	Professor, Northern Illinois University	DeKalb, IL, USA
Wang	Morgan	Professor, University of Central Florida	Orlando, FL, USA
Yang	Eunmi (Olivia)	Graduate Student, Iowa State University	Ames, IA, USA
Yore	Larry	Distinguished Professor, University of Victoria	Victoria, BC, Canada

Part II
Setting the Agenda: Science Education and Science-based Research

Chapter 2
Why "Gold Standard" Needs Another "s": Results from the Gold Standard(s) in Science and Literacy Education Research Conference

Larry D. Yore and Pietro Boscolo

Silver, gold, diamond, and platinum are symbols of quality. International advertising and marketing stress their rarity, beauty, symbolic, and emotional considerations. Infrequently do these promotional efforts mention that rarity can result from natural scarcity or from controlled access to the supply of the materials, and few ever mention the concepts of pragmatics and value as a proportional consideration of quality, cost, and utility. Concerns about quality have been heard in the language and literacy, learning and instruction, measurement and statistics, and science education research communities since the 1980s. Voices of reason have occasionally risen above the din of the simplistic *either*/*or* positions in the irrational quantitative–qualitative debates. The opposing sides of purists in this unproductive endeavor appear more interested in impressing one another rather than informing and persuading the opposition about quality research and benefits of comingling methods to better match the problem space and available instrumentation and technology. Furthermore, these rhetorical arguments (i.e., oratorical and discursive techniques designed to persuade) do not appear to recognize (a) the contemporary modern view of science (postpositivist); (b) education as a social science rather than a natural science; (c) the strengths and rigor required of the new learning sciences; and (d) the need for a long-term research agenda that targets a problem space and topic, addresses worthwhile and perplexing questions, and persists in the inquiry using appropriate investigations, which evolve and progress toward sound evidence-based arguments, generalized knowledge claims, and explanations involving causality and mechanism (Johnson & Onwuegbuzie, 2004; Phillips, 2006; Yore, 2003).

The debates continued as both purist quantitative and qualitative researchers conducted serial investigations with little visible growth and without apparent utilization of and connection to experiences and results from earlier inquiries. This can be seen

L.D. Yore
University of Victoria

P. Boscolo
University of Padova

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

in the research literature where, for example, you can find sequences of aptitude–treatment–interaction (ATI) inquiries in which one new attribute after another was arbitrarily substituted for the previous attribute in two-way analyses of variance approaches, and a series of grounded theory investigations of the same topic without using prior findings to frame hypotheses, venture tentative answers, make predictions, or craft an interpretative framework for the next inquiry. These concerns applied equally to both research camps, but neither side appeared to recognize the risks and ultimate outcomes from the government and funding agencies and the ever-decreasing trust in education research by policy makers, decision makers, and other stakeholders—thus, the No Child Left Behind Act of 2001 (NCLB, 2002) and the Education Sciences Reform Act of 2002 (ESRA, 2002) in the United States.

The Gold Standard(s) of Quality Research in Science and Literacy Education Conference, also known as the 2nd Island Conference, attempted to stimulate honest and open deliberations about these two US federal laws and related research issues from international and interdisciplinary perspectives. The international perspective provided a distant objectivity and content and methodological perspectives not popular in North American research communities. The diverse collection of experienced and new qualitative, quantitative, and mixed-methods researchers from literacy and language, learning and instruction, measurement and statistics, and science education focused on moving forward from the qualitative–quantitative debates and with the benefits of 3 years of Gold Standard experience.

2.1 Background Context

Learning and education research in general have received considerable attention in the United States in this decade (US National Research Council [NRC], 2000, 2002, 2004, 2007). Embedded in this context of committee reports and ongoing education reforms, two important laws were enacted by the US federal government: 2001 HR 1 and 2002 HR 3801, commonly called *No Child Left Behind* and the *Gold Standard for Education Research*. Both laws reflected politicians' and taxpayers' sincere disappointment in the effects of public schools and their deep skepticism about the quality and rigor of education research. These political, ideological documents were only temporally, not logically, connected to the NRC committees' reports (2000, 2002). The ESRA connected to the NCLB goals of equal educational opportunities and elimination of inequities in school achievement; it led to the establishment of the Institute of Education Sciences (IES) in the US Department of Education (US ED) to replace the Office of Educational Research and Improvement (Brickhouse, 2006). IES compared and linked education research efforts to the successes and progress in medicine and social welfare research and policy, and set a priority to establish evidence-based school practices, teaching strategies, instructional materials, programs, and policies for education.

2.1.1 Learning and Education Research

The NRC Committee on Developments in the Science of Learning (US NRC, 2000) reported on the deliberations of a select committee that summarized the research results of cognitive sciences and constructivist perspectives and recognized the importance of language in learning. This report established three general principles about how people learn:

- People come to learning with prior conceptions about the world (natural and people-built) that must be engaged or challenged if new or refined conceptions are to be developed.
- Enhanced competence requires prior foundational knowledge, conceptual frameworks, and storage, retrieval, and application strategies.
- Learning requires metacognition to be aware, monitor, and control meaning making and transference of learning to new situations.

This report stated, “Students often have limited opportunities to understand or make sense of topics because many curricula have emphasized memory rather than understanding” (pp. 8–9).

The NRC Committee on Scientific Principles for Education Research (US NRC, 2002) identified six principles for improving the quality of education research: (a) significant questions that can be investigated empirically, (b) linked to relevant theory, (c) use methods that allow direct investigation, (d) provide coherent and explicit reasoning chains, (e) replicate and generalize across studies, and (f) allow and encourage professional scrutiny and critiques. This report also provided design features of funding agencies. It suggested that funding agencies should be driven by a commitment to the scientific principles, insulated from political micromanagement, provided a long-range funding envelope, and committed to transparency. Furthermore, the report specified that funding agencies should be staffed by people skilled in science, leadership, and management; have structures in place to guide research agendas, inform funding decisions, and monitor progress; provide insulation from inappropriate, external interference; focus on balanced research that addresses short-term, medium-term, and long-term issues; and adequately fund investment in research infrastructure. Collectively, the 2002 NRC report outlined middle-of-the-road solutions that attempted to mediate rumors of pending legislation and did not privilege either the extreme qualitative or quantitative perspectives (Phillips, 2006).

Advancing Scientific Research in Education (US NRC, 2004) outlined strategic objectives focused on upgrading research approaches, promoting quality research, building the knowledge base, and professional development of researchers. This NRC Committee on Research in Education report reflected the ESRA by promoting randomized controlled trials (RCTs) and emphasizing cause–effect relationships. At first glance, both of these upgrading strategies appear to privilege quantitative approaches without consideration of the broader spectrum of research approaches and types of data. But upon deeper consideration, it becomes clear that the committee believed that quality (qualitative and quantitative approaches) could be enhanced by addressing features of the funding agencies and by selecting rigorous methods. Regarding funding agencies, the report encouraged the establishment and clear delineation of criteria for research; selection of evaluation panels that represent a wide range of expertise, frameworks, and groups; minimization of potential conflicts-of-interest and narrow perspectives; and provision of professional development for panel members to ensure valid and consistent judgments. It suggested that rigorous methods, including both quantitative and qualitative approaches, should stress alignment amongst research question, problem space, and educational setting; ensure appropriate resources for large-scale studies; and support scaling innovations and building capacity that are far more complex than simply a multiplier of sample size. The committee reported that building the knowledge base involved (a) establishing explicit ethical standards, infrastructures, and security for data sharing and (b) encouraging research journals to require authors to make data available and to require structured abstracts. The report suggested that professional development for researchers should start by encouraging doctoral programs that stress research competencies and deep, substantive, methodological knowledge and skills, and provide research internships.

2.1.2 Concerns within the Science Education Research Communities

The NRC Committee on Science Learning, Kindergarten through Eighth Grade (US NRC, 2007), suggested that much science education research has been based on outmoded views of cognition that did not recognize fully learners' prior knowledge and reasoning abilities about ideas and events and that people's informal experiences, reasoning, and intuition provide starting points for developing understanding, plausible reasoning, critical thinking, and reflections. Many commonly held views of cognition (a) stressed conceptual learning, which did not include psychomotor performances and affective outcomes; (b) stressed learners' deficits rather than their diverse assets; (c) underemphasized language as a cognitive tool; (d) discounted social transmission and scaffolding amongst learners and more expert peers, adults, and mentors; and (e) were unaware of the need for metacognition (US NRC, 2000, 2005, 2007). These reports differentiated between conceptual growth and change; they suggested that learning trajectories may not be smooth but rather a sawtooth pattern punctuated with learning advances and forgetting retreats. Conceptual growth—additions to well-established conceptual networks—occurs easily; but conceptual changes are far more difficult, requiring learners to integrate unfamiliar ideas and reorganize prior conceptual networks. Therefore, effective learning may involve changes to learners' conception of the nature of science and metacognitive improvements regarding how knowledge is stored and retrieved for future use. Furthermore, these reports subtly encourage researchers and teachers to view teaching in service of effective learning—rather than the traditional perspective that effective teaching causes learning.

The NRC (US NRC, 2000, 2005, 2007) viewed learning as an interactive, constructive process and recognized the importance of language in developing, shaping, and reporting understanding. People’s experiences and their environment can facilitate their language growth, communication abilities, and academic discourses—but activity alone is not sufficient to promote effective learning. The hands-on science reforms of the 1960s led to *activity mania* and the associated belief that lack of learning could be overcome with just one more activity, without any consideration of processing and discussing the experiences. The current international science education reforms—Science Literacy for All—generally accept the symbiotic relationship between fundamental literacy and the derived understanding of the big ideas of science (Norris & Phillips, 2003; Yore, Pimm, & Tuan, 2007). Furthermore, natural language (home language) is the starting point toward acquiring the science language (conceptual and procedural terminology) and scientific metalanguage (argument, theory, hypothesis, model, inference, observation, measurement, etc.). Moving from home language to school language and on to scientific language is essential in becoming science-literate and will reflect the person’s worldview of science (Yore, 2008; Yore & Treagust, 2006).

Science education has been engaged in heated disputes around the preferred and dominant research approach for exploring the relationships amongst science learning, teaching, and assessment for over 25 years. Science educators and researchers sense a duality in their fundamental existence—science (natural sciences) and education (social sciences)—that has caused a split identity, persistent conflicts, and dilemmas for many. Unfortunately, the purists from the extreme of the duality believed in the either/or solution rather than a pragmatic comingling of the ontological, epistemological, methodological, and other theoretical beliefs/assumptions characteristic of each approach (Johnson & Onwuegbuzie, 2004).

Critics of the scientific experimental approaches, which did not reflect the uncertainty of observations and measurements and the tentativeness of statistics-based claims, assumed an extreme postmodernist (multiplist or relativist) stance in which multiple interpretations could flow from the same datasets and information files and did not accept that these claims needed to be submitted to a public evaluation to judge validity, trustworthiness, and utility. This was viewed as the *slippery slope* by some researchers. Johnson and Onwuegbuzie (2004) suggested that quantitative purists maintain that social science inquiry should be objective and stated:

> [R]eal causes of social scientific outcomes can be determined reliably and validly....These researchers have traditionally called for rhetorical neutrality, involving a formal writing style using the impersonal passive voice and technical terminology, in which establishing and describing social laws is the major focus.... [While qualitative] purists contend that multiple-constructed realities abound, that time- and context-free generalizations are neither desirable nor possible, that research is value-bound, that it is impossible to differentiate fully causes and effects, that logic flows from specific to general, ... and that knower and known cannot be separated because the subjective knower is the only source of reality. (p. 14)

This binary ontological–epistemological view of the realist, absolutist worldview and idealist, relativist worldview represents extreme positions along the knowledge-construction continuum and provides little consideration of the middle-of-the-road modernist naïve realist, evaluativist worldview (Phillips, 2006; Yore, Hand, & Florence, 2004). The naïve realist, evaluativist worldview recognizes different interpretations of data because of the interpreter's lived experiences, theoretical perspectives, and prior knowledge; but this worldview requires all knowledge claims to be submitted to public evaluation based on the available evidence and canonical knowledge. This centrist position encourages comingling of research approaches (mixed methods) and elevates analytical and dialectic argumentation in educational inquiries to primary position: "Has the overall case made by the investigator been established to a degree that warrants the tentative acceptance of the theoretical or empirical claims that are being put forward?" (Phillips, p. 24). These types of argument make evidence and logic essential and avoid the rhetorical forms based on eloquent language and linguistic–persuasive devices. Toulmin (1958) suggested that sound and compelling arguments involve evidence-based claims within well-articulated theoretical backings that stand the test of criticism.

Within this unyielding climate and with little indications of the war moving toward rational resolution, a variety of editorials, committee reports, and finally public policies addressed the need for quality research, evidence, and results on which to base education policies, decisions, and professional education and to evaluate practices and instructional materials (Phillips, 2006; US IES, n.d.). Many of the concerns about education research have to do with rigorous inquiry, informed choice, plausible reasoning and logic, and appropriate argument.

Lawson (2005) provided an outline to judge "good science" in quantitative and qualitative science and mathematics education research involving the formation of "interesting and important question[s in an agenda that moves from descriptive-level] who, what, and where [questions to causation-level] why and how" questions (p. 1). The *who, what,* and *where* questions can be addressed with inductive inquiries intent on formulating more acute *why* and *how* questions, discerning patterns, and producing hypotheses. Once the later questions become the inquiry focus, the researcher needs to formulate tentative answers in the form of speculations, predictions, and hypotheses based on the established experience with, knowledge about, and prior research on the problem space that can be tested; decisions about the tentative answers can be generated deductively. These speculations need to be tested using creative methods to operationally define the variables, collect data in reasonably controlled situations, and interpret results against those results expected if the speculation were true and if the speculation were false. The *if/and/then* hypothetico-deductive nature of this inquiry is fundamental to high-level inquiries and plausible reasoning. The decision to support or not support the speculation should encourage the researcher to either move ahead with other inquiries utilizing the speculation and associated explanation of more acute and focused questions or seek a better, revised speculation and explanation to guide the design of further inquiries. Lawson expected researchers to outline applications of their results because educational inquiry is applied research. His criteria implied that any study needs to be judged in the context of the canonical knowledge, ongoing research agenda, and practical applications as well as the quality of the individual study. Furthermore, nothing in the quality criteria focused on the type of data (qualitative or quantitative); rather, they emphasized that the study is about rigorous testing of a proposed relationship and explanation.

Yore (2003) expressed concern about argument, evidence, and generalization of empirical (quantitative and qualitative) science and mathematics education research. He pointed out the research is as much about argument as it is about inquiry. Analytical and dialectical arguments are central to empirical approaches that:

> critically examine and evaluate the numerous and at times iterative transformations of evidence into explanations [to produce descriptive and causal claims].… Analytical arguments inductively or deductively form a set of premises to a conclusion [of the form]: If p then q, p therefore q. [While dialectical] arguments are those that occur during discussion of debate and involve reasoning with premises that are not evidently true. (Duschl & Ellenbogen, 1999, p. 1)

Unfortunately, some pseudoscientific research studies substitute rhetorical arguments involving eloquent oratorical presentations filled with abstract or invented terms, fuzzy thinking, and discursive techniques to impress and persuade an audience without much, if any, evidence for the knowledge claims.

Compelling arguments require clear connections amongst data, theoretical backings, warrants, evidence, claims, counterclaims, and rebuttals (Toulmin, 1958). Collecting data is easy, but not all data are evidence for a knowledge claim in science and literacy education research! Data and measurements collected reflect the theoretical foundations used to design the inquiry or flow for the problem context under consideration (Kelly & Lesh, 2000). Which of these data and measurements are evidence for a claim requires warranted, deductive interpretations flowing from the theoretical backings or inductive discernments utilizing grounded theory techniques (Yore, 2003). How observations, measurements, and information morph into evidence for a claim is central to all research approaches. Lester and Wiliam (2000) suggested that the logical relation of the information to the claim involved a classificatory relationship between data and question—a comparative relationship that considers competing and alternative claims—and a statistical relationship in which chance occurrence must be considered.

A switch from quantitative to qualitative research approaches requires a parallel switch in logic associated with the methods (Roberts, 1982). Qualitative research data collection and interpretation processes frequently utilize (a) abduction, to extract a pattern from data in a holistic manner or gestalt that cannot be derived from the component parts, or (b) induction, to systematically identify regularities, patterns, or trends across events or information sources and to formulate conjectures, assertions, rules, and tentative knowledge claims. Abduction may involve metaphorical reasoning that injects creativity and potential confusion if researchers and audience do not share the selected metaphor. Induction increases the semantic information (specific to general), while decreasing or eliminating possibilities depending on the interpreter’s prior knowledge and experience (Johnson-Laird, 1988). Inductive expansion of information and generalization involves uncertainty and no ultimate proof. Quantitative research data collection and interpretative processes normally utilize deductive or hypothetico-deductive reasoning in which predictions and data, or data derivatives, are compared utilizing direct comparisons, graphic techniques, or statistics. These deductive forms of plausible reasoning do not lead to absolute proof or rejection of the a priori speculations and associated hypothesis, model, or theory but rather lead to claims within specific, probabilistic ranges of uncertainty or confidence intervals. The identification of associations, assertions, and relationships are not enough; high-quality research must provide explanatory mechanisms and cause–effects link(s), which may involve post hoc, nonexperience considerations not part of the original research design (Phillips, 2006).

The utility and power of research results must consider parallel but appropriate qualitative attributes (dependability, credibility, believability, confirmability), quantitative attributes (reliability, validity, significance, objectivity), and general utility and application. Generalizability is likely the single biggest concern that consumers of literacy and science education research express. There are growing concerns within both quantitative and qualitative research communities about generalizability and methods to produce generalized results. State and federal governmental and private organizations have been formed to certify instructional practices and materials as meeting the Gold Standard based on much of *Evidence Matters* (Mosteller & Boruch, 2002). Phillips (2006) stated, “It is this volume that ought to have been the target of criticism and debate in scholarly symposia rather than the [2002] NRC report” (p. 21).

Munby (2003) voiced concern about rigor in science education research that is echoed in mathematics education (Simon, 2004). Once the decision that the central research question is worthy of investigation and the development of the problem space and available technologies and instrumentations are aligned with a qualitative or quantitative approach selected, the next most pressing issue is rigor—not blind obedience to mechanistic methodological procedures. Rigor does not ensure correctness! No research methods course or handbook can fully provide a fail-safe, step-by-step process for conducting an inquiry into all research questions and writing a compelling and informative report of the research. Any that does will provide an ineffective approach, which will lead to inefficient and unproductive engagement in research, but not authentic inquiry (Simon). Munby suggested that research inquiry, resulting argument, and report represent a tightly woven fabric of justifications and theoretical foundations (intertextuality) used to screen questions, select methods, frame data analysis, construct an explicit and coherent chain of reasoning, and supplement data, generate claims, and advance knowledge.

2.2 Deliberations and Outcomes

Within this controversial context, the *Gold Standard(s) of Quality Research in Science and Literacy Education Conference* (NSF Award #0437198) was proposed and planned to promote honest and open deliberations of the Gold Standard within the language and literacy, learning and instruction, statistics and measurement, and science education communities through an invited list of attendees composed of established researchers, junior faculty members, and graduate students from Asia, Australasia, Europe, and North America. The conditions of the conference invitation were an active interest in research and science literacy, and a willingness to participate in open considerations of quality issues. All conference delegates participated as presenters, copresenters, discussants, and general provocateurs.

The 47 participants were divided into 6 working groups that were led by one of the conference planning team members and recorded by one of the graduate assistants with the able assistance of academic support staff. The authors of this chapter were charged with monitoring and documenting deliberations across the groups. Working sessions were interspersed across the conference schedule and amongst the parallel paper presentations and workshops. Large-group sharing and focused deliberations were scheduled to establish intermediary results, and a final summary presentation of cross-group summaries was held to verify the authors’ assertions and the conference’s recommendations for future actions.

The results of the 2nd Island Conference identified procedural rigor and clarity, clarity of variables, mining and secondary analysis of existing databases and information resources, generalization of research results, research ethics, mixed-methods and other innovative approaches as issues to enhance research quality and improve the Gold Standard(s). A very large majority of the participants suggested that no single standard could enhance the quality of research for the variety of problems being investigated; therefore, the plural of *standard* will be used in the remainder of this chapter.

2.2.1 Rigor

Procedural rigor and clarity have become somewhat problematic with the diversity and flexibility of research approaches (Kilpatrick, 2001; Lester & Wiliam, 2000; Ragin, Nagel, & White, 2004; Simon, 2004). Greater attention to assumptions and underlying procedures are needed. Beliefs about the robustness of data interpretation do not replace considerate planning, preparation, and practice. Misguided design decisions and ill-prepared researchers cannot be neutralized with eloquent data analysis and reporting.

The applications of statistics packages and discourse and conversation analysis in research on literacy and science learning provide examples. Inexperienced researchers *plug* numerical data into statistical software and select an application without awareness of the underlying assumptions and type of data collected (numerical, ordinal, interval, ratio). Discourse analysis permits teachers and researchers to compare texts to be read by students as well as their written texts (e.g., a synthesis of different texts on a topic) in terms of propositional structure. Conversation analysis allows for temporal and sequential dissection of oral language to produce a “fine-grain analysis of moment-to-moment interaction and the sequences of linguistic/discourse actions that create meaning” (Graesser, Gernsbacher, & Goldman, 2003, p. 12). Thus, they can be invaluable tools for analysis of written and spoken discourse in literacy and science learning and also as a window into students’ representations of what they are learning at both the recall and mental model levels.

However, discourse and conversation analysis requires a careful and flexible use. Goldman and Wiley (2004) argued that such analyses permit a researcher to highlight learners' misconceptions but do not provide a quick solution to the problem of which instructional method is better for a specific subject. Many researchers are choosing to do discourse and conversation analysis without the necessary preparation, linguistic background, insights into language and demands of the target context, and realization of time demands required by proper application of this data analysis technique (Fang, 2006; Grimshaw, 2003; Halliday & Matthiessen, 2004). Others enter into computer-assisted analysis of text and video files—discourse analysis software (e.g., Atlas TI™, Nudist™, and XSight™) and video analysis systems (e.g., StudioCode™, Transanna™, and Videograph™)—without fully understanding the operator demands and coding requirements or serving an appropriate internship with an expert user. Furthermore, researchers must utilize explicit and transparent sampling procedures so as to select the appropriate data interpretation method(s) and units of analysis.

Data collection and interpretation must fit the problem, purpose, and future considerations. Researchers must realize that not all data are evidence and that effective and creative data collection does not happen by chance—they are the result of the well-prepared mind. Procedural awareness, expertise, and consistency must be considered at all phases of the inquiry; data identification, data collection, data storage, data retrieval, data interpretation, etc. are critical issues in high-quality research approaches. Rigor goes beyond mechanics to include underlying assumptions, insights arising from expert guidance, logic, and reasoning leading to justified arguments and evidence-based knowledge claims.

2.2.2 Clarity

Definitions and measures of variables, outcomes, and contextual features need to be considered in terms of the target audience for or end users of the research results. If the audience is simply a small collection of like-minded researchers and professors, then the definitions need to address the conventions and traditions of these academic communities. However, if the end users are politicians, bureaucrats, school administrators, teachers, and public stakeholders, then the definitions and measures selected must be understandable and have credibility with these communities. The target audience and end users may influence the selection such that the tests and measures have reasonable—not necessarily great—psychometric and theoretical foundations but are familiar and have credence with less academically savvy consumers. Furthermore, there are variables and constructs used in specific research communities that originate from distinctly different scholarly traditions and, when embedded in a cognitive science perspective, may have lost some of their distinctiveness. Ford and Yore (in press) stated:

> Metacognition, from the psychology tradition, involves the learner's awareness and management of personal learning; while reflection, from the progressive education tradition, involves contemplating what you have done or what you are doing; and critical thinking, from the philosophy tradition, involves rational inquiry into your thinking to improve your thinking.

Today, it is difficult to determine what researchers mean by these three constructs because the focus of learning has expanded to include performance and affective outcomes, as well as conceptual outcomes, and new research areas, such as self-regulated learning. The conference deliberations revealed that a construct central to this conference—science literacy—was not well accepted across the international and diverse communities represented (Yore et al., 2007).

Assessment and evaluation appear to be used differently by the various communities when considering individuals, groups, and systems. An example is the relationship between assessment *of* learning or accountability or large-scale assessment, on the one hand, and assessment *for* learning or the construction and use of assessment tools responsive to the instructional objectives of a classroom, on the other hand (Black & Wiliam, 1998). The enactment of NCLB has stressed the importance of the relationship of assessment to instructional practice. Recently, Gitomer and Duschl (2007) proposed a framework for designing coherent assessment systems with two related aims: providing information to policy makers (large-scale assessment) and supporting classroom learning in different subjects (in particular, science). As regards the relationship between the two types of assessment, a basic distinction is introduced between external and internal coherence. Assessment systems are externally coherent when they reflect theories and models of learning. For example, assessing conceptual change in the classroom should be consistent with a model of conceptual change. Two or more components of an assessment system—namely, large-scale and local assessment—are internally coherent if they share the same underlying view of students’ learning and instruction. Internal consistency may be obtained if teachers are informed about the results of large-scale assessment and, if they are able, helped to adapt or modify their teaching and assessment practices according to these results.

It was apparent that assessment involved documentation and data collection for local purposes and that evaluation involved assessment and judgment against explicit criteria and standards. Assessment empowers and informs actions, but evaluation did not need to have such close connections to learning and instruction. Therefore, clear and concise scoring rubrics and procedures need to be developed and verified that reflect the end use and users: learners, teachers, and administrators. The assessment techniques in literacy and science education are much more creative and diverse than the traditional objective test items. However, researchers need to clarify their techniques and illustrate the avoidance of cultural, linguistic, and content bias connections to and rationale for using them as proxies for literacy and science understanding.

2.2.3 Secondary Analysis, Synthesis, and Data Mining

The conference deliberations identified the inefficient use of education databases and information sources and the need to make better use of these expensive and time-demanding resources. The NRC report (US NRC, 2004) recommended making data and information files available to other researchers working in similar problem spaces. This will require collaborative efforts from funding agencies, research institutions, and researchers to provide support, develop methods and perfect procedures to share data in meaningful, uniform, and useable ways—similar to what was done in the Human Genome Project and many other biochemistry projects. Infrastructures, techniques, and security systems need to be developed to store, retrieve, and reuse these data via existing analysis techniques and new approaches. Structural equation modeling (SEM), hierarchical linear modeling (HLM), meta-analysis, metasynthesis, and secondary reanalysis of large datasets and collective sets of quantitative and qualitative data should be used to better explore relationships, disentangle confounded results, and produce generalized results. The use of computer-based analysis software and systems (e.g., Atlas TI, Nudist 6, Nvivo 7, XSight, StudioCode, Transanna, Videograph, etc.) and business systems software (e.g., enterprise architecture) can be used to discern useful patterns in large text-files, video-files, and demographic databases.

Other reanalyses might start with teasing out results from large-scale, general achievement tests (reading comprehension of various texts, combined science achievement) that are more focused on well-defined constructs (reading comprehension of informational science text, recall of science information, conceptual understanding of science) to more precisely explore reading's contribution to lower-level and high-level science achievement. Such reanalyses require security access to individual test items and raw item responses. Large-scale international, national, statewide, and province-wide tests (e.g., NAEP, TIMSS, PISA, Stanford 9, Iowa Tests of Educational Development, Iowa Tests of Basic Skills, etc.) are frequently utilized to make political statements related to the comparative quality and performance of education systems; unfortunately, these rich datasets are not used fully to clarify problems and inform solutions. Furthermore, secondary analyses and reanalyses can be used for model building and evaluation against theoretical frames and reality to confirm hypotheses and to explore potential relationships. Caution is needed to address confounding variables, lurking variables, and models that do not reflect reality.

2.2.4 Generalizations

An issue related to the Gold Standards is the production of generalizations across quantitative and qualitative studies that will address the need of politicians, bureaucrats, policy makers, and other decision makers for evidence-based claims that apply to a variety of educational contexts and settings. Many nonrandomized research claims are limited to the samples and contexts explored and cannot be generalized to broader populations and contexts like random samples and randomized controlled trials. Meta-analysis and secondary reanalysis are well-established techniques to determine a summary effect across collections of quantitative studies or data focused on similar treatments, outcomes, and methods. Less well-established secondary analyses of qualitative data and metasynthesis results have been developed in medical and health care research communities that could provide similar insights and generalized results in education research. These endeavors will be much more successful if researchers facilitate the synthesis and cross-study generalizations by clarifying their target constructs and by utilizing uniform or standardized approaches, anchors (e.g., questions, items, performances), common information sources or artifacts, explicit coding procedures, and clear categories to allow studies to be linked via common data points.

An informative illustration of generalizations across research studies utilizing meta-analysis, secondary reanalysis, and systematic interpretation of quantitative and qualitative research results can be found in the report of the National Literacy Panel on Language-Minority Children and Youth and searchable database studies included (August & Shanahan, 2006b). This report was notable because of its clarity, procedural rigor, and shared database. The central purpose of this project was “to identify, assess and synthesize research on the education of language-minority children and youth with respect to their attainment of literacy” (August & Shanahan, 2006a, p. 1). They explicitly outlined the research questions, their theoretical framework and target definitions, the procedures for conducting the review and synthesis (information sources, selection criteria, search procedures, studies identified, external verification, and analyses), and the generalizations asserted from the five working subcommittees. The transparency of the purpose, focus, procedures, and outcomes are essential to allow open and full evaluation of the results by the research communities, end users, and other stakeholders. A critique of this report illustrated the value of the informed public debate and evaluation and how alternative theoretical frames could influence the synthesis procedures and likely the outcomes (Grant, Wong, & Osterling, 2007).

2.2.5 Research Ethics: Policies, Procedures, and Practices

Research ethics is becoming a much more important, complex, and perplexing consideration in human-focused research and appears to limit the type of inquiries possible. Some jurisdictions have adopted a *one-size-fits-all* human subjects approval policy for medical, social, and natural sciences that consists of a single review panel, procedure, and process for all ethics approval applications. Such approaches apply high-risk assessments to all approvals ranging from drug trials to action research studies. Occasionally, the implementation of these policies and the review panels are sidetracked into risk management and into providing advice about research design that go beyond ethics, fundamental safety, and informed and voluntary participation.

The policies, procedures, and panels are reasonably well equipped to address traditional experimental designs—although true random sampling, where every participant has equal opportunity to be selected for the experimental and control groups, is almost impossible under fully informed and voluntary consent requirements.

Unfortunately, some research approaches, such as classroom-based design experiments and community-based studies, do not lend themselves to a priori approval procedures. Common research approaches to improve professional practice and implement instructional innovations—action research studies and teaching experiments (Kelly & Lesh, 2000)—require the flexibility to revise procedures, cancel planned actions, and collect unexpected information. These methods have inherent power over free choice, confidentiality, and conflict-of-interest issues. Other approaches—such as community-based and participatory research focused on social justice, indigenous peoples, and underserved and underrepresented participants—involve potential conflicts between the existing power structure and approval for controversial issues with potential negative outcomes for the authorities.

Many research ethics concerns have been based on anecdotal records of negative events involving a combination of legal, moral, and ethical considerations with little empirical exploration or documented resolutions (Pritchard, 2002). The *Journal of Empirical Research on Human Research Ethics* was established to address the lack of empirical evidence upon which to base research ethics policies, procedures, and practices appropriate to specific cultures, contexts, and research topics (Sieber, 2006). Furthermore, researchers have started to respond to research ethics policies and research ethics review panels that restrict authentic and fundamental inquiries into critical language and science education problems involving culture, language, and worldviews (see Anthony et al., Chap. 24). These actions have involved formal efforts to revise government policy, changes to local interpretation and implementation of policy, and efforts to educate review panel members about standards of quality research.

2.2.6 Mixed-methods Approaches

The conference deliberations addressed the need for mixed-methods or multimethod approaches. Research purists reject mixed-methods research; but others believe such approaches are needed, important, useful, and complement traditional approaches (Johnson & Onwuegbuzie, 2004). These approaches comingle quantitative and qualitative methods that move the overall design into the middle of the research continuum so as to more fully address the complex nature of research questions and to recognize development of the problem space (Phillips, 2006). The mingling of methods can be seen in problem spaces that require developmental inquiry before addressing cause–effect issues, in research programs as questions and approaches evolve from descriptive to causal forms, or in nested studies where relationships are established and then explored in depth to establish causal mechanisms and explanations.

Literacy and science education researchers use both quantitative and qualitative approaches in mixed-method research projects. They use qualitative methods

> to obtain information on meaning, affect, and culture, while quantitative methods are used to measure structural, contextual, and institutional features, [hybrid strategies involving]

> qualitative methods to construct typologies of case narratives from in-depth survey data and then use modal narratives as categories in quantitative analysis

and nested approaches involving large-group surveys with follow-up interviews of a random or purposeful subsample and large-group assessments with targeted performance tasks for selected students across achievement stratifications (Ragin et al., 2004, p. 14).

Mixed methods capture the strengths of both qualitative and quantitative approaches, more fully address the development and evolution of problem spaces and research questions, and recognize the availability of instrumentation and technologies for data collection and interpretation. The qualitative–quantitative debates missed the need to match research approach to the research question—instead, they modified the research question to fit a preferred approach—and the central issue about quality, rigorous inquiries. The either/or debates did not address the ontological and epistemological foundations of various research logics and the resulting pattern of argumentation; nor was it fully realized that the research approach must match the research question and purpose, not the reverse—modifying the research focus to fit the preferred research approach. Mixed-methods approaches are becoming more common in literacy and science education research (see Levin & Wagner, Chap. 11; Norton-Meier et al., Chap. 9).

2.3 Closing Remarks

What is next? We believe that the research communities must develop a framework of common understandings about research purposes, development of problem spaces, available technologies and instrumentations, designs, arguments, claims, and applications. A significant majority of the 2nd Island Conference participants believed that the one-size-fits-all Gold Standard, or RCT, can limit the problems/questions addressed and that quality research needs to consider both the progress and the alignment of problem and design. Deliberations frequently focused on the types of problems and research that are not endorsed by the Gold Standard—such as fundamental explorations, historical and philosophical inquiries, longitudinal studies—and other designs focused on the early descriptive issues. Therefore, the conference promoted standards that retain and respect developmental inquiry so as to define and clarify problem spaces and worthwhile issues and to apply new inquiry technologies and approaches. Research criteria and approaches differ across the international spectra; researchers should be aware of these differences in traditions, perspectives, and conventions. Research communities need to resist external intervention by funding agencies, parent organizations, and governments. Furthermore, research communities must address public awareness of education and the dissemination of results flowing from quality research to the variety of audiences (academics, politicians, bureaucrats, teachers, administrators, taxpayers, parents, etc.) who need these findings to make policy and decisions, enact effective curricula and instructional practices, and advocate for effective education and schools.

Serious enhancement of literacy and language, learning and instruction, statistics and measurement, and science education research must not consider just the issues of rigor, clarity, utilization of existing datasets, generalizations, ethic standards, and mixed methods but also the fundamental philosophy of knowledge issues. Quality research moves toward explanations of relationships, provides insights into the cognitive and pedagogical mechanisms, and allows generalizations and predictability across contexts. Unfortunately, fewer researchers consider the limitations flowing from the type of data collected, the ontological assumptions of their worldview, and the epistemological beliefs embedded in their methodologies that allow explanations and generalizations to be achieved.

The fundamental nature of data (e.g., nominal, ordinal, interval, ratio scale) is infrequently considered. Researchers need to revisit the underlying assumptions of the interpretation systems used and the precision of the results reported from these types of data. Degree of certainty of knowledge claims and assertions and the accuracy of measures are reflected in researchers' word choices (hedges) and in the significant figures of numerical results reported. Disregard of the underlying assumptions for parametric and nonparametric statistical procedures and the data requirements of the interpretative system selected are becoming too common in research courses and reports. Researchers' informed choices are critical to design decisions, data collection, compelling arguments, and substantiated claims and implications. Unless researchers are willing to consider these fundamental issues in honest and open discussion, there is little likelihood of overcoming public skepticism about education research and recognizing the potential social good that rigorous, appropriate, generalized results can make to society.

Brickhouse (2006), then editor of the journal *Science Education,* following attendance at the 2nd Island Conference stated:

> I would like to use this [90th] anniversary [of Science Education] as an opportunity to respond to some of the political pressures science education researchers face, particularly those in the United States, regarding the methods they use. I want to raise the question of whether these new criteria for quality research will lead to good research.... I call for a more rigorous, critical dialogue within our community about both methods and aims as a way to improve research in science education. (p. 1)

She focused her criteria for good research in terms of the ethical issues of accomplishing goals beneficial to individuals and society (worthiness, social justice, public awareness, persuasion). Her standards included, but were not limited to, the following:

1. Evaluation of learning speaks to important educational aims for science education.
2. There is a careful and honest description of who is and who is not benefiting in science education studies.
3. There is potential for influencing policy and practice. (p. 2)

Brickhouse closed her considerations with insights into how a single Gold Standard would negatively influence science education and decrease the diversity of research questions considered and inquiry methods utilized. She suggested that diversity in science education, like an ecosystem, is beneficial. These comments echoed the deliberations of the 2nd Island Conference about education, learning, and instruction as diverse problem spaces; the choice of research questions and methods depends on the development of understanding and prior research in the problem space and other potentially related spaces, and the availability of technologies and instrumentations related to the problem and embedded variables.

Phillips (2006), in a 2005 keynote address to the European Association for Research on Learning and Instruction, cautioned that the noble endeavor of education research could come to a standstill, replaced by ghettoes of isolated research findings and personal opinions of questionable value and validity if the research communities do not engage in serious deliberations to ensure quality research. The general education research community has not converged toward some reasonable and rational position. Phillips outlined the research positions along a bipolar continuum in which the extreme left pole was represented by the postmodernist and poststructuralist worldviews and the extreme right pole was represented by the traditional reductivist and positivist worldviews.

Yore et al. (2004) believed that discussions about the nature of science had converged onto these extreme positions without recognizing the middle-of-the-road modern (postpositivist) worldview. They posited that each worldview subsumes specific ontological assumptions about metaphysics and fundamental elements of knowledge building and epistemological beliefs about how knowledge is constructed. The traditional worldview assumes a realist, absolutist position; the modern worldview assumes a naïve realist, evaluativist position; and the postmodern worldview assumes an idealist, relativist/multiplist position.

Phillips (2006) contended that educational research oscillates between *physics envy* and *all opinions are equally valid*; both of these extreme positions represent a legendary exact science, which likely did not exist, and a sociopolitical stance that confounds power, social justice, political action, and knowledge construction. Unfortunately, neither position appears to assess methodological strengths and weaknesses, pragmatics, alignment of problem, canonical knowledge, and available instrumentation. The extreme ontological assumptions about human behavior and learning—in which it follows well-defined and precise relationships or in which it is chaotic and idiosyncratic—lead to these extreme poles of the epistemic research continuum. If one assumes that human behavior and learning have some regularity within defined contextual situations, then a modern, postpositivist position and the associated variety of methods that move inquiries toward generalized claims and explanations are possible and justified.

In 2007, the National Association for Research in Science Teaching (NARST) devoted a president-sponsored symposium and a research committee-sponsored symposium to the consideration of “A Critical Look at Science Education as a Field of Research” and “The Gold Standard of Science Education Research—Does One Size

Fit All Problems?". Lawson (2007), a panelist in the symposium, applied an interpretative framework to classify the epistemology used in articles published in the volumes of the *Journal of Research in Science Teaching* (JRST, 1965–2005 at 5-year intervals) to get a sense of the state of science education research published in the NARST-sponsored journal. The epistemic framework identified three levels: Level 1—involves observational activities to determine *what* and utilizes induction to formulate descriptions; Level 2—involves struggling with observations to propose causal hypotheses and tentative explanations that can be tested using hypothetico-deductive reasoning; and Level 3—involves the construction of umbrella theories that integrate ideas and provide general explanations, which in turn can be deductively used to design ways to test these ideas. This framework assumed a progression of scientific inquiry metalanguage (ontological and epistemic terms) across the levels (predictions, hypotheses, theories, etc.).

Computer-generated word counts of epistemic terms (individual and combinations) in text-files of all articles published in the target years revealed a steadily increasing use of *theory* from 1965 (18.0%) to 2005 (86.7%), less dramatic and inconsistent increases for *hypothesis* (32.0 → 48.9%) and *prediction* (8.0 → 31.1%), and relatively few articles containing combinations of two or three of these epistemic terms (8.0 → 46.7%; 4.0 → 26.6%; 0.0 → 17.8%; 2.0 → 17.8%, respectively). An in-depth analysis of a random sample of the 2005 articles "revealed that most authors were generating and testing hypotheses and/or theories (presumably guided by Level 2 or Level 3 epistemology) albeit in a largely implicit and sometimes haphazard way" (Lawson, 2007, p. 1). Lawson reminded researchers of Novak's 1963 classic guidance in the first issue of *JRST*, in which Novak stated, "The purpose of research in science education, nevertheless is the same as that in other fields of science, e.g., to advance the conceptual systems which have been developed to explain events in the universe about us" (p. 3). This axiom of intent has been the central focus of many editorials in science and mathematics education research journals (Brickhouse, 2006; Lawson, 2005; Munby, 2003; Simon, 2004; Yore, 2003). Furthermore, the advancement of conceptual systems cannot be judged by a single contribution of a researcher; rather, it must consider the researcher's research program, inquiry agenda, and progression of knowledge building.

The 2nd Island Conference planning committee left the conference realizing the importance of diverse scholars talking to one another rather than talking past each other. The research debates did not recognize the full membership of cognitive sciences (learning sciences) to include historical and philosophical inquiries and tried to privilege one approach over all others. The Gold Standard RCTs may establish relationships between variables, but explanations of cause–effect mechanisms and generalizing claims can be achieved by other methods (Phillips, 2006). The planning committee explicitly mentioned that advocacies for evidence-based decisions do not fully recognize the sociopolitical context or fabric of education and, therefore, the other cultural values and ideological grounds that influence the policies or decisions. The Gold Standard may promote an unexpected positive outcome in which the research communities deliberate issues not easily explored by empirical traditions, stimulate improved practices with the full range of quantitative–qualitative methods, and enable researchers to explore theoretical issues that do not have immediate and direct applications. This new momentum may help researchers to become more concerned about the user groups (teachers, curriculum developers, etc.) and affected groups (learners, etc.). Researchers must expand their responsibilities to consider impact factors, risk–gain relations, and cost–benefit analyses so that learners can be helped with minimum risk and harm. Furthermore, interventions must be doable by normal teachers in real classrooms. These issues are considered and elaborated in the remainder of this book.

References

August, D., & Shanahan, T. (2006a). Introduction and methodology. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the national literacy panel on language-minority children and youth* (pp. 1–42). Mahwah, NJ: Lawrence Erlbaum.

August, D., & Shanahan, T. (Eds.) (2006b). Developing literacy in second-language learners: Report of the national literacy panel on language-minority children and youth. Mahwah, NJ: Lawrence Erlbaum.

Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Brickhouse, N. W. (2006). Celebrating 90 years of science education: Reflections on the gold standard and ways of promoting good research [Editorial]. *Science Education*, *90*(1), 1–7.

Duschl, R. A., & Ellenbogen, K. (1999). Middle school science students' dialogic argumentation. *Proceedings of the 2nd international conference of the European Science Education Research Association "Research in science education: Past, present, and future"*, Kiel, Germany. Retrieved from http://www.ipn.uni-kiel.de/projekte/esera/book/regf.htm

Education Sciences Reform Act of 2002. Pub. L. No. 107–279, 116 Stat. 1940. (2002).

Fang, Z. (2006). The language demands of science reading in middle school. *International Journal of Science Education, 28*(5), 491–520.

Ford, C. L., & Yore, L. D. (in press). Toward convergence of metacognition, reflection, and critical thinking: Illustrations from natural and social sciences teacher education and classroom practice. In A. Zohar & J. Dori (Eds.), *Metacognition in science education: Trends in current research*. Dordrecht, The Netherlands: Springer.

Gitomer, D. H., & Duschl, R. A. (2007). Establishing multilevel coherence in assessment. In P. A. Moss (Ed.), *Evidence and decision making* (Vol. 106, pp. 288–320). Malden, MA: Blackwell.

Goldman, S. R., & Wiley, J. (2004). Discourse analysis: Written text. In N. K. Duke & M. H. Mallette (Eds.), *Literacy research methodologies* (pp. 62–91). New York: Guilford.

Graesser, A. C., Gernsbacher, M. A., & Goldman, S. R. (Eds.) (2003). *Handbook of discourse processes*. Mahwah, NJ: Lawrence Erlbaum.

Grant, R. A., Wong, S. D., & Osterling, J. P. (2007). Developing literacy in second-language learners: Critique from a heteroglossic, sociocultural, and multidimensional framework [Essay book review]. *Reading Research Quarterly*, *42*(4), 598–609.

Grimshaw, A. D. (2003). Genre, registers, and contexts of discourse. In A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.), *Handbook of discourse processes* (pp. 25–82). Mahwah, NJ: Lawrence Erlbaum.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar* (3rd edn.). London: Arnold.

Johnson-Laird, P. N. (1988). *The computer and the mind: An introduction to cognitive science*. Cambridge, MA: Harvard University Press.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26.

Kelly, A. E., & Lesh, R. A. (Eds.) (2000). *Handbook of research design in mathematics and science education*. Mahwah, NJ: Lawrence Erlbaum.

Kilpatrick, J. (2001). Where's the evidence? *Journal for Research in Mathematics Education*, *32*(4), 421–427.

Lawson, A. E. (2005). Conducting high quality educational research [Editorial]. *International Journal of Science & Mathematics Education*, *3*(1), 1–5.

Lawson, A. E. (2007, March). *How "scientific" is science education research?* Paper presented at the annual meeting of the National Association for Research in Science Teaching, New Orleans, LA.

Lester, F. K., & Wiliam, D. (2000). The evidential basis for knowledge claims in mathematics education research. *Journal for Research in Mathematics Education*, *31*(2), 132–137.

Mosteller, F., & Boruch, R. (Eds.) (2002). *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings Institution Press.

Munby, H. (2003). Educational research as disciplined inquiry: Examining the facets of rigor in our work [Guest editorial]. *Science Education*, *87*(2), 153–160.

No Child Left Behind Act of 2001. Pub. L. No. 107–110, 115 Stat. 1425. (2002).

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, *87*(2), 224–240.

Phillips, D. C. (2006). A guide for the perplexed: Scientific educational research, methodolatry, and the gold versus platinum standards. *Educational Research Review*, *1*(1), 15–26.

Pritchard, I. A. (2002). Travelers and trolls: Practitioner research and institutional review boards. *Educational Researcher*, *31*(3), 3–13.

Ragin, C. C., Nagel, J., & White, P. (2004). Workshop on scientific foundations of qualitative research. Available from http://www.nsf.gov/pubs/2004/nsf04219/start.htm

Roberts, D. A. (1982). The place of qualitative research in science education. *Journal of Research in Science Teaching*, *19*(4), 277–292.

Sieber, J. E. (2006). The evolution of best ethical practices in human research. *Journal of Empirical Research on Human Research Ethics*, *1*(1), 1–2.

Simon, M. A. (2004). Raising issues of quality in mathematics education research. *Journal for Research in Mathematics Education*, *35*(3), 157–163.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

United States Institute of Education Sciences. (n.d.). *What Works Clearinghouse overview: Who we are*. Retrieved June 13, 2008, from http://ies.ed.gov/ncee/wwc/overview/

United States National Research Council. (2000). *How people learn: Brain, mind, experience, and school—Expanded edition*. Committee on Developments in the Science of Learning. J. D. Bransford, A. L. Brown, & R. R. Cocking (Eds.). *Commission on Behavioral and Social Sciences and Education*. Washington, DC: The National Academies Press.

United States National Research Council. (2002). *Scientific research in education*. Committee on Scientific Principles for Education Research. R. J. Shavelson & L. Towne (Eds.). *Center for Education, Division of Behavioral and Social Sciences and Education*. Washington, DC: The National Academies Press.

United States National Research Council. (2004). *Advancing scientific research in education*. Committee on Research in Education. L. Towne, L. L. Wise, & T. M. Winters (Eds.). *Center for Education, Division of Behavioral and Social Sciences and Education.* Washington, DC: The National Academies Press.

United States National Research Council. (2005). *How students learn: Science in the classroom*. Committee on How People Learn, A Targeted Report for Teachers. M. S. Donovan & J. D. Bransford (Eds.). *Division of Behavioral and Social Sciences and Education*. Washington, DC: The National Academies Press.

United States National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Committee on Science Learning, Kindergarten through Eighth

Grade. R. A. Duschl, H. A. Schweingruber, & A. W. Shouse (Eds.). Board on Science Education, Center for Education, *Division of Behavioral and Social Sciences and Education*. Washington, DC: The National Academies Press.

Yore, L. D. (2003). Quality science and mathematics education research: Considerations of argument, evidence and generalizability [Guest editorial]. *School Science & Mathematics*, *103*(1), 1–7.

Yore, L. D. (2008). Science literacy for all students: Language, culture, and knowledge about nature and naturally occurring events. *L1—Educational Studies of Language & Literacy*, *8*(1), 5–21. Retrieved from http://l1.publication-archive.com/show?repository = 1&article = 213

Yore, L. D., Hand, B., & Florence, M. K. (2004). Scientists’ views of science, models of writing, and science writing practices. *Journal of Research in Science Teaching*, *41*(4), 338–369.

Yore, L. D., Pimm, D., & Tuan, H.-L. (2007). The literacy component of mathematical and scientific literacy. *International Journal of Science & Mathematics Education*, *5*(4), 559–589.

Yore, L. D., & Treagust, D. F. (2006). Current realities and future possibilities: Language and science literacy—empowering research and informing instruction. *International Journal of Science Education*, *28*(2/3), 291–314.

Chapter 3
Research and Practice: A Complex Relationship?

Robin Millar and Jonathan Osborne

The past decade has seen fundamental questions about the nature and quality of educational research, and its relationship to practice and policy, placed prominently on the agenda in many countries. In the United Kingdom, the 1996 Teacher Training Agency lecture by David Hargreaves, then of the University of Cambridge, is widely seen as having played a key role in setting the agenda and influencing the direction of the ensuing debate. In his lecture, Hargreaves (1996) asked if teaching could be regarded as a research-based profession and concluded that it could not. This he attributed largely to the nature and quality of the outcomes of educational research: “Given the huge amounts of educational research conducted over the past fifty years or more, there are few areas which have yielded a corpus of research evidence regarded as scientifically sound and as a worthwhile resource to guide professional action” (p. 2).

As to how the situation might be improved, Hargreaves (1996) drew on a comparison between educational research and medical research. First, he observed:

> In medicine, as in the natural sciences, research has a broadly cumulative character. Research projects seek explicitly to build on earlier research – by confirming or falsifying it, by extending or refining it, by replacing it with better evidence or theory, and so on. Most educational research is, by contrast, non-cumulative.... A few small-scale investigations of an issue produce inconclusive and contestable findings of little practical relevance. (p. 2)

Second, he pointed to “a very sharp difference in the way the two professions approach applied research. Much medical research is not itself basic research ... but a type of applied research which gathers evidence about what works in what circumstances” (p. 2).

This vision of the role of research—as providing evidence about what works—can be seen as part of a wider movement for evidence-based practice across many areas of professional work (Davies, Nutley, & Smith, 2000). The idea first became

R. Millar
University of York

J. Osborne
Stanford University

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

well established in medicine in the early 1990s, and its influence has subsequently extended to other aspects of health care (such as nursing), welfare policy, criminal justice, social policy, and social work—as well as to education. The aspiration to create a rather different and more effective form of educational research—perhaps even a science of education—is not, of course, new. In the second half of the 19th century, Bain (as cited in Nisbet, 1980) was making a very similar case for educational research methods more closely modeled on those used in the natural sciences. And there have been several waves of enthusiasm for this kind of emphasis in the intervening 130 years, each countered by opponents highlighting the epistemological differences between the natural and the social sciences. None of the previous waves of enthusiasm has been long-lived, not least because the *scientific* methods advocated were not shown to be capable of delivering the outcomes or practical improvements claimed for them.

On the one hand, it is difficult, as a science educator, not to feel somewhat dissatisfied with what educational research on the learning and teaching of science has been able to offer to practitioners and policy makers. Lijnse (2000) wrote of his frustration when, as a newly appointed lecturer in physics education, he assumed that the research literature would offer him some practical guidance when faced with the challenge of devising a course in introductory quantum mechanics, only to discover that “hardly any such help appeared to be available” (p. 309). Despite the large body of research on the learning and teaching of topics like Newtonian mechanics, electric circuit theory, the particulate model of matter, and so on, we cannot recommend to a teacher preparing to teach any of these topics at secondary school a teaching intervention for which we could honestly say there was clear and compelling research evidence of its efficacy. If we deem this not to be feasible for epistemological reasons, it would seem to raise significant questions about the practical usefulness of this kind of research. On the other hand, if we judge it to be feasible, albeit difficult, why have so few studies, relatively, tried to produce such evidence?

In this chapter, we explore these issues and ask: What can we reasonably expect research to provide by way of warrants for acting in specific ways, for using one specific teaching intervention rather than another for a given topic? This we do by drawing on examples that have been undertaken in the domain of science education. First, we suggest that some key terms in the discussion of the research–practice interface are not as clearly delineated as they need to be—and that some issues about the interrelationships between knowledge arising from research and the practical actions of teachers in the classroom or laboratory need to be teased out rather more carefully if we are to talk, write, and indeed think clearly about them. In particular, we think it is important to distinguish the possible contribution of research to the *design* of instruction from its possible contribution to the *warranting* of actions and decisions. Then we look in a little more depth at the second of these; namely, the role of research in providing a warrant for action in the specific context of science education, that is to say, in the context of trying to improve learners’ understanding of a specific science topic or idea, or their capability in executing a specific science-related skill.

3.1 The Relationships Between Research and Practice

Perhaps not surprisingly, the phrase *evidence-based practice in education* has met with a mixed response from many educators. Often this seems to be based on a perception that the advocates of evidence-based practice want to make research evidence the sole basis for practical decisions, leaving little or no role for professional judgments. This, however, is not how evidence-based medicine is seen by its advocates. Sackett, Rosenberg, Gray, Haynes, and Richardson (1996) characterized it as:

> [t]he conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence-based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research. (p. 71)

They are especially careful to make clear that this is not "'cookbook' medicine" (p. 72); professional expertise and judgment have an essential role in making decisions about the treatment of individual patients.

Nonetheless, the reaction of some educators and teachers to the term evidence-based practice had led Hargreaves by 1999 to argue that

> [w]e should, perhaps, not use the term 'evidence-based'.... Decisions are not based on research evidence alone.... To avoid any implication that teachers or educational policy makers should not, in making decisions, take account of (i) the quality and strength of the research evidence and (ii) the contextual factors relating to that decision, we should, I suggest, speak of evidence-informed, not evidence-based, policy or practice. (1999, p. 246)

Some have suggested even softer language for the research–practice relationship, preferring terms such as "evidence-influenced" or "evidence-aware" (Davies et al., 2000, p. 11). The relationship between research and practice is, however, more complex than these labels immediately suggest. There is an important distinction to be made between the influence of research on the *design* of a teaching intervention and the extent to which research provides *evidence of the effectiveness* of a teaching intervention. There are appreciably more examples of science education research drawing attention to the need for improvement in specific aspects of practice than of research demonstrating clearly and convincingly how to teach a topic more effectively. As a result, there are rather more instances of practice that could reasonably claim to be research evidence-informed—that is, that attempts to address the weaknesses identified by research—than of practice whose use is warranted by research findings.

Decisions about the detail of practice are at the heart of every teacher's professional work. Teachers routinely have to make decisions and choices about how they are going to go about the teaching of a topic that is on the course or syllabus they are following. Figure 3.1 (Millar, Leach, Osborne, & Ratcliffe, 2006) tries to summarize the influences on decisions of this sort; it is quite general and hence applicable to any teaching in any subject area. The first choice is between using an existing teaching intervention (perhaps the way you taught the topic last year or a sequence suggested by a published course) or developing a new teaching intervention (possibly drawing on your past practice, or colleagues' advice, or

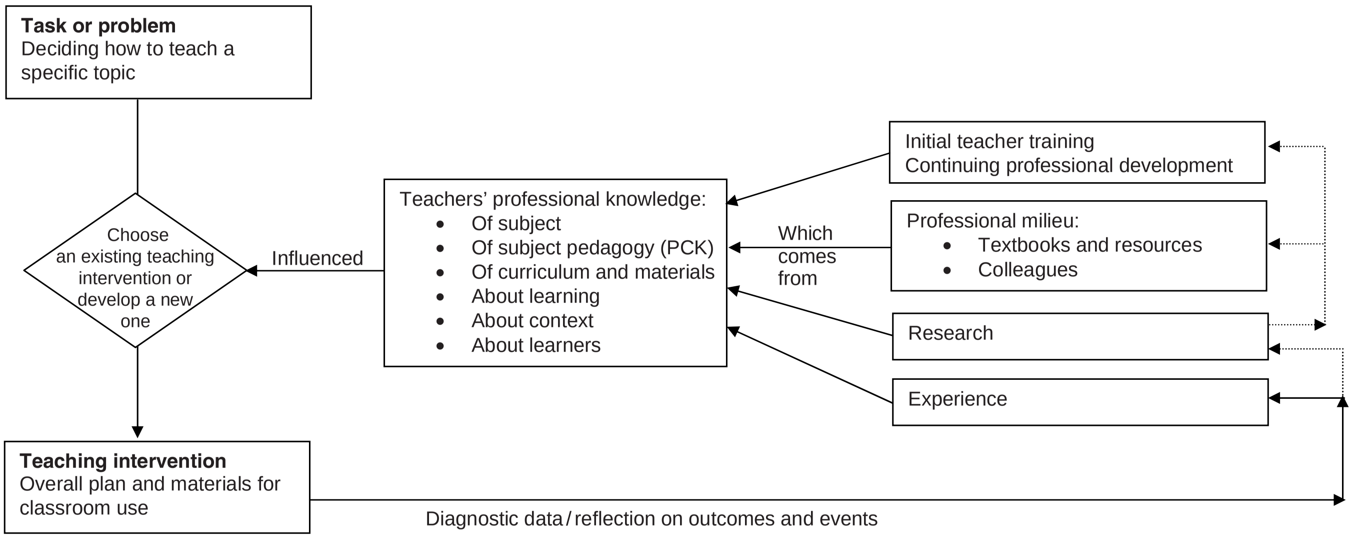

Task or problem
Deciding how to teach a specific topic
Choose an existing teaching intervention or develop a new one
Influenced
Teachers' professional knowledge:
• Of subject
• Of subject pedagogy (PCK)
• Of curriculum and materials
• About learning
• About context
• About learners
Which comes from
Initial teacher training
Continuing professional development
Professional milieu:
• Textbooks and resources
• Colleagues
Research
Experience
Teaching intervention
Overall plan and materials for classroom use
Diagnostic data / reflection on outcomes and events

Fig. 3.1 Influences on teachers' decision making

published materials). This decision is influenced by the professional knowledge the teacher brings to bear on the situation—as are the many more detailed choices and decisions that are required whichever of these two basic options is followed. The relevant professional knowledge includes knowledge of the subject matter to be taught, knowledge about how to teach it (often called pedagogic content knowledge), knowledge of the context of the teaching and of the curriculum being followed, knowledge of the available teaching resources for the topic, and knowledge and views about the characteristics of learners like those in the group to be taught. Much of this knowledge is tacit and not consciously articulated. All of it has been acquired by the teacher from a variety of sources, including their initial teacher education and any continuing professional development (CPD) they have experienced during their teaching career, colleagues, published teaching materials and resources, reported research findings, and their own experience—both informal and gained through any more systematic efforts to evaluate their previous practice (through, e.g., action research).

Research, if available and if the teacher is aware of the findings, is only one of several potential influences on teachers' professional knowledge. Findings and ideas from research contribute to teachers' professional knowledge in both direct and indirect ways. The indirect ones are via their influence on the providers of initial teacher education and CPD, on the authors of textbooks and other teaching materials, and on colleagues with whom the teacher interacts. The ways in which—indeed even the fact that—research informs these influences may or may not be recognized by teachers planning their teaching of a topic. Finally, once all the decisions have been taken and a sequence of lessons is actually taught, the teacher will collect data on outcomes, certainly informally and perhaps more formally (e.g., through a piece of action research), which then feeds into his or her experience and informs future decisions about how to teach this topic. So the label *research evidence-informed* covers a complex web of influences. There are many ways in which research might influence decisions and choices about practice.

The task facing a science education researcher who wishes to develop a teaching intervention informed by research on the teaching and learning of a given science topic is also complex. Often research identifies weaknesses in the outcomes of current practice but is silent (or speculative) about how they might be remedied. And often research draws conclusions in terms of *general* claims about learning or teaching, which must then be implemented with the *specific* content of the topic. The resulting intervention will be a composite of features that are seen by its designer(s) as necessary consequences of taking a particular research finding seriously, alongside features that are seen as almost entirely contingent. The former should not be altered in implementing the intervention whereas the latter could be altered without any serious consequences. Almost never is a commentary provided to tell a potential user which features are of which sort or to explain the designer's decisions about how to apply the more general principles that informed the design of the intervention to *this particular* content. The issues involved in designing a teaching intervention that could be termed research evidence-informed are discussed in some detail by Scott, Leach, Hind, and Lewis (2006).

To call a teaching intervention research evidence-based is, if we follow the usage in medicine of the term *evidence-based*, to make a rather different kind of claim. It is essentially the claim that there is evidence, from research (i.e., collected appropriately and systematically and open to public scrutiny), that indicates the intervention is effective in achieving its aims—perhaps indeed that it is significantly more effective than other alternatives. The teaching intervention whose effectiveness is being evaluated could, in principle, be informed by research or by practitioner knowledge, or a mixture of the two—though this does not have any particular significance for the underlying argument here. The role of research is simply to provide evidence of the effectiveness of an intervention whatever the influences on its design. In the remainder of this chapter, we focus on issues surrounding the feasibility of evidence-based practice in science education. In this sense, our central question is: To what extent can research provide a convincing warrant for acting in one way rather than another in the science classroom—for choosing (or recommending that someone else should choose) to use one teaching intervention or approach rather than another?

3.2 Can Research Provide a Warrant for Specific Practices in Science Education?

Rather than explore the relationship between research and practice in science education in general terms, we approach the central issues through a number of specific examples. We begin with three real examples of teaching interventions or approaches for which the warrant for their use provided by the research findings to support their use is strong. The three examples we have chosen are: wait-time when posing questions to a student or a class; the routine use of formative assessment; and the Cognitive Acceleration through Science Education project materials. The findings from the first two are not solely applicable to science education, and the findings of the third have applicability beyond the domain of science education as well. For each case, we discuss the quality of the evidence for action that the research provides. Typically, researchers look to issues of validity (the extent to which the measurement provides a proper measure of the effects of any action), reliability (the extent to which the measurement is a consistent measure of what it purports to measure), and replicability (the evidence that an effect has been shown in more than one study). Research that meets these three criteria is generally considered to provide high-quality evidence to justify its claims.

3.2.1 Example 1: Wait-time

This research looked at the length of time that teachers of science are prepared to wait after addressing a question to their students—defined as wait-time. The work was conducted by Rowe (1974) in the early 1970s and drew on a sample of 221 audio transcripts of science lessons plus another 100 sent in by teachers. The data revealed that the average time that teachers of children of all ages left for their charges to respond to their questions was less than 1 second (0.9 s on average), and that wait-time rarely exceeded 3 seconds. Ninety-six teachers were then trained to extend their wait-times and avoid mimicking student answers. The surprising finding was that student responses increased from an average of 8 to 27 words, the mean number of unsolicited responses per lesson went up from 5 to 17, failure to respond diminished from 7 to 1 occasion per lesson, the quality of students' reasoning improved, and the range and type of student contributions quadrupled. In addition, there was a notable effect on teachers. They became more flexible in the types of response they gave, asking fewer but better questions that sought less to test pupils' knowledge and more to support, develop, and probe their understanding. The change in pupil response also led teachers to modify their expectations of what pupils can achieve.

The first point to make of this research is that most of the variables (time before response, length of student utterance, number of unsolicited responses, and failure to respond) are easily defined and measured. There can be little disagreement about their validity which, in turn, would undermine the reliability of the study. The size and range of the sample would also suggest that the findings are a product of a definite effect rather than an unreliable measurement of abnormal cases. Doubt might be cast on the effect of this practice on student reasoning as the validity of measurements of student reasoning is more open to question. However, such findings do fit with common sense as, in normal dialogue, few individuals respond to any cognitively demanding question in less than 1 second. What is more, this research has been replicated by Tobin (1986), who found essentially identical effects with mathematics teachers and language arts teachers. In his study, 10 classes were randomly assigned to a group that received feedback and assistance to maintain an average teacher wait-time of between 3 and 5 seconds during a sequence of mathematics lessons. A control group of 10 teachers maintained a regular wait-time and received placebo feedback. The study was also replicated in a sequence of language arts lessons. In both cases, the use by an average teacher of extended wait-times in whole-class instructional settings was associated with higher mathematics achievement and improvements in the quality of teacher and student discourse. Thus, measured against the three criteria of validity, reliability, and replication, we would argue that the evidence provided by this research as a warrant for well-specified practice is clear, unambiguous, and difficult to contest.

3.2.2 Example 2: Formative Assessment

Formative assessment refers to assessment practices that reveal the differences between the actual state of the student and the desired state, which then inform and influence the subsequent pedagogic actions of the teacher. The nature of the evidence for the value of formative assessment differs from that for wait-time. It comes from many disparate studies of a range of practices that fall under the broad heading of formative assessment. A more systematic examination of this evidence can be found in two major reviews (Black, 1993; Black & Wiliam, 1998a), which synthesized a large body of potentially relevant literature. The review findings were then summarized in a small pamphlet (Black & Wiliam, 1998b) written specifically for teachers.

Reviews of this nature attempt to achieve validity by stating explicitly the methods used to identify the relevant research and the criteria used for selection or rejection (these criteria were not clearly articulated in the 1998 Black and Wiliam review, though they did provide a list of 76 journals that were searched). Clarity about methods makes replication, at least in principle, a possibility. Black and Wiliam (1998a) identified 250 articles from 681 publications that appeared initially to be relevant. From these articles, they began their review by using 8 examples that were selected for pragmatic reasons to make a strong case for the practice of formative assessment and to identify issues for further exploration. A case could be made, therefore, that they were selective in the evidence they presented with little discussion of any research evidence that showed a null or even negative effect for the practice of formative assessment (though negative aspects are discussed later in their paper). Another criticism is the authors' own acknowledgment that the number of quantitative studies with adequate rigor is "of the order of 20 at most" (p. 53).

What the review does attempt is to draw on research evidence to provide an explanatory mechanism for why certain practices might be effective (e.g., the provision of feedback on performance relative to task-specific criteria but not marks). Even then, the authors are forced to acknowledge that some will see this approach as flawed "because any one tactic will vary in its effect within the holistic context in which it works" (Black & Wiliam, 1998a, p. 38). Another weakness is that a subset of the research on formative assessment emanates from the mastery learning literature where "it is impossible to establish from the research reports which features were implemented … let alone which were effective" (p. 43).

Nevertheless, their review would appear to tell a consistent (reliable) story. Whatever doubts might be cast on the validity or reliability of an individual study, the replication of the findings in diverse studies points emphatically to the view that the regular use of formative assessment leads to significant learning gains for students, particularly for students of lower ability. The fact that many (far more than the eight specifically used to introduce the review) studies exist—albeit differing in their sample size, specific intervention, and outcome measures—does broadly address the requirements of replicability. Indeed, such syntheses are now becoming an important aspect of the work of centers, such as the Evidence for Policy and Practice Information and Co-ordinating Centre based at the Institute of Education in London. The strength of the warrant here for the use of formative assessment lies in the range and extent of the studies, in the synthesis of the studies, and in the interpretation extracted by Black and Wiliam (1998a). Their analysis led them to conclude that "the research reported here shows conclusively that formative assessment does improve learning … [and that the principles that underlie the evidence of] substantial improvements are robust" (p. xx). Nevertheless, even such a body of research is qualified with lack of clarity about the context and the specific actions of the teachers such that there is, in the words of the authors, “no single royal road” (p. xx).

3.2.3 Example 3: Cognitive Acceleration through Science Education

The Cognitive Acceleration through Science Education (CASE) teaching intervention is a set of lessons based on Piagetian ideas and designed to develop children’s cognitive capabilities. It is the outcome of a program of over 30 years of research—a feature that makes it exceptional compared to most educational research. In the mid 1970s, Shayer, Küchemann, and Wylam (1976) first estimated the Piagetian levels of thinking of the UK school population using a representative sample of students from ages 11 to 16. This work enabled them to develop instruments for measuring the scientific reasoning ability of children; these instruments were extensively tested for their reliability. It might be noted, however, that the validity of these measures of children’s cognitive capabilities has been contested (Metz, 1995).

Arguing that the imperative for research was to develop interventions that would accelerate children’s thinking to higher levels, Shayer and Adey developed CASE, an extensive intervention program for 11–14-year-old children consisting of a detailed and tightly specified set of activities to be used by the teacher at fortnightly intervals (Adey, Shayer, & Yates, 1989). These activities specifically addressed logico-mathematical reasoning skills over a period of 2 years.

A study investigating the effects of an intervention based on these materials used a quasi-experimental design. The researchers worked with a set of schools and teachers who were introduced to the theoretical background and aims of the project and specifically trained in the use of teaching materials. The intervention was found to have significant positive effects on student learning in science (Adey & Shayer, 1990) that had increased 1 year later (Shayer & Adey, 1992a). Two years later, the intervention had a significant positive effect on these students’ performance in national examinations in science (Fig. 3.2) and, more surprisingly, in English and mathematics for the boys though not for the girls (Shayer & Adey, 1992b). A study of a later cohort also showed similar gains for girls (Shayer & Adey, 1993).

The schools for this work had to be specifically recruited; and the results are based on a sample of 11 schools, which are compared with 16 control schools. Hence, this study, though using an experimental design, does not have all the attributes of an RCT. To do that would have required random selection of schools and teachers to implement the intervention, and it is very difficult to see how schools would have agreed to such a request. Finally, the sample of treatment schools might be considered too small by some.

The findings of this study have, however, been replicated through the annual collection of datasets, which show that schools using the CASE intervention achieve higher learning gains than would be predicted from a knowledge of the

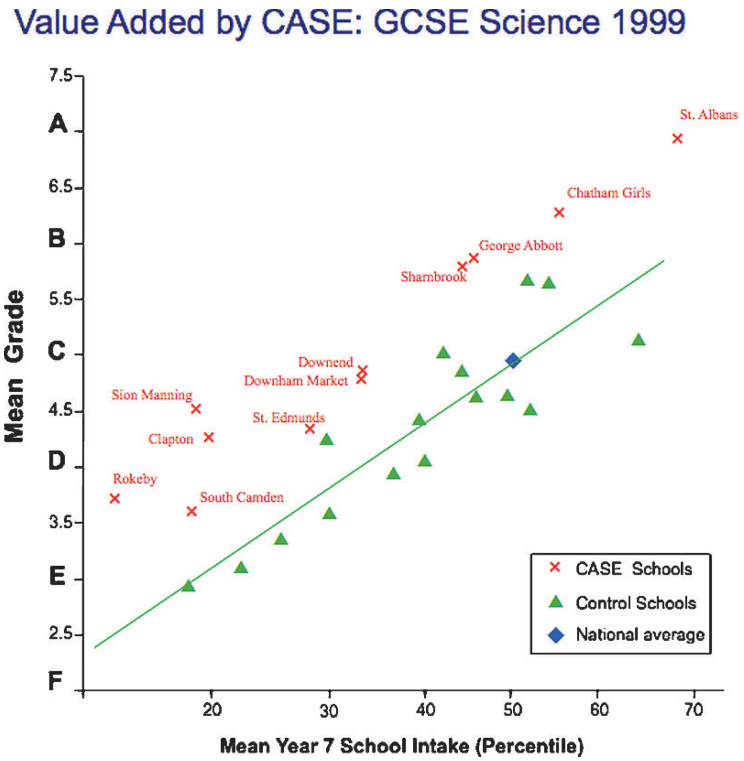


Fig. 3.2 Results for students taking national science examinations (GCSE) at age 16 for CASE schools and control schools (Shayer & Adey, 1993)

characteristics of their student intake. Overall, the outcome of this work is possibly the most conclusive set of evidence that exists for the effects of a general intervention program within the domain of research in science education in that it consists of (a) a tightly defined and scripted intervention whose materials were later made available by a commercial publisher; (b) an extensive program of professional development to assist teachers in understanding the nature of the program and its implementation; (c) a clearly defined set of outcome measures (national tests) that are commonly used to assess student performance; and (d) a set of extensive instruments (whose reliability had been tested) developed from a previous program of research for measuring students' ability to reason. Whilst it is possible to argue for a more rigorous research design, it is difficult to see how more could have been achieved without considerably greater investment in human and material resources. As ever with experimental interventions, questions can be raised about the causal mechanism underlying the outcome obtained (Leo & Galloway, 1996) and the validity of the instruments used to measure outcomes. Nevertheless, the rigor of the study, its well-defined background, and the consistent picture of the outcomes would suggest that this must stand as a benchmark of the kind of evidence that science education research can achieve.

3.3 Strength of Warrant and Impact on Practice

Looking at the three examples discussed above as a group, we can say that all have had some impact on practice in certain schools—but none could be said to have become a routine part of normal educational practice. Indeed, none has been adopted in the majority of UK schools. The advocates of experimental methods of evaluation, centering on the use where possible of RCTs, might argue that such modest impact is largely a consequence of the relatively weak evidence that nontrials' designs can provide; for example, the intervention was implemented by teachers who were predisposed to the methods or in specially chosen schools rather than a group that had been randomly selected. If this is so, then what kind of evidence or warrant for practice might such an RCT offer and in what way might it be an improvement?

To probe the issues around the use of trials a little further, we now consider what might be involved in developing, and evaluating by a trial, a better way to teach some key points in basic electric circuit theory—a topic that has been shown consistently by research in many countries to be difficult for many students (Shipstone, 1985). We will discuss an imaginary study simply because we do not know of any RCT of a teaching intervention for this or any other specific science topic.

The starting point is the problem the new intervention is intended to address. Many research studies have pointed to the difficulty students have with the idea of electric current as a quantity that has the same value at all points round a series circuit and that is not used up as it flows through the components in the circuit (Shipstone, 1985). As this idea is fundamental to electric circuit theory and much of what we would want to teach subsequently depends upon it, a teaching intervention that can be shown to be a more effective way of enabling more students to grasp this core idea would be seen by many science educators as an important advance in both knowledge and practice.

The first point we would make is that, although research consistently shows that this topic poses learning difficulties for many students, this does not in itself tell us how teaching might be made more effective. A teacher or a curriculum developer with this aim in mind would need to identify some general teaching strategy (or strategies) on which an intervention might be based. Undertaking this task would require a careful consideration of the extensive, and sometime conflicting, literature on how students learn. Candidates might be: explicit elicitation of students' ideas followed by specific activities to cause cognitive conflict with nonscientific ideas (Posner, Strike, Hewson, & Gertzog, 1982), a strong emphasis on models and analogies (Treagust, Harrison, & Venville, 1996), an explicitly dialogic form of teaching allowing students many opportunities to articulate their current ideas (Mercer, 1996), the use of computer animations and simulations to provide memorable images to support learning (Linn, Davis, & Bell, 2003), or a practically based approach making considerable use of tasks with a predict–observe–explain structure (White & Gunstone, 1992). In practice, an intervention might draw on more than one of these. Whichever is (or are) chosen, there are then many more fine-grained decisions to be taken to work out the details of applying the chosen *general* strategy to *this particular* body of content.

Let us imagine that these choices and decisions have been made in the light of the available knowledge and experience (see Fig. 3.1). We would now have a proposed intervention that its designer believes can improve the teaching of this topic. How might this claim be evaluated in a rigorous manner? First, we should note that it would not be sensible to contemplate a large-scale, rigorously designed research evaluation of any new intervention or approach until there was clear evidence from less tightly designed studies to suggest that it works. To do otherwise would require too large an investment of resources in a speculative trial. Preliminary evidence might come from its implementation by the person who had designed it; for example, a teacher seeing improvements with her or his own classes as a result of having adopted a new way of teaching something. The natural next step would be to see if others not involved in the development of the intervention also observed similar improvements. Let us assume that this has been done for our imaginary intervention and so a large-scale RCT evaluation is both justified and planned.

The next question is: What should be used as the control or to what should the intervention be compared? In the case of an intervention designed to improve the teaching of a standard curriculum topic, the salient comparison is usually with current practice. This means, however, that the intervention experienced by the control group is not uniform across the sample of classes involved and the likely consequence is that intended outcomes will vary somewhat (perhaps considerably) within the control group. This is not, however, an insuperable problem although it may mean that a larger sample size is needed to gain statistically significant evidence of improvement if the effect size is modest, as it is for most educational innovations.

It is important to recognize that, even if a teaching intervention is specified in detail, its actual implementation by different teachers working with different classes and in different schools will necessarily be different. Anyone who has even taught the same course to two classes in the same school year will know that the same lesson is always different each time you teach it. The reason is that teaching is necessarily responsive to the learners—and students' reactions, responses, and interventions can lead any given lesson in quite widely differing directions. This means that there is an issue of specifying exactly what the intervention being evaluated is. We are probably best to regard it as the package of materials (which could include oral elements, such as training programs, and written lesson materials and guidance on their intended use) rather than as the actions that ensue in the classroom—though we might want to check that the implemented intervention bore enough resemblance to the intended intervention for it to be regarded as an instance of the intended intervention in action. If it passed this test, the advocates of trials would then regard the inevitable differences between implementations of the intervention as the very factors that randomization is designed to deal with.

The trial would then report on whether or not the intervention leads to better outcomes *on average*—in more situations, or for more students, more of the time.

It is worth noting in passing—and a commonly neglected point in most arguments for RCTs—that the first time anything is taught is not a good occasion to evaluate the practice. Few teachers have a clear concept of the goals and aims of any new teaching intervention until they have been through it at least once and have evaluated which practices or components of the strategy matter most in implementing it. Thus, it is unlikely that a first implementation provides useful evidence of the effectiveness of any intervention.

The next practical challenge in setting up a trial is to develop an outcome measure that deals evenhandedly with both the experimental and control groups. That is to say, we need an instrument (perhaps a test of some kind) that measures students' knowledge and understanding at the end of the intervention (perhaps both immediately after it and some weeks or months later to detect more lasting learning) in a way that is fair to both those who followed the intervention and those in the control group. This is not an easy matter to achieve. Often a new teaching intervention does not simply seek to teach more effectively the same thing as has previously been taught but to alter the teaching emphasis by giving more time to certain aspects of a topic than others, or to aim for understanding rather than recall, or to emphasize understanding of how ideas are applied in practice, or whatever. If so, it may then be difficult to devise an outcome measure that treats the experimental and control groups equitably. It may be necessary to identify a common core that applies to both despite other differences. One striking feature of science education is that no standard or commonly agreed outcome measures exist for any major topic. The need for them is demonstrated by the way in which some published assessment tools, like the Force Concept Inventory (FCI, Hestenes, Wells, & Swackhamer, 1992), have been used by researchers. But such tools are also subject to quite significant criticism—not least because the construction of any test instrument requires choices about the learning outcomes of most worth, which inevitably involve values. Also, instruments like FCI have not been subjected to the kind of rigorous scrutiny of factorial structure and content validity that would be standard practice for measures of attainment or learning outcome in other subject areas, in particular the kinds of standard measures used by psychologists. So using a trusted, off-the-shelf, outcome measure is not an option for the science education researcher.

Nevertheless, let us imagine that this challenge has been met and an outcome measure has been developed that all those involved agree treats both experimental and control groups fairly. The next issues to be faced relate to the structure and organization of schools. Most teaching interventions are intended to be used with classes, not with individual learners. So evaluation needs to be carried out also with classes—as the learning outcomes may be very different for class groups and for individual learners. Rather than randomly allocating individual students to the experimental and control groups, we have to allocate classes in a *cluster-randomized* design. This does not raise any additional technical issues, though it does mean that the number of students involved in a trial in order that statistically significant differences can be detected is appreciably larger and is, as a corollary, the cost.

The allocation of classes to experimental or control groups raises another issue: the teacher's preference. Some teachers will be attracted by the intervention that is to be evaluated whereas others may prefer the way they currently teach the topic. These preferences may be very soundly based in an awareness of their personal strengths and weaknesses and in the kind of teaching approach they can manage well. In an individually randomized design, it would be easy to allocate students randomly to teachers, all of whom were teaching in their preferred manner. In a cluster-randomized design, it is more likely that classes will be allocated to either the experimental or control group—and the teachers told that they are using the intervention or (in line with the decision reached earlier in this case) to teach in their normal manner. If the intervention is one that its designers hope, in the long run, will be voluntarily chosen by teachers in the light of the findings of the evaluation trial, then an RCT in which teachers are allocated to either the experimental intervention or control group does not answer our question. We do not find out what learning outcomes can be obtained by teachers opting for the intervention but rather what outcomes are achieved when they are told to use it. Hence, the outcome is only likely to be of interest for interventions that will, if shown to be effective, be made mandatory at a national or regional level.

Let us now jump ahead to the point where a trial has been conducted and students' scores on the agreed outcome measure are being compared. One result—the one the designers hope for—is that the average score of students in the experimental group is higher by a statistically significant amount than that of students in the control group. What conclusions can we draw from such a result? Certainly, we may be able to conclude that *this particular intervention* is likely to lead to higher scores on *this outcome measure* than the alternatives it has been compared to. We cannot, however, draw well-supported inferences as to *why* this has occurred. We definitely cannot conclude unequivocally that it is a consequence of the underlying strategy on which the design of the intervention was based. For it could be that its success is due instead to more specific, critical details of the lesson sequence and activities (Viennot, 2003). Without further trials of the same underlying strategy applied to other content areas, we can only base our view of this strategy on professional judgment and not on empirical evidence. Yet the capacity of trials to provide such evidence is precisely what supporters base their advocacy upon. Also, any teacher who wanted to argue against using the intervention could rationally challenge the validity of the outcome measure—as its appropriateness rests ultimately on a professional judgment rather than being entailed by objective evidence.

What if the outcome of the trial is less positive, that is, there is no significant difference in average score between the experimental and control groups? Can we conclude that the intervention is not worth using? Again, the situation is less clear-cut. Perhaps the underlying strategy on which the experimental intervention is based is sound, but it has not been implemented well with this particular content. Or perhaps the lack of commitment to the intervention by some of those required to implement it has cancelled out any gains achieved by those who did favor it.

Once again, any conclusions that we draw about the features of the intervention that led to its lack of success are based on professional judgment and are not logically entailed by the evidence.

We have discussed this imaginary trial in detail in order to raise some critical questions about the logic of RCTs, in particular about the extent to which they really can provide clear evidence of what works. The principal claim of RCT advocates is that they provide strong evidence of causal effects. Our argument is that, even if an intervention can be defined clearly and implemented reasonably consistently and an adequate outcome measure agreed, it can at best provide strong evidence of a causal effect that is so circumscribed as to be of little practical value to anyone. It can only tell us if there is a measurable causal effect of the *specific intervention trialed* (in its entirety) in a situation where the teacher using it has no choice about the matter. In most real situations, this is not close to what we are really interested in. We want to know if an intervention based on some clear design principles works when it is implemented by teachers who believe in it.

3.4 The Response of Teachers to Research Evidence

The fact that RCTs cannot offer unequivocal evidence about the causal mechanism at work in the intervention is, we would argue, a fundamental reason why RCTs cannot provide a Gold Standard for evidence on educational questions. All they can do is establish a covariation between a given intervention and a particular measure of its effects. The underlying causal mechanism remains uncertain or speculative because a correlation is not, of itself, evidence of a causal connection; even if it is agreed that a causal link is likely, the RCT provides no information on the mechanism. From a philosophical view, the argument is then susceptible to the Duhem–Quine thesis of underdetermination: there are in principle an infinite number of hypotheses that could explain any observed outcome (see, e.g., Curd & Cover, 1998, pp. 255–408). An awareness of this, albeit often tacit, has a significant impact on how people respond to evidence of an educational effect. As Koslowski (1996) has argued:

> Even when the presence of co-variation is buttressed by the presence of plausible mechanism, causation is still not certain. The likelihood of a causal relation also depends on consideration of alternative accounts. Explanations are not evaluated in isolation; they are judged in the context of rival accounts. … One finds an explanation increasingly compelling to the extent that alternative causes of the effect have been ruled out or controlled for. Conversely, to the extent that alternative causes remain viable, one is less certain that the target cause was at work in the situation in question. (p. xx)

In social contexts, there are always many competing hypotheses. An array of alternative hypotheses is possible because of the multiple factors at play and the multivariate relationships in the complex process called *schooling* or *teaching–learning*. In the case of wait-time, for instance, why should such a simple modification lead to such a notable change? Teachers will commonly argue, and did in our work (Millar et al., 2006, ch. 8), that the context of a research study was exceptional and that the same intervention would not work with their school, or their students, or that the teachers were specially chosen. Hence, similar effects would be unlikely to be attained in other classrooms. RCTs are, of course, designed to address such arguments; but the lack of any clear evidence of the causal role of specific features of the intervention enables such doubts to be generated. And, without additional evidence to the contrary, they are logically sufficient to justify a decision not to adopt the intervention.

Likewise, the CASE intervention requires teachers to use a program of activities; many of which, with a focus on logico-mathematical operations, are different from normal practice and alien to the teaching style of many teachers. The underlying causal hypothesis is neither simple nor transparent, leading teachers to doubt whether their attempts to implement such an approach would lead to similar outcomes in their context. In addition, in this case, the positive effects are not immediately evident but are seen in the long term.

In both examples, rival accounts may be generated by teachers and others who see either the nature of the students or the disposition of the teachers, or both, as offering an alternative explanatory account of the findings. Chinn and Brewer (1993) found seven distinct forms of response to unexpected data—only one of which is to accept the data and change your theories and, one would hope, your actions. The other six responses involve discounting the data in various ways in order to protect the preexisting theory. The more entrenched the belief (as is the case with ideas that have been reinforced through daily use), the less likely are theory-changing responses.

Why might this be so? In part, it is born of a natural reluctance to change—the unwillingness of individuals to accept that their standard practice can be improved sufficiently to justify the effort and the personal challenge of changing it. Change, as Claxton (1988) pointed out, involves threats to any individual's sense of competence (as new techniques are unfamiliar and untested), control (as the outcomes and reactions of the students are uncertain), and confidence (as there is no base of previous experience on which to rely). In the case of wait-time, the new practice would require teachers to interrupt a personal style of questioning that is habituated in their practice. Consciously breaking the habit of a lifetime is both difficult and discomforting, and this encourages the reinterpretation of research evidence to justify continuing with current practices.

In addition, the "crucible of practice" (US National Research Council, 2002, p. 25) that research informs is not a homogeneous entity. Rather, the mix of ingredients that constitutes effective practice varies from one classroom to another and from one teacher to another, making their adoption and replication, as seen by the teacher, problematic. Without sufficient details of the mechanisms by which the intervention functions, teachers can always point to the contingent nature of the activity of teaching (i.e., the necessity of responding to the specifics of context and student response) and argue that warrants offered by research are not generalizable to their particular context. In contrast, in medical situations where treatment and response are understood as phenomena governed by laws and theories of the natural sciences rather than the social sciences, the perception of contingency—and indeed the influence of contingency—is less, if not absent.

Even if unequivocal evidence could be produced, it is one thing to change teachers' knowledge and values but another to change teachers' pedagogic practices. Indeed, there is a lack of clarity in the literature as to whether it is best to begin by seeking to change teachers' knowledge and values (Harland & Kinder, 1997; Putnam & Borko, 2000) or by changing their practice (Guskey, 2002). For even where there has been considerable success in communicating to teachers ideas arising from research—such as with formative assessment where Black and Wiliam and their collaborators have developed a set of pamphlets (Black, Harrison, Lee, Marshall, & Wiliam, 2002; Black & Wiliam, 1998b) and a book (Black, Harrison, Lee, Marshall, & Wiliam, 2003)—it would be rash to claim that these ideas have led to a general change in teachers' practice.

Any teacher, like any other individual, will only change his or her professional practice if they have some level of dissatisfaction with existing practice. If, for instance, there is no dissatisfaction with existing practice (Posner et al., 1982) and if there are doubts either about the plausibility of the research evidence or the potential value of the new suggested practice, then there is little incentive to change (Fullan, 2001). Moreover, change in practice is rarely reducible to a single action but, rather, requires a set of actions for which teachers need an understanding of their theoretical rationale and a course of training in their use. However, there is a body of evidence (Adey, Landau, Hewitt, & Hewitt, 2003; Joyce & Showers, 2002; Loucks-Horsley, Hewson, Love, & Stiles, 1998), including our own (Bartholomew, Osborne, & Ratcliffe, 2004), that points to a range of outcomes when teachers are given a course of professional development in a new practice. Not all teachers change; indeed, some do not change at all. In addition, teachers tend to modify an innovation to fit more traditional patterns of instruction with which they are familiar (Ogborn, 2002; Osborne, Duschl, & Fairbrother, 2002; Ratcliffe, Hanley, & Osborne, 2007). Where interventions are simple and straightforward, this may not be true; but many interventions are complex and rely on their interpretation and translation by the teacher who has the responsibility of enacting them in the classroom.

Current knowledge, derived from research, would suggest that a more complex view of professional learning is required to bring about substantial and sustained change (Adey et al., 2003; Bell & Gilbert, 1996; Fullan, 2001; Hoban, 2002; Loucks-Horsley et al., 1998; Spillane, 1999). These researchers see teaching as a dynamic relationship with students and with teachers where change involves uncertainty, room for reflection in order to understand the emerging patterns of change, a sense of purpose that fosters the desire to change, a community to share experiences, opportunities for action to test what works or does not work in their classrooms, conceptual inputs to extend teachers' knowledge and experience, and sufficient time to adjust to the changes made. Embedding a new approach in the teaching of science as a normative practice requires changes in pedagogy to be adopted not just by individuals in isolation but, rather, by whole school departments working collaboratively. Thus, the requirement for good or excellent quality in the warrants produced by educational research is at best a necessary condition for change but not a sufficient condition.

This is not to say that research cannot offer signposts to the reflective teacher. However, teachers, compared to doctors, rarely have such well-defined objectives as the remediation of the patient's illness or, at the least, its alleviation. In the context of science teaching, a goal might be to teach students that the size of the electric current is the same at any point in a simple electric circuit. However, unlike the medical patient—where it is relatively straightforward to measure if they have a temperature or (with modern imaging technology) bone degeneration, liver malfunction, tumors, etc.—whether a student has understood a scientific idea is not amenable to a simple one-off test. Test items commonly lack validity or reliability, or both.

The work of the Evidence Based Practice in Science Education (EPSE) project on diagnostic tests found that with five items designed to assess students' knowledge of simple electric circuits, only 14% of a sample (*N* = 173) 14-year-old children were able to answer all five correctly (Millar & Hames, 2002). What judgment is a teacher to make in such a context? That the educational experience has failed? Or alternatively, should he or she accept the view that these students plus a further 10% who answered four out of five correctly have achieved understanding? Or that the 43% who gave correct answers to three or more of the five questions have understanding? Such a decision is a matter of values and not one for which research can provide a single or unequivocal answer.

Moreover, teachers, at any given time, are often working not with a single goal but a multiplicity of goals of both a short-term and long-term nature. In teaching about conservation of current in a circuit, particularly if the lesson is practically based, in addition to the conceptual goal, the teacher may want the students to learn how to connect elementary circuits, how to insert an ammeter with the correct polarity, or how to take reliable measurements by repeating them several times. In addition, the teacher may have affective goals in mind of offering an engaging experience or of challenging students' common preconceptions. In addition, longer-term goals might be to develop a model of electric current, to help students see that an electric current is simply a means of transferring energy from one location to another or that accurate measurement is a core feature of scientific practice. How are such goals to be measured? And, more importantly, how can the design of any RCT take proper account of such multiple goals in a way that leads to compelling evidence of a causal effect—as choices about the relative weighting of the learning outcomes in any outcome measure depend on the values of those who make them and may not be universally shared?

Our aim in this chapter has been not to argue that RCTs have no value in educational settings. Rather, our view is that they have a contribution to make; but it is a relatively modest one as their range of applicability is seriously circumscribed—to the evaluation of interventions that can be implemented relatively uniformly, for which the desired outcomes can be measured in ways that command broad consensus, and where the longer-term aim, if the intervention is shown to work, is to make it mandatory. If these conditions are not all met, then an RCT, however well designed, will fail to persuade practitioners to change and will only persuade policy makers if they are already minded to make the change. In our view, RCTs cannot, for the reasons we have rehearsed in more detail earlier, provide irrefutable evidence for choosing a specific intervention to teach any given science topic or skill. Yet we believe, as do many science educators, that it is in improving the effectiveness of teaching of the subject matter of science that the greatest improvements in science education lie. As a community of research and practice, we need a stronger emphasis on an engineering approach of research and development—the research evidence-informed development of new approaches followed by their careful evaluation. But we should be wary of placing undue confidence in experimental studies and RCTs as the means of showing which of them work and of persuading teachers in greater numbers to adopt those that appear to work. To allow one method to become hegemonic—to be seen as a Gold Standard to which all others should aspire—would be to fail to recognize the limitations of RCTs and the contribution that other methods can make to the research evidence base.

References

Adey, P. S., Landau, N., Hewitt, G., & Hewitt, J. (2003). *The professional development of teachers: Practice and theory*. Dordrecht, The Netherlands: Kluwer.

Adey, P. S., & Shayer, M. (1990). Accelerating the development of formal thinking in middle and high school students. *Journal of Research in Science Teaching, 27*(3), 267–285.

Adey, P. S., Shayer, M., & Yates, C. (1989). *Thinking science: The curriculum materials of the CASE project*. London: Thomas Nelson & Sons.

Bartholomew, H., Osborne, J., & Ratcliffe, M. (2004). Teaching students "ideas-about-science": Five dimensions of effective practice. *Science Education, 88*(5), 655–682.

Bell, B., & Gilbert, J. (1996). *Teacher development: A model from science education*. London: Falmer.

Black, P. J. (1993). Formative and summative assessment by teachers. *Studies in Science Education, 21*(1), 49–97.

Black, P. J., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2002). *Working inside the black box*. London: King's College London.

Black, P. J., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning*. Maidenhead, Berkshire, UK: Open University Press.

Black, P. J., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Black, P. J., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. London: King's College.

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63*(1), 1–49.

Claxton, G. (1988). *Live and learn: An introduction to the psychology of growth and change in everyday life*. Milton Keynes, Buckinghamshire, UK: Open University Press.

Curd, M., & Cover, J. A. (Eds.) (1998). *Philosophy of science: The central issues*. New York: W.W. Norton.

Davies, H. T. O., Nutley, S. M., & Smith, P. C. (Eds.). (2000). *What works? Evidence-based policy and practice in public services*. Bristol, Avon, UK: The Policy Press.

Fullan, M. (2001). *The new meaning of educational change* (2nd edn.). London: Cassell.

Guskey, T. R. (2002). Professional development and teacher change. *Teachers and Teaching, 8*(3), 381–391.

Hargreaves, D. H. (1996). *Teaching as a research based profession: Possibilities and prospects* [The Teacher Training Agency Annual Lecture]. London: The Teacher Training Agency.

Hargreaves, D. H. (1999). Revitalising educational research: Lessons from the past and proposals for the future. *Cambridge Journal of Education*, *29*(2), 239–249.

Harland, J., & Kinder, K. (1997). Teachers' continuing professional development: Framing a model of outcomes. *British Journal of In-Service Education*, *23*(1), 71–84.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*(3), 141–158.

Hoban, G. (2002). *Teacher learning for educational change*. Buckingham, UK: Open University Press.

Joyce, B., & Showers, B. (2002). *Student achievement through staff development* (3rd edn.). White Plains, NY: Longman.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: MIT Press.

Leo, E. L., & Galloway, D. (1996). Conceptual links between cognitive acceleration through science education and motivational style: A critique of Adey and Shayer. *International Journal of Science Education*, *18*(1), 35–49.

Lijnse, P. L. (2000). Didactics of science: The forgotten dimension in science education research. In R. Millar, J. Leach, & J. Osborne (Eds.), *Improving science education: The contribution of research* (pp. 308–326). Buckingham, UK: Open University Press.

Linn, M. C., Davis, E. A., & Bell, P. (2003). *Internet environments for science education*. Mahwah, NJ: Lawrence Erlbaum.

Loucks-Horsley, S., Hewson, P., Love, N., & Stiles, K. E. (1998). *Designing professional development for teachers of science and mathematics*. Thousand Oaks, CA: Corwin.

Mercer, N. (1996). The quality of talk in children's collaborative activity in the classroom. *Learning & Instruction*, *6*(4), 359–377.

Metz, K. E. (1995). Reassessment of developmental constraints on children's science instruction. *Review of Educational Research*, *65*(2), 93–127.

Millar, R., & Hames, V. (2002). EPSE Project 1: Using diagnostic assessment to improve science teaching and learning. *School Science Review*, *84*(307), 21–24.

Millar, R., Leach, J., Osborne, J., & Ratcliffe, M. (2006). *Improving subject teaching: Lessons from research in science education*. London: Routledge.

Nisbet, J. (1980). Educational research: The state of the art. In W. B. Dockrell & D. Hamilton (Eds.), *Rethinking educational research* (pp. 1–10). London: Hodder & Stoughton.

Ogborn, J. (2002). Ownership and transformation: Teachers using curriculum innovations. *Physics Education*, *37*(2), 142–146.

Osborne, J., Duschl, R. A., & Fairbrother, R. (2002). *Breaking the mould? Teaching science for public understanding*. London: Nuffield Foundation. Available from http://www.nuffieldfoundation.org/fileLibrary/pdf/teachingspu01.pdf

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*(2), 211–227.

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher*, *29*(1), 4–15.

Ratcliffe, M., Hanley, P., & Osborne, J. (2007). *Evaluation of twenty-first century science GCSE Strand 3: The teaching of twenty-first century science GCSE, and teachers' and students' views of the course*. Southampton, Hampshire, UK: University of Southampton.

Rowe, M. B. (1974). Wait-time and rewards as instructional variables, their influence on language, logic, and fate control: Part one - wait-time. *Journal of Research in Science Teaching*, *11*(2), 81–94.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *British Medical Journal, 312*(7023), 71–72.

Scott, P., Leach, J., Hind, A., & Lewis, J. (2006). Designing research evidence-informed teaching sequences. In R. Millar, J. Leach, J. Osborne, & M. Ratcliffe (Eds.), *Improving subject teaching: Lessons from research in science education* (pp. 60–78). London: Routledge.

Shayer, M., & Adey, P. S. (1992a). Accelerating the development of formal thinking in middle and high school students II: Postproject effects on science achievement. *Journal of Research in Science Teaching*, *29*(1), 81–92.

Shayer, M., & Adey, P. S. (1992b). Accelerating the development of formal thinking in middle and high school students III: Testing the permanency of effects. *Journal of Research in Science Teaching*, *29*(10), 1101–1115.

Shayer, M., & Adey, P. S. (1993). Accelerating the development of formal thinking in middle and high school students IV: Three years after a two-year intervention. *Journal of Research in Science Teaching*, *30*(4), 351–366.

Shayer, M., Küchemann, D. E., & Wylam, H. (1976). The distribution of Piagetian stages of thinking in the British middle and secondary school children. *British Journal of Educational Psychology*, *46*, 164–173.

Shipstone, D. M. (1985). Electricity in simple circuits. In R. Driver, E. Guesne, & A. Tiberghien (Eds.), *Children's ideas in science*. Milton Keynes, Buckinghamshire, UK: Open University Press.

Spillane, J. P. (1999). External reform initiatives and teachers' efforts to reconstruct their practice: The mediating role of teachers' zones of enactment. *Journal of Curriculum Studies*, *31*(2), 143–175.

Tobin, K. (1986). Effects of teacher wait time on discourse characteristics in mathematics and language arts classes. *American Educational Research Journal*, *23*(2), 191–200.

Treagust, D. F., Harrison, A. G., & Venville, G. J. (1996). Using an analogical teaching approach to engender conceptual change. *International Journal of Science Education*, *18*(2), 213–229.

United States National Research Council. (2002). *Scientific research in education*. Committee on Scientific Principles for Education Research. R. J. Shavelson & L. Towne (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Viennot, L. (2003). Relating research in didactics and actual teaching practice: Impact and virtues of critical details. In D. Psillos, P. Kariotoglou, V. Tselfes, E. Hatzikraniotis, G. Fassoulopoulos, & M. Kallery (Eds.), *Science education research in the knowledge-based society*. Dordrecht, The Netherlands: Kluwer.

White, R. T., & Gunstone, R. (1992). *Probing understanding*. London: Falmer.

Chapter 4
Moving Beyond the Gold Standard: Epistemological and Ontological Considerations of Research in Science Literacy

Donna E. Alvermann and Christine A. Mallozzi

What goes around, comes around is a maxim that seemingly applies more and more often to the current debate in the United States over what constitutes the *scientific* label in education research. At the time of writing this chapter, the No Child Left Behind Act of 2001 (NCLB, 2002), which calls for, among other things, scientifically based reading methods and materials, is up for reauthorization. With it have come challenges to the federal government's role in legislating what counts as scientifically valid research through the Education Sciences Reform Act of 2002 (ESRA, 2002). The provisions of this law, at least as enacted, have effectively equated scientifically valid research to randomized controlled trials (RCT)—or what is commonly known as the Gold Standard in education research circles. Prior to the passage of ESRA, the National Research Council (NRC) in its publication *Scientific Research in Education* had criticized the proposed bill for attempting to mandate "a list of 'valid' scientific methods … [a list which] erroneously assumes that science is mechanistic and thus can be prescribed" (US NRC, 2002, p. 130). More recently, groups—such as the Knowledge Alliance (a Washington, DC, firm representing a mix of researchers and research and development centers), the American Educational Research Association, and the Software & Information Industry Association—have voiced their opposition to ESRA's definition of scientifically valid research. Perhaps not surprisingly, language in a recent House of Representatives draft of a bill to reauthorize NCLB would omit references to randomized studies. In its place, the proposal would define scientifically valid research as being "rigorous, systematic, and objective … [and] appropriate to the methods used" (Viadero, 2007, The Gold Standard section, para 6).

Whether or not this attempt to move away from the *one-size-fits-all* Gold Standard makes its way into reauthorized legislation is yet to be seen. In the interim (and for the purpose of this chapter), we intend to explore how methodological border crossings among researchers in language, literacy, and science education can enrich curricular conversations about teaching and learning in science classrooms. To chart this terrain, we begin by providing a cursory view of

D.E. Alvermann and C.A. Mallozzi
University of Georgia

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

the relation of language and literacy to science teaching and learning. We then offer a window into our thinking on how Gold Standard policies have sanctioned certain kinds of research and curricular development while discouraging other types, thus potentially narrowing the range of information about science literacy practices that teachers have at their disposal. To address this situation, we examine the assumptions underlying five different dimensions or styles of doing research for the express purpose of looking for ways to open up, at least partially, what we view as an overly restrictive, one-size-fits-all approach to science literacy research in the United States.

4.1 Language and Literacy in Relation to Science Teaching and Learning

In a guest editorial coauthored by researchers representing the fields of language, literacy, and science education, Hand et al. (2003) laid out an argument for viewing language as both a means and an end in science literacy:

> Language is an integral part of science and science literacy—language is a means of doing science and to constructing science understandings; language is also an end in that it is used to communicate inquiries, procedures, and science understandings to other people so that they can make informed decisions and take informed actions. (p. 608)

This means–end argument, which points to the integral relation of language to science literacy, is also reflected in Norris and Phillips' (2003) efforts to connect literacy and science in an epistemological sense:

> Literacy in the fundamental sense is based upon the same epistemology that underlies science and that the reasoning required to comprehend, interpret, analyze, and criticize any text resembles in its major features the reasoning at the heart of all of science. When it is also recognized that science is in part constituted by text and the resources that text makes available, and that the primary access to scientific knowledge is through the reading of text, then it is easy to see that in learning how to read such texts a great deal will be learned about both substantive science content and the epistemology of science. (pp. 236–237)

In broadening the notion of science literacy, Lemke (2004) would have us extend our thinking to embrace the literacies of science, which include reading and writing verbal texts in relation to any number of other semiotic markers (e.g., chemical symbols, mathematical formulas, whiteboard diagrams, animated simulations, a display on a hand-held calculator, a teacher's gestures and tonal register, the actions of other students as they manipulate a demonstration apparatus, a teacher's lecture notes online and hyperlinked, and so on). From Lemke's point of view, these multiple modes of meaning making, rather than being redundant, are *individually* presenting complete and relevant information—a situation that ultimately calls on learners to integrate and cross-contextualize the information they *read* multimodally. Not an easy task to accomplish and one that requires no small amount of mediation on a teacher's part. This is a tall order for anyone; but when it is interpreted as a veiled effort to turn science teachers into reading teachers, trouble ensues.

In the first half of the 20th century, long before multimodal learning had entered our lexicons, the well-known psychologist William S. Gray popularized the aphorism "every teacher a teacher of reading" (as cited in Moore, Readence, & Rickelman, 1983, p. 424) to call attention to the comprehension difficulties older students were experiencing in their efforts to read the textbooks their teachers had assigned. Gray's objective was to goad content area teachers into taking responsibility for helping students learn to comprehend their assigned texts. Despite its continued popularity to this day as a rallying cry among reading specialists (and sometimes building principals), by and large "every teacher a teacher of reading" has fallen on deaf ears—and for good reason. It is too often perceived by content area teachers as tantamount to downgrading or marginalizing their expertise in disciplinary content (Fisher & Ivey, 2005; Moje, Young, Readence, & Moore, 2000).

Emphasizing the reading process over science content is a stance we prefer to avoid. Instead, we argue for an approach to subject matter teaching and learning that views science literacy not as an add-on but, instead, as integral to the disciplinary content that science teachers value and want to share with their students (Moje et al., 2004). This approach also recognizes the potential synergy between literacy and understanding in science (US NRC, 2007), and it easily incorporates what students find engaging about the multiple forms of text available to them (e.g., digital, linear print, visual, aural, symbolic, iconic, performative) both in and out of school. In a sense, it could be argued that student engagement with multiple modes of communication has hastened literacy educators' interest in multiliteracies.

More than a decade ago, the New London Group (1996), working within a multiliteracies framework, drew attention to the need for integrating multiple modes of communication (e.g., linguistic, visual, oral, auditory, kinesthetic) in light of a culturally diverse world grown significantly more attached to new communication technologies—although multiliteracies in and of themselves need not necessarily involve digital technologies (Lankshear & Knobel, 2006). Typically, the term *multiliteracies* denotes more than "mere literacy" (Cope, Kalantzis, & New London Group, 2000, p. 5), which remains language- and print-centered. Although a relatively new term, multiliteracies is a concept that was foreshadowed in at least two science education publications dating back to the early 1990s: *Science for All Americans* (American Association for the Advancement of Science [AAAS], 1990) and *Benchmarks of Science Literacy* (AAAS, 1993). In both publications, science literacy was viewed as having to do with broad, explanatory processes rather than narrowly defined reading skills:

> science literacy requires understandings and habits of mind that enable citizens … to make some sense of how the natural and designed worlds work, to think critically and independently, to recognize and weigh alternative explanations of events and design trade-offs, and to deal sensibly with problems that involve evidence, numbers, patterns, logical arguments, and uncertainties. (1993, Introduction para 3)

If science is describing and explaining the natural world and people's designs for that world, then science literacy can be thought of as enabling the development of scientific knowledge as well as providing a variety of means to communicate that knowledge to others.

Traditional definitions of literacy commonly refer to reading and writing in relation to a variety of texts that are language-specific tools for communicating. In today's highly technical and global society, texts are more broadly construed and language is no longer the sole mode for communicating, thus causing traditional definitions of literacy to be perceived as both limited and unresponsive to society's needs (Cunningham, Many, Carver, Gunderson, & Mosenthal, 2000). The United Nations Educational Scientific and Cultural Organization (UNESCO, 2004), in an effort to define literacy more broadly, issued the following statement:

> Literacy is the ability to identify, understand, interpret, create, communicate and compute, using printed and written materials associated with varying contexts. Literacy involves a continuum of learning in enabling individuals to achieve their goals, to develop their knowledge and potential, and to participate fully in their community and wider society. (p. 13)

Although we would not summarily dismiss UNESCO's definition of literacy, we question whether it goes far enough. For instance, in a report titled *Beyond the 3 Rs: Voter Attitudes Toward 21st Century Skills*, 80% of the 800 registered voters polled in the United States claimed "that the kind of skills students need to learn to be prepared for the jobs of the 21st century is different from what they needed 20 years ago" (Partnership of 21st Century Skills, 2007, p. 1). Support for the public's perception of the need for new skills among members of the future workforce can be seen in the items on the reading and science tests administered by the Programme for International Student Assessment (Organisation for Economic Co-operation and Development, 2003) as well as in the cross-disciplinary work of literacy, mathematics, and science educators (Anderson, Lin, Treagust, Ross, & Yore, 2007).

Because literacy is often viewed as singular in form and a metaphor for competency in basic reading and writing skills, it is useful to bear in mind the historical trajectory of the term multiliteracies, which the first author has traced in detail elsewhere (Alvermann, in press) and recapped here. Traditionally, the autonomous model of literacy has been (and continues to be) prevalent in school-based literacy practices in the United States. This model views reading and writing as neutral processes that are largely explained by individual variations in cognitive and physiological functioning, notwithstanding Gee's (1990) seminal publication, *Social Linguistics and Literacies: Ideology in Discourses* and Heath's (1983) influential research that demonstrated it is how children are socialized into different literacies (their different ways with words and whether those ways match the school's approach to literacy instruction) that matter. The autonomous model assumes a universal set of reading and writing skills for decoding and encoding printed text. In a critique of this model, Street (1995) drew from his anthropological field work on literacy in Iran during the 1970s. Briefly, he questioned the assumption that reading and writing are neutral processes, thereby laying the groundwork for an ideological model of literacy:

> A great deal of the thinking about literacy … has assumed that literacy with a big 'L' and single 'y' [is] a single autonomous thing [with] consequences for personal and social development.… One of the reasons for referring to this position as an autonomous model of

> literacy is that it represents itself as though it is not a position located ideologically at all, as though it is just natural. One of the reasons why I want to call the counter-position ideological is precisely in order to signal that we are not simply talking here about technical features of the written process or the oral process. What we are talking about are competing models and assumptions about reading and writing processes, which are always embedded in power relations. (pp. 132–133)

A point worth underscoring is that viewing literacy as ideologically embedded does not lessen the importance of the cognitive aspects of reading and writing, nor does it eliminate the need to attend instructionally to the technical skills associated with the autonomous model. Rather, as viewed from Street's (1995) perspective, the ideological model subsumes the autonomous model and simultaneously incorporates an array of social and cultural ways of knowing that can account for seemingly absent but always present power structures. Issues of power, whether visible or invisible, are endemic in the sciences, particularly in cases where people's lives and the future of life on earth as we know it are at stake (Ford & Forman, 2006). How such issues are taken up (or ignored), of course, will depend in no small way on the degree to which teachers recognize the value of science literacy for enabling students to think critically and independently in order to weigh alternative explanations and thus "deal sensibly with problems that involve evidence, numbers, patterns, logical arguments, and uncertainties" (AAAS, 1993, para 3).

Finally, because no discussion on language and literacy in relation to science teaching and learning would be complete without some mention of theoretical frameworks, we focus first on a theory that is particularly germane to multiliteracies, namely, a social semiotic theory of multimodality (Lemke, 1989). Used by Kress and colleagues (Kress, 1996; Kress & van Leeuwen, 1996) to study communication in its widest sense—linguistic, visual, oral, gestural, musical, kinesthetic, and digital—social semiotic theory attempts to explain how people employ various resources (signs) available to them through different modes to represent what they wish to communicate to others. For example, consider the different modes an elementary science teacher might invoke in representing the hoped-for effect of hybrid cars on the earth's atmosphere. She or he might hold up a picture of a hybrid car that exudes sleekness and shine, or show a video that captures the voices and facial expressions of pleased hybrid owners, or interact with an online digital graph that compares the hybrid car's gas consumption to the nonhybrid's. This teacher's use of multiple modes to represent how hybrid cars might contribute to a cleaner environment could be said to indicate what is salient about that relationship for her or him. It could also indicate what she or he perceives are the students' interests and developmental needs.

By examining the multimodal representations people make of available resources, it is possible to infer what matters to them (Jewitt & Kress, 2003). Drawing inferences about multimodal representations is but one of several disciplinary practices in science; others include, though not exclusively, argumentation (von Aufschnaiter, Erduran, Osborne, & Simon, 2008; Osborne, Erduran, & Simon, 2004; Tippett, in press), experimentation (Lehrer & Schauble, 2006; Lehrer, Schauble, & Petrosino, 2001), intertextuality (Varelas & Pappas, 2006), and genre writing (Akkus, Gunel, & Hand, 2007; Prain & Hand, 1996; Wallace, Hand, &

Yang, 2004). Each of these practices is associated with various ways of knowing and doing science literacy research.

But to return to theoretical matters, we want to focus next on Ford and Forman's (2006) effort to develop a framework within Vygotsky's (1978) sociocultural theory that would assist them in evaluating what counts as disciplinary learning in science classrooms. This new framework builds on Lave and Wenger's (1991) notion of active participation (apprenticeship) in learning, not unlike what Dewey (1916) had envisioned at the start of the 20th century. In contrast to how others had responded in the past to Dewey's and Lave and Wenger's calls for participatory learning, Ford and Forman drew on the notion of student agency (Holland, Lachicotte, Skinner, & Cain, 1998). Specifically, they developed a framework that enables researchers to address questions such as: How does pedagogically attending to the social and material practices of doing science provide students with a grasp of those practices? In other words, what do students need to take away from a science lesson to demonstrate they are acquiring knowledge about how the discipline works?

Although we can envision this question appealing to researchers interested in accounting for more than the mere acquisition of behavioral skills and mental constructs in learning science, we can also imagine the difficulties one might currently experience in obtaining funds to study it. This would seem to be the case especially if US federal dollars for education research continue to flow in the direction of those who embrace the Gold Standard RCT approach over more qualitatively nuanced inquiries involving student agency in learning science.

4.2 Restrictive Policies in the Gold Standard Era

Critical of what he perceived as unwarranted political interference both in funding matters and literacy instruction by policy makers in the United States and Great Britain, Street (2003) wrote vehemently against the enactment of policies that support increasingly narrowed definitions of literacy:

> Policy makers … bringing the 'light' of literacy to the 'darkness' of the 'illiterate,' and educationalists … similarly arguing for the economic and social benefits of a narrowly defined and disciplined 'literacy' can simply argue that all of those counter-examples of the complexity and meanings of literacy in people's everyday lives are not relevant to their agenda. Local, everyday, home literacies are seen within that frame as failed attempts at the real thing, as inferior versions of the 'literacy' demanded by the economy, by educational institutions, by the politics of centralizing and homogenizing tendencies. (pp. xi–xii)

Framing the problem somewhat differently, Coburn (2006) also warned, though perhaps too late, of factors that need scrutinizing before so-called solutions to perceived problems are enacted into law:

> Policy problems do not exist as social fact awaiting discovery. Rather, they are constructed as policymakers and constituents interpret a particular aspect of the social world

> as problematic. How a policy problem is framed is important because it assigns responsibility and creates rationales that authorize some policy solutions and not others. (p. 343)

Both of these critiques point to how policy is socially constructed in response to a perceived problem and sometimes in direct opposition to how those individuals who must eventually carry out a particular policy might see the problem. As a result, the potential for resentment to build exists, although inequities in funding priorities tend to keep this risk lower than might otherwise be anticipated. Worth bearing in mind is that ill-founded policies in literacy and science education may result from inaccurate definitions of the shared goal: science literacy (Yore, Pimm, & Tuan, 2007). A further possibility is that even when accurate definitions of that goal exist, funding may not follow.

Yearly, the US federal government allocates billions of dollars to departments and agencies for research purposes. For example, the National Institutes of Health (US NIH, 2007) receive approximately $28 billion, which supports research projects and institutes directly or the management of research in the medical and human sciences. The National Science Foundation (US NSF, 2007) receives monies just shy of $6 billion; approximately 95% of that money supports research activities, resources, and facilities. In sharp contrast, the Institute of Education Sciences (IES), the research arm of the US Department of Education (US ED) is given approximately $550 million to sustain its program of research and evaluation (US ED, 2007).

Although the research budget of the IES may not match other agencies in scale, it is here where a clear governmental research agenda can be seen. The US ED's *Fiscal Year 2008 Budget Summary* states: "A cornerstone of NCLB is investment in research ... IES ensures that the Federal investment in education research and data collection is well-managed and relevant to the needs of educators and policymakers" (US ED, 2007, p. 68). Of the approximately 70 times the word *research* appears in the context of the budget summary, in only 10 places is it paired with the words *scientifically based*, *scientific evidence*, or like terms that use a derivative of the word *science*. More than just an interesting curiosity, one might interpret this terminology to mean that the US ED counts some education research as scientific and other research as unscientific—an interpretation well documented in a growing number of critiques that question the usefulness of the Gold Standard's one-size-fits-all approach to research (Freeman, deMarrais, Preissle, Roulston, & St. Pierre, 2007; Maxwell, 2004; Raudenbush, 2005). While one could argue that there are some problem spaces and research questions that require developmental explorations prior to large-scale inquiries, as qualitative researchers we find that argument objectionable. The positioning of qualitative research for many years was viewed as exploratory, which then set it up to be *prescientific*, thus not scientific. We do not want to support the notion that first we do exploratory qualitative research and the *real* science begins later.

Funded by IES, the What Works Clearinghouse (WWC) is charged with disseminating research that provides "the strongest evidence of effects: primarily well conducted randomized controlled trials and regression discontinuity studies, and secondarily quasi-experimental studies of especially strong design" (US IES, n.d., para 1). In terms of the potential for curricular impact at the state and local levels, studies are either stamped with a green-means-go Meets Evidence Standards label, a cautionary-yellow Meets Evidence Standards with Reservations for strong quasi-experimental and randomized trials with some problems, or a halting-red Does Not Meet Evidence Screens for everything else. Because the WWC is in place to "provide evidence-based information for policymakers, researchers, and educators" (US ED, 2007, p. 68), the message to those audiences is that studies with randomized controlled trials, regression models, and quasi-experimental models are passable and that anything else is unworthy of consideration. Of particular note, Millar and Osborne (see Chap. 3) provide additional insights into different types of research and offer three examples of results (wait-time, formative assessment, cognitive acceleration) that have influenced professional practice.

As noted earlier, research findings that are informed primarily by the Gold Standard tend to sanction certain kinds of curricular development while discouraging other types, thus narrowing the range of science literacy practices that teachers have at their disposal. To explore ways of addressing this situation, we examine next the assumptions underlying five different styles of doing research and discuss their potential for offsetting, at least in part, the negative effects of a one-size-fits-all approach to scientific inquiry.

4.3 Five Ways of Knowing (and Doing) Research

In this final section, we explore qualitative and quantitative researchers' assumptions about five different dimensions or styles of doing research. These assumptions reflect different worldviews of what counts, or should count, as scientifically valid research. Our goal is to describe why certain assumptions underlying research that adheres to the Gold Standard make it difficult to disseminate to a community of practitioners, such as science educators and particularly to those within that community who are knowledgeable about various aspects of science literacy. We take exception to the claim that only Gold Standard research should find its way into science classrooms, arguing instead that in some instances findings from research other than that which is associated with RCTs may be equally well, if not better, positioned to influence the world of practice.

4.3.1 Correspondence versus Coherence Theory

Researcher bias that is weighted more toward one end than the other of an imaginary continuum separating correspondence theory from coherence theory is reflective of the age-old philosophical debate about theories of truth. According to Stanovich (2003), correspondence theories support the realistic worldview that "there is a real world out there that exists independently of our beliefs about it" (p. 107). This view, he claimed, is held by most reading researchers and is in stark contrast to coherence theories of truth, which rely on constructivist principles and require that "beliefs fit together in a reasonably logical way" (p. 107). Coherence theories appeal mostly to qualitative researchers with idealist world views, in his opinion, because they resonate with narratives or stories passed down from one generation to the next (folk psychology). He concluded, perhaps a bit prematurely, that the very fact that correspondence theories do not follow a narrative logic makes them less compelling in the market of ideas "out there" for dissemination—his argument being that "a correspondence theory of truth often necessitates the frustration of the strong human need for narrative coherence in explanation" (p. 108). Similar differences can be seen in the nature of science debates—traditionalist (absolutist realist), modernist (evaluativist naive realist), postmodernist (relativist idealist (Staver, 1998; Yore, Hand, & Florence, 2004).

Coherence theories, with their emphasis on constructivism, are sometimes criticized because in their most radical form the sole truth-criterion is whether or not one's beliefs make sense to the person constructing them. This can be problematic in science education, particularly, where authority resides within the disciplinary community and not the individual, as Linn and Eylon (2006) have noted:

> The constructivist perspective stresses the effort that students expend to make sense of the natural world. It resonates with research showing that the intuitive ideas students formulate tend to remain in the repertoire even when normative ideas are added.... It is consistent with reports that students often seem unperturbed when shown that their ideas do not converge or align with scientific principles. (p. 521)

To our way of thinking, this is neither an indictment of coherence theory nor a denunciation of constructivism. For as Linn and Eylon go on to point out in their review of the research on science education, evidence exists to suggest that when students hold worldviews that value coherence over correspondence, they are more apt to look for contradictions—a finding that would seem to have real-world implications. For example, as literacy educators with an abiding interest in critical theory, we maintain that teachers who create conditions that encourage students to question and evaluate all texts, even those espousing normative ideas about scientific principles, are involving them in ways that map quite nicely onto how scientists go about their work. According to the *National Science Education Standards* (Florence & Yore, 2004; US NRC, 1996; Yore, Florence, Pearson, & Weaver, 2006), students who generate questions for the purpose of finding answers to questions derived from curiosity about everyday experiences are participating in scientific inquiry.

Thus, in contrast to what Stanovich (2003) would have us believe, we maintain that the so-called constructivist bias of qualitative researchers need not stand in caricatured opposition to the more correspondence-oriented worldviews of those whose research aligns better with the Gold Standard. If anything, a bias toward making contradictory ideas fit together narratively, in a reasonably logical way, might be interpreted as the impetus that predisposes scientists to think outside the box.

4.3.2 Analytic Reductionism versus Holism

Ways of knowing that involve inquiring into the subprocesses of a phenomenon—whether it is reading, writing, arguing, representing, or some other form of communication—are fractionated by design in order to reduce the complexity associated with understanding that phenomenon. According to Stanovich (2003), the assumption is that scientific progress is doubtful when critics of this approach to doing research refuse to consider the more simple explanations for complex processes (e.g., reading, writing, and learning science). In contrast, reductionists are willing to live with the simplicity, at least temporarily. Claiming that such simplification eliminates the very essence of the process under investigation, the holists (as those who are opposed to analytic reductionism are called) refuse to go along with the reductionists' gamble that "a gain in explanatory power will not result in a loss of contextualized understanding" (p. 111). This refusal, from Stanovich's point of view, becomes an even greater problem when holistic literacy researchers seek to undermine analytic progress by proposing what he called "subtractive critiques" (p. 112) or counterarguments, which for all intents are aimed at preventing the more simplified explanation from being broadly endorsed in the field. We would interject, however, that rigor and evidence are of concern to all researchers, regardless of whether they ground their work in analytic reductionism or holism. Researchers from both approaches, to our way of thinking, are interested in preciseness and in contextualized data warranted as evidence. At the same time, we acknowledge that there are differences in terms of when and how such concerns come into play.

At the very least, Stanovich (2003) argued that holistic critiques should add to rather than subtract from the work of the reductionists. By way of example, he called attention to how additive critiques from the research on cross-linguistic differences led eventually to the contextualization of the analytic reductionist understanding of orthography and its link to reading achievement. Drawing on other examples from biology (e.g., John Maynard Smith's disagreements with Evelyn Fox Keller on the mechanisms of gene action), he concluded that the continuum having analytic reductionism and holism as its endpoints is probably best bridged by a mix of the two. This ecumenical view, he submitted, might lead to improved dissemination and encompassing of the advances in reading research that scientifically valid inquiry is meant to ensure. It might also, he noted, acknowledge the fact that "science is a delicate epistemological game … [wherein] many of its modes of operation represent dispositions rather than rules" (p. 121). Based on our own experiences both as researchers and observers of other researchers, we would have to say that ways of knowing are often in productive tension with ways of being. Thus, taking an ecumenical view, as Stanovich suggested, could conceivably address complexity across a program of research but at the same time allow for a reductionist study or two along the way.

Epistemological and ontological dispositions similar to those just noted are at play as well in Simon, Erduran, and Osborne's (2006) work on argumentation. They ground their work in Toulmin's (1958) classic model of argumentation, which specifies four components—data, claim, warrants, backing—that are basic to an argument pattern and must be in place before an extended model of competing claims and rebuttals can be adjudicated within a disciplinary community, such as science. In a year-long professional development study involving 12 middle school teachers in the greater London area, the task was to teach a series of lessons in which students were asked to argue for or against a funding proposal to build a new zoo. Analytic tools developed by the researchers (based on Toulmin's four components) were used to analyze audio- and videotaped observations of the teachers' efforts at implementation. Findings showed that the teachers varied considerably in their level of comfort with the goals of argumentation. Ford and Forman's (2006) critique of that study is of interest here because it points to the possibility that a reductionist argumentation model, however well formed and regardless of its attention to an important disciplinary practice of science, simply did not provide teachers with the bigger picture:

> If students and teachers are not fully aware of why argumentation … is useful, then they have failed to … participate in a disciplinary practice. If [they] are not aware of or interested in authoring truth claims, then training them to use argumentation would not be sufficient for inculcating a disciplinary approach in science education. (p. 19)

The parallels between this critique and Stanovich's (2003) earlier assessment of the difficulties encountered when researchers working from a holistic perspective refuse to go along with a fractionated account of the reading process are worth commenting on if for no other reason than both situations draw attention to the need to bridge the imaginary continuum separating analytic reductionism and holism. Minus such bridging, it is all too easy for researchers as well as teachers to slip comfortably into a noncritical discourse that avoids conflict, contradiction, and critical thinking—all very much a part of scientific thinking and science literacy.

4.3.3 Probabilistic Prediction versus a Case-based Approach

Of the five dimensions that Stanovich (2003) described as causing confusion among teachers and the public at large, none is perhaps more troubling to him than the one he labeled probabilistic prediction versus a case-based approach. The problem is that neither of the endpoints on this continuum is an appropriate replacement logic or model for the other. Probabilistic prediction, according to Stanovich, can seem alien to teachers and other laypersons who reason why try it if a study's results cannot be guaranteed to work reasonably well with every student. Why take a chance, they ask, when high-stakes testing and adequate yearly progress (AYP) are givens; and the penalties imposed on non-AYP schools can be formidable.

What worries Stanovich (2003) is not so much a feeling of inadequacy in explaining probabilistic prediction to a lay audience but rather a concern that this audience will turn to the more intuitively appealing, case-based approach for the wrong reason. In clarifying that he is not opposed to the case-based approach on general principles, he wrote:

> The teacher's replacement logic is appropriate when a case-based approach is actually called for—when the teacher is not making an aggregate decision. But often teachers (and the public) carry over the case-based style into situations where the decision is clearly an aggregate one—[for example], choosing group activities in a classroom, choosing instructional time allocation, choosing a school's or district's curriculum, or choosing the allocation of in-service training time. There, the probabilistic style is the appropriate one, and the replacement logic is inappropriate—there is no replacement program that could beat the number of children who are helped by the baseline treatment. (pp. 115–116)

Although we grant that Stanovich's worry stems from real-world experience—in his 2003 article cited here, he described a scenario in which the inevitable question *Does it work for everyone?* comes up when he speaks to teacher audiences, we simply could not find evidence from our search of the case-study research in science literacy that teachers overgeneralize results. What we did find, interestingly, was this: When science literacy is explored in naturalistic contexts of the science classroom, case-study researchers are careful to point out qualifiers that establish boundaries for their claims. They also are apt to study how individual students, rather than groups of students, make meaning of a science activity (e.g., von Aufschnaiter et al., 2008), or how a single teacher chooses from available curricular offerings in literacy instruction (Hasbrouck, Woldbeck, Ihnot, & Parker, 1999) or science (Tsai, 2002) as opposed to an entire school district (e.g., see a mixed-methods study conducted by Roehrig, Kruse, & Kern, 2007).

Case studies, of course, can and do focus on more than one individual, and not all case-study researchers are equally careful to point out qualifiers that establish boundaries for their claims. Still, we have to wonder if the concern that Stanovich (2003) expressed about probabilistic prediction versus a case-study approach might be more tied to his personal experiences than to any large-scale misappropriation of case-study findings by classroom teachers. Several other fundamental issues surrounding probabilistic prediction and case-based approaches are addressed in terms of quantitative meta-analysis and qualitative syntheses by Rossman and Yore (see Chap. 26).

4.3.4 Robust-process Explanations versus Actual-sequence Explanations

Robust-process explanations are more often a topic of discussion in the philosophy of biology literature than in the literature on reading, according to Stanovich (2003), although he contended they deserve attention in the reading field as well. "To a philosopher, someone seeking a robust-process explanation is looking for a causal model that defines a class of possible worlds—the set in which a posited series of causal linkages holds" (p. 116). For example, drawing from the work of Kim Sterelny (a philosopher of evolutionary biology), Stanovich described how evoking large-scale principles through causal modeling can answer questions such as, *How might we explain the fact that the birthrate in New Zealand dropped dramatically after World War 1?* The answer from a robust-process point of view is this:

> For example, it is known that technological change during wartime spawns increasing urbanization, which in turn spawns lower birthrates. One could then show that these large-scale causal mechanisms were at work in New Zealand at the time. A robust-process explanation in this case would identify New Zealand as a member of the class of countries in which the purported causal conditions were fulfilled. (p. 116)

As Stanovich went on to note, while this explanation would satisfy a scientist focused on robust processes, it would not adequately address an audience at the other end of the continuum wanting to know the actual-sequence explanation (e. g., the "precise micromechanisms" (p. 117) that were at work in New Zealand following World War I). An audience looking for an actual-sequence explanation would be interested in the details of the New Zealand situation specifically and, therefore, would be interested in knowing the microhistories of the birth and death rates of individuals. In other words, a scientist working under the assumptions of the robust-process approach would accept New Zealand as *the type of country in which* a particular phenomenon might occur, whereas someone interested in the actual-sequence approach would accept New Zealand as *the* country of focus.

In an analogous example drawn from the field of literacy education, Stanovich (2003) speculated that, while robust-process explanations might disseminate well among researchers interested in studying young children's developmental reading proficiencies, these same explanations would not disseminate well among teachers interested in knowing how a *particular* child reached a certain point of proficiency in reading (the actual-sequence end of the continuum). To Stanovich's way of thinking, the two endpoints of this dimension's imaginary continuum work in concert regardless of the degree to which they would find acceptance among different audiences.

4.3.5 Consilience versus Uniqueness

In this, the fifth and last dimension or style of doing research that Stanovich (2003) put forth, he defined *consilience* as the "unification of knowledge by the linking of facts and fact-based theory across disciplines to create a common groundwork of explanation" (after a book of the same name authored by E. O. Wilson and cited in Stanovich, pp. 117–118). Literacy researchers who do the best work in the field of generalized theory building, he claimed, are those who have sought connections to other fields, specifically cognitive neuroscience, computer simulation, and linguistics. Becoming a better consumer of scientifically valid research so that one can make the right decisions about which programs and practices to implement will only happen, he reasoned, if the evidence that is amassed in any given area (e.g., the scientific study of reading) is achieved through consilience and then successfully disseminated to those who can put that evidence into practice. This was the central goal of the 1st Island Conference in which psychology, philosophy, linguistics, and pedagogy were connected to science literacy for all (Hand et al., 2003; Yore, Hand, Goldman et al., 2004).

The counterpoint to consilience is *uniqueness*, which Stanovich (2003) described as “the faddish tendencies in the field of education to search for magic bullets and miracle cures deriving from theories that do not cohere with the knowledge being developed by allied disciplines” (p. 118). When the results from so-called faddish or magic-bullet research are disseminated more widely than those backed by evidence obtained through consilience, he reasoned, then it typically leads to a weakening of models built on consilience considerations. How does this happen? We have observed in our own field of research (literacy education) that an academic climate that rewards faculty for moving their program of research forward all too often sends the message (intended or not) that conducting research for the purpose of confirming or adding to a study’s explanatory power is rather static behavior, almost like standing still.

This rather grim picture raises a more basic question: What counts (or should count) as scientifically valid knowledge in the allied disciplines? Is it not the case that consilience achieved through certain bodies of knowledge (e.g., cognitive neuroscience, computer simulation, linguistics) will have a dissemination index that is quite different from consilience achieved through certain other bodies of knowledge (e.g., cultural studies, social semiotics, transformative social action research)? If the answer to that question is *yes* and if the evidence claims from both sets of unified bodies of knowledge are equally well supported, then it would be primarily a judgment call as to which body of knowledge is more or less scientifically valid and thus deserving of dissemination. Of course, Gergen and Gergen (2000) have reminded us, and it bears repeating here, that:

> [i]t borders on the banal to suggest that everything can be valid for someone, sometime, somewhere. Such a conclusion both closes off dialogue among diverse groups and leads to the result that no one can speak about another. Such an outcome would spell the end of social science inquiry. Dialogue is invited, then, into how situated validity is achieved, maintained, and subverted. (p. 1032)

4.4 Closing Thoughts

In some ways Stanovich’s (2003) five dimensions or styles of research seem well suited to inviting the kind of dialogue that Gergen and Gergen (2000) envisioned. By providing more than a convenient heuristic for accommodating paradigm differences between quantitative and qualitative researchers, the five dimensions offer valuable insight into the degree to which researchers of varying epistemological and ontological perspectives might move around on the five continua and still be viewed as doing scientifically valid research—and, by extension, as producing evidence that merits dissemination.

In other ways, the five dimensions seem unlikely to promote the desired dialogue, particularly if scientific validity is perceived as less of a concern by some researchers than by others. For example, among qualitative researchers, credibility achieved through spending sufficient time in the field, peer debriefing, analyzing negative cases, and member checking (Guba & Lincoln, 1989)—not randomized experiments, however well controlled—are likely to be the preferred points of discussion. Even among quantitative researchers, there is disagreement as to the need for a unitary concept of validity (Lissitz & Samuelsen, 2007). Then, too, as Stanovich (2003) made clear, he holds disparate views about how much movement is permissible on any given dimension or style. For example, on three of the five dimensions (analytic reductionism versus holism, probabilistic prediction versus a case-based approach, and robust-process versus actual-sequence explanations), he maintained one could take any position between the two endpoints of each continuum and still be viewed as doing scientifically valid research. Not so, however, on the remaining two dimensions (correspondence versus coherence theory and consilience versus uniqueness). In those instances, he would require "at least some minimal adherence to correspondence and consilience values as essential to the scientific attitude" (p. 121).

To argue that such permissiveness will be met by skepticism by the very policy makers who gave us the Gold Standard is certainly one response to Stanovich's (2003) five ways of knowing (and doing) research. Our inclination, however, is to use the five continua as a starting point for discussions among researchers in the science literacy community. We are cautiously optimistic that a dialogue could ensue and might be an important first step in widening the field's perceptions of what constitutes valid research in the Gold Standard era. Bringing more voices into the discussion might also lead to productive changes in policies that affect research priorities and funding decisions—two areas that are of profound interest to science literacy researchers and indirectly to curriculum specialists and teachers. Keeping communication lines open about different ways of knowing (and doing) science literacy research is our contribution to what we perceive as an ongoing discussion. Along the way, we expect to encounter travelers of many sorts—other researchers, teacher educators, classroom teachers, administrators, parents, and policy makers—who can only add to whatever understandings we presently possess.

References

Akkus, R., Gunel, M., & Hand, B. (2007). Comparing an inquiry-based approach known as the science writing heuristic to traditional science teaching practices: Are there differences? *International Journal of Science Education, 29*(14), 1745–1765.

Alvermann, D. E. (in press). Sociocultural constructions of adolescence and young people's literacies. In L. Christenbury, R. Bomer, & P. Smagorinsky (Eds.), *Handbook of research on adolescent literacy*. New York: Guilford.

American Association for the Advancement of Science. (1990). *Science for all Americans: Project 2061*. New York: Oxford University Press. Available from http://www.project2061.org/publications/sfaa/online/sfaatoc.htm

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy: Project 2061*. New York: Oxford University Press. Available from http://www.project2061.org/publications/bsl/online/index.php?txtRef = &txtURIOld = %2Fpublications%2Fbsl%2Fonline%2Fbolintro%2Ehtm

Anderson, J. O., Lin, H. L., Treagust, D. F., Ross, S. P., & Yore, L. D. (2007). Using large-scale assessment datasets for research in science and mathematics education: Programme for International Student Assessment (PISA). *International Journal of Science & Mathematics Education*, *5*(4), 591–614.

Aufschnaiter, C., von, Erduran, S., Osborne, J., & Simon, S. (2008). Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge. *Journal of Research in Science Teaching, 45*(1), 101–131.

Coburn, C. E. (2006). Framing the problem of reading instruction: Using frame analysis to uncover the microprocesses of policy implementation. *American Educational Research Journal*, *43*(3), 343–349.

Cope, B., Kalantzis, M., & New London Group (Eds.). (2000). *Multiliteracies: Literacy learning and the design of social futures*. London: Routledge.

Cunningham, J. W., Many, J. E., Carver, R. P., Gunderson, L., & Mosenthal, P. B. (2000). How will literacy be defined in the new millennium? [RRQ Snippet]. *Reading Research Quarterly*, *35*(1), 64–71.

Dewey, J. (1916). *Education and democracy*. New York: Macmillan.

Education Sciences Reform Act of 2002. Pub. L. No. 107–279, 116 Stat. 1940. (2002).

Fisher, D., & Ivey, G. (2005). Literacy and language as learning in content-area classes: A departure from "Every teacher a teacher of reading". *Action in Teacher Education*, *27*(2), 3–11.

Florence, M. K., & Yore, L. D. (2004). Learning to write like a scientist: Coauthoring as an enculturation task. *Journal of Research in Science Teaching*, *41*(6), 637–668.

Ford, M. J., & Forman, E. A. (2006). Redefining disciplinary learning in classroom contexts. Review of Research in Education, *30*(1), 1–32.

Freeman, M., DeMarrais, K., Preissle, J., Roulston, K., & St Pierre, E. A. (2007). Standards of evidence in qualitative research: An incitement to discourse. *Educational Researcher, 36*(1), 25–32.

Gee, J. P. (1990). *Social linguistics and literacies: Ideology in discourses*. London: Falmer.

Gergen, M. M., & Gergen, K. J. (2000). Qualitative inquiry: Tensions and transformations. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (2nd edn., pp. 1025–1046). Thousand Oaks, CA: Sage.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

Hand, B., Alvermann, D. E., Gee, J. P., Guzzetti, B. J., Norris, S. P., Phillips, L. M., et al. (2003). Message from the "Island group": What is literacy in science literacy? [Guest editorial]. *Journal of Research in Science Teaching*, *40*(7), 607–615.

Hasbrouck, J. E., Woldbeck, T., Ihnot, C., & Parker, R. I. (1999). One teacher's use of curriculum-based measurement: A changed opinion. *Learning Disabilities Research & Practice*, *14*(2), 118–126.

Heath, S. B. (1983). *Ways with words: Language, life, and work in communities and classrooms*. Cambridge, UK: Cambridge University Press.

Holland, D. C., Lachicotte, W., Skinner, D., & Cain, C. (1998). *Identity and agency in cultural worlds*. Cambridge, MA: Harvard University Press.

Jewitt, C., & Kress, G. R. (Eds.). (2003). *Multimodal literacy*. New York: Peter Lang.

Kress, G. R. (1996). *Before writing: Rethinking the paths to literacy*. London: Routledge.

Kress, G. R., & Leeuwen, T., van. (1996). *Reading images: The grammar of visual design*. London: Routledge.

Lankshear, C., & Knobel, M. (2006). *New literacies: Everyday practices and classroom learning* (2nd edn.). Berkshire, UK: Open University Press.

Lave, J., & Wenger, E. (1991). *Situated learning. Legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.

Lehrer, R., & Schauble, L. (2006). Scientific thinking and scientific literacy. In W. Damon, R. Lerner, K. A. Renninger, & E. Sigel (Eds.), *Handbook of child psychology* (6th edn., Vol. 4, pp. 153–196). Hoboken, NJ: John Wiley & Sons.

Lehrer, R., Schauble, L., & Petrosino, A. J. (2001). Reconsidering the role of experiment in science education. In K. D. Crowley, C. D. Schunn, & T. Ikada (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 251–278). Mahwah, NJ: Lawrence Erlbaum.

Lemke, J. L. (1989). Social semiotics: A new model for literacy education. In D. Bloome (Ed.), *Classrooms and literacy* (pp. 289–309). Norwood, NJ: Ablex.

Lemke, J. L. (2004). The literacies of science. In E. W. Saul (Ed.), *Crossing borders in literacy and science instruction: Perspectives on theory and practice* (pp. 33–47). Newark, DE: International Reading Association & National Science Teachers Association.

Linn, M. C., & Eylon, B. S. (2006). Science education: Integrating views of learning and instruction. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd edn., pp. 511–544). Mahwah, NJ: Lawrence Erlbaum.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437–448.

Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher*, *33*(2), 3–11.

Moje, E. B., Peek-Brown, D., Sutherland, L. M., Marx, R. W., Blumenfeld, P. C., & Krajcik, J. S. (2004). Explaining explanations. In D. S. Strickland & D. E. Alvermann (Eds.), *Bridging the literacy achievement gap, grades 4–12* (pp. 227–251). New York: Teachers College Press.

Moje, E. B., Young, J. P., Readence, J. E., & Moore, D. W. (2000). Reinventing adolescent literacy for new times: Perennial and millennial issues. *Journal of Adolescent & Adult Literacy*, *43*(5), 400–411.

Moore, D. W., Readence, J. E., & Rickelman, R. J. (1983). An historical exploration of content area reading instruction. *Reading Research Quarterly*, *18*(4), 419–438.

New London Group. (1996). A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review*, *66*(Spring), 60–91.

No Child Left Behind Act of 2001. Pub. L. No. 107–110, 115 Stat. 1425. (2002).

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, *87*(2), 224–240.

Organisation for Economic Co-operation and Development. (2003). *The PISA 2003 assessment framework – mathematics, reading, science and problem solving: Knowledge and skills*. Paris: Author.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching*, *41*(10), 994–1020.

Partnership of 21st Century Skills. (2007). *Beyond the 3 Rs: Voter attitudes toward 21st century skills*. Tucson, AZ: Author.

Prain, V., & Hand, B. (1996). Writing for learning in secondary science: Rethinking practices. *Teaching and Teacher Education*, *12*(6), 609–626.

Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher*, *34*(5), 25–31.

Roehrig, G. H., Kruse, R. A., & Kern, A. (2007). Teacher and school characteristics and their influence on curriculum implementation. *Journal of Research in Science Teaching*, *44*(7), 883–907.

Simon, S., Erduran, S., & Osborne, J. (2006). Learning to teach argumentation: Research and development in the science classroom. *International Journal of Science Education*, *28*(2/3), 235–260.

Stanovich, K. E. (2003). Understanding the styles of science in the study of reading. *Scientific Studies of Reading*, *7*(2), 105–126.

Staver, J. R. (1998). Constructivism: Sound theory for explicating the practice of science and science teaching. *Journal of Research in Science Teaching*, *35*(5), 501–520.

Street, B. V. (1995). *Social literacies: Critical approaches to literacy development, ethnography, and education*. Harlow, UK: Longman.

Street, B. V. (2003). Foreword. In J. Collins & R. K. Blot (Eds.), *Literacy and literacies: Texts, power, and identity* (pp. xi–xv). Cambridge, UK: Cambridge University Press.

Tippett, C. D. (in press). Argumentation: The language of science. *Journal of Elementary Science Education*.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Tsai, C. C. (2002). A science teacher's reflections and knowledge growth about STS instruction after actual implementation. *Science Education*, *86*(1), 23–41.

United Nations Educational Scientific and Cultural Organization. (2004). *The plurality of literacy and its implications for policies and programmes* (UNESCO Education Sector position paper). Paris: Author.

United States Department of Education. (2007). *Fiscal year 2008 budget: Summary and background information*. Retrieved May 7, 2008, from http://www.ed.gov/about/overview/budget/budget08/summary/08summary.pdf

United States Institute of Education Sciences. (n.d.). *What Works Clearinghouse overview: Standards.* Retrieved May 6, 2008, from http://ies.ed.gov/ncee/wwc/overview/review.asp?ag = pi

United States National Institutes of Health. (2007). *Summary of the FY 2008 President's budget*. Retrieved May 6, 2008, from http://officeofbudget.od.nih.gov/PDF/Press%20info-2008.pdf

United States National Research Council. (1996). *The national science education standards*. Washington, DC: The National Academies Press. Available from http://www.nap.edu/catalog.php?record_id = 4962

United States National Research Council. (2002). *Scientific research in education*. Committee on Scientific Principles for Education Research. R. J. Shavelson & L. Towne (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

United States National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Committee on Science Learning, Kindergarten through Eighth Grade. R. A. Duschl, H. A. Schweingruber, & A. W. Shouse (Eds.). Board on Science Education, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

United States National Science Foundation. (2007). *FY 2008 budget request to Congress*. Retrieved June 17, 2008, from http://www.nsf.gov/about/budget/fy2008/index.jsp

Varelas, M., & Pappas, C. C. (2006). Intertextuality in read-alouds of integrated science-literacy units in urban primary classrooms: Opportunities for the development of thought and language. *Cognition and Instruction*, *24*(2), 211–259.

Viadero, D. (2007). 'Scientific' label in law stirs debate. *Education Week*, *27*(1), 23.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wallace, C. S., Hand, B., & Yang, E. M. (2004). The science writing heuristic: Using writing as a tool for learning in the laboratory. In E. W. Saul (Ed.), *Crossing borders in literacy and science instruction: Perspectives on theory and practice* (pp. 355–368). Newark, DE: International Reading Association & National Science Teachers Association.

Yore, L. D., Florence, M. K., Pearson, T. W., & Weaver, A. J. (2006). Written discourse in scientific communities: A conversation with two scientists about their views of science, use of language, role of writing in doing science, and compatibility between their epistemic views and language. *International Journal of Science Education*, *28*(2/3), 109–141.

Yore, L. D., Hand, B., & Florence, M. K. (2004). Scientists' views of science, models of writing, and science writing practices. *Journal of Research in Science Teaching*, *41*(4), 338–369.

Yore, L. D., Hand, B. M., Goldman, S. R., Hildebrand, G. M., Osborne, J. F., Treagust, D. F., et al. (2004). New directions in language and science education research. *Reading Research Quarterly*, *39*(3), 347–352.

Yore, L. D., Pimm, D., & Tuan, H. L. (2007). The literacy component of mathematical and scientific literacy. *International Journal of Science & Mathematics Education*, *5*(4), 559–589.

Chapter 5
Longitudinal Studies into Science Learning: Methodological Issues

Russell Tytler

The Gold Standard for education research promotes randomized controlled trials (RCTs) that can produce generalizable knowledge claims across similar problems and situations. Unfortunately, the Gold Standard does not fully recognize the need for developmental research to better understand the problem space, formulate theory and approaches to teaching and learning, and formulate and pursue associated research questions. This developmental research has been a precursor to the development of interventions together with the necessary instrumentation and technologies required to fully investigate these through the more formal evaluative processes imagined by the Gold Standard. This chapter focuses on longitudinal studies that cover a continuum from such developmental research to research that uses control-experimental features to evaluate interventions. These studies attend to a set of issues dealing with developmental progressions and learning trajectories that require investigation over an extended period of time. It will be argued that these longitudinal studies of a variety of methodological types represent quality research in that rigorous design and implementation produce evidence-based claims. The chapter examines the nature of the relationship between evidence and claims in these studies, to show the possibility of building in control features every bit as strong as those in classic Gold Standard designs. Further, it will be argued that, given the complexity of learning pathways, a simplistic interpretation of RCTs conducted over the shorter term can be misleading in terms of both internal and external validity claims.

5.1 Background

There have been relatively few longitudinal studies of student learning in science, despite the fact that the gains sought for in education generally are long term and permanent (Arzi, 1988, 2004; White & Arzi, 2005). Increasingly, learning is

R. Tytler
Deakin University

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

becoming understood as a complex process involving the interaction of many elements and the building of understandings gradually over the longer term, rather than the sequential imposition of specific conceptions. A sequence of reports by the US National Research Council (1999, 2005, 2007) based on neuroscience and cognitive science research emphasizes the complex interactions between prior knowledge, maturation, experience, and instruction. It is claimed that "to be successful in science, students need carefully structured experiences, instructional support from teachers, and opportunities for sustained engagement with the same set of ideas over weeks, months, and even years" (2007, p. 338). As with curriculum and teaching, these views imply the need for research into student learning that is long term; yet studies of this kind are rare since they are difficult to justify and maintain in a climate where research interests and personal circumstances are unpredictable, research grant funding is normally for a maximum of 3–5 years, and Ph.D. cycles are focused on bounded studies because of timescale issues.

White (1987, 2001) called for longitudinal studies tracing the emergence and development of conceptions. Studies of the development of student conceptions in different domains have almost all been cross-sectional in design and with patterns of growth established by comparing responses of different age students. These studies often involve a mix of qualitative and quantitative analyses with comparative counts of categories of response. The methodological problems with these cross-sectional studies are twofold: first, it is difficult to guarantee that the cohorts are identical in composition to make valid comparisons; and second, while broad patterns might emerge to chart general developmental principles, these cannot be translated to the learning and development pathways of individual students. Black and Simon (1992) argued, on the basis of the need to chart progression of children's knowledge in science, that any proposal to explore progression "is hampered both by the lack of an effective theory of conceptual change and by the absence of substantial evidence about the changes in pupils' ideas with time" (p. 48). Shymansky et al. (1993, 1997) described a complex pattern of conceptual gains and losses for teachers and students over separate interventions. In my own work (1998a, 1998b), a similar complexity of conceptual growth was shown over a 6-month period with understandings depending on context and explanations gradually becoming more consistent with time after the intervention ended.

These complexities in student learning call into question any evaluation of a short-term teaching intervention that does not acknowledge time-related growth or diminution in understanding. Long-term studies are needed to evaluate the effectiveness of teaching interventions in supporting student learning over the longer term. Shymansky, Yore, and Anderson's (2004) study of a long-term, elementary schoolteacher, professional development program is an example of research that took this time factor seriously, both for exploring teacher change and investigating student attitudes and learning over the longer term. Yet most evaluations of teaching and learning interventions, either classroom sequences or large-scale programs are short term and do not trace ongoing benefits of short-term gains nor do they identify longer-term effects that may occur separately from those that are short term.

Two recent issues of international journals (*Canadian Journal of Science, Mathematics, and Technology Education* and *Research in Science Education*) were devoted to long-term studies in science (Shapiro, 2004a; Tytler, Arzi, & White, 2005). The longitudinal studies in these special issues differed from each other in a number of respects: in the extent and type of intervention being described, in whether the focus is on coming to some conclusion about a structured intervention or about developing insight into learning processes, in the way that evidence is used to establish claims, and in the way validity is established in these designs. This chapter draws on these studies and selected others to explore the different logic that applies to their design, methods, and analysis in order to identify the affordances and constraints of the different designs. The analysis identifies the different purposes of such studies including the evaluation over time of time-constrained interventions, the exploration and evaluation of interventions that unfold over time, and the exploration of learning pathways using different degrees of intervention to stimulate learning.

Some of these studies are intended to evaluate the rollout of programs built on preexisting theory while others are more concerned with the building of theory about learning or about curriculum. The studies thus occupy different niches or problem spaces within the broader research agenda referred to by Shelley, Yore, and Hand (see Chap. 1). Each study is strongly grounded in science education literature, and each occupies a different niche in the research quest to understand how best to support student learning. With their different purposes come different types of research questions and different functional requirements. Accordingly, the research designs in most cases differ from the restricted methodological agenda represented by the Gold Standard notion of RCTs. This chapter describes these different longitudinal designs accepting this variation in purpose and addresses the question of how we should speak of *quality* in the methodology attaching to these designs. The discussion includes consideration of how these designs address threats to the internal and external validity of the findings (Campbell & Stanley, 1963).

In selecting the studies, I have followed Arzi's (1988, 2004) definition of longitudinal: that the study must follow an individual or cohort over more than 1 year such that short-term learning effects have time to dissipate and the focus is on more lasting changes. The discussion centers on the different types of intervention and how these sit within the logic of the study, the methodological framework, the nature of the cohort, the causal or other relations that are explored in the study, and the way time appears in the logic of the design. I will be drawing in particular on the analysis and commentary on some of these issues by White and Arzi (2005) and Arzi (2004).

The questions I address in relation to these longitudinal designs are:

- How does intervention feature in different types of study?
- How is the time dimension conceived of in these studies?
- How can we characterize the different purposes of these studies?
- What are the causal or other logical relations explored in these studies?

- How do the designs relate to the purpose of the studies?
- How might we define quality in the methodologies underpinning these designs?

5.2 The Studies Represented in this Exploration

In this section I provide a brief overview of nine studies that explore learning in science over the long term. These cases are written to particularly highlight the variation in methodological features. The studies are drawn mainly from the two special issues described above, and most have been previously reported in the literature.

Adey (2005) described a program of research that explored the long-term effects of cognitive acceleration programs that have been devised and implemented in the United Kingdom. The first and best known of these is the Cognitive Acceleration in Science Education (CASE) program, which involved a 2-year intervention in school (Years 7 and 8) based on Piagetian tasks and designed to develop students' ability to process information. The study used a quasi-experimental design involving matched groups and compared students' scores on (a) cognitive development and science tests and (b) the General Certificate of Secondary Education (GCSE) public examination scores after 4 or 5 years. Despite there being no significant short-term differences, significant differences in examination scores generally, not only related to science, were claimed after this time between the experimental and control groups. The methodological issues discussed by Adey involved the potential contamination of control groups over the timescales studied, the problem of matching groups presumed to have parallel experiences over time outside the intervention, the validity of baseline tests, and the theoretical problem of characterizing the core feature of the intervention in order to establish the nature of the causal relationship.

Novak's (2005) study continued over 12 years and involved tracking the science understandings of children who had been exposed in Grades 1 and 2 to a 2-year intervention of audio tutorial science instruction in fundamental science topics. The intervention was based on Ausubelian notions of the importance of establishing a conceptual framework through which subsequent experience can be interpreted effectively. The children's subsequent science understandings were explored in interviews over 12 years and analyzed using a concept mapping approach specially designed to characterize the data. Comparison with a control group, consisting of children at the same school starting a year later and without the intervention, showed students receiving the instruction had more valid conceptual links and fewer invalid conceptual links, and that the difference between the groups increased over the 12 years. Methodological issues involved the problem of establishing matched groups, sustaining the data collection over this length of time (in this case, using teams of research students), and the problem of reducing complex data to comparable measures. The study was quasi-experimental in design but with many qualitative features in gathering and categorizing children's responses.

Johnson (2005) described a 3-year study of a chemistry curriculum in a secondary school in which student conceptions were traced through interview and linked back to features of a carefully designed teaching intervention consisting of a staged series of units focused on the idea of substance. The study traced the growth in student conceptions of core ideas and established patterns of change that could be linked to particular teaching sequences, such as the introduction of particle ideas. The correspondence of features of the intervention and advances in learning were established through matching qualitative advances in student conceptual understandings to particular features of the instructional sequence. These patterns of intervention and response were also used to generate insights into the fundamental nature of understanding phenomena. As an example, Johnson argued—on the basis of increased understanding of evaporative phenomena following the introduction of particle ideas—that access to a particle model of matter is a necessary aspect of imagining how liquid can be changed into gas of vastly greater volume. Methodological issues concern how best to represent the changes in understanding that occurred and how to relate these convincingly to particular aspects of a complex and extended teaching program.

Hubber (2004, 2005, 2006) followed six students' conceptions of light and seeing over 3 years of instruction in the upper secondary school. Students' changing conceptions were tracked to show a steady refinement in their understandings and to develop insight into the conceptual change process. Changes in student conceptions were found to be intimately related to the nature of the teaching sequence and strategies and showed a slow but steady learning trajectory. The links between the intervention and learning outcome was made, as with Johnson's (2005) study, on the evidence of patterns of student responses related to the learning sequence. The nature of the evidence in this case was different, involving multiple interviews and close tracking of students' responses to ideas discussed as part of the intervention. In the last year, students' mental models of light were probed as well as their views on the role of models in science. The results were able to be related to earlier models they held and also to the history of their understandings of light. Hubber established a link between students' epistemological sophistication and their learning; the students with a more advanced conception of the role of models in science also developed more sophisticated conceptions.

Helldén's (2004, 2005) research into pathways of student understandings was based on interviews of 28 students from ages 9 to 19. There was no explicit intervention; but students' ideas about particular ecological processes were traced, repeating the same probes over a number of years. The study demonstrated a gradual growth in the sophistication and increasing differentiation of students' ideas. Helldén found a noticeable consistency of imagery and metaphor for individual students that was repeated in interviews across the years. In many cases he was able to show this to be related to powerful childhood experiences that students at a later age were able to identify as underpinning their views. Thus, for instance, a student who consistently used the example of eggshells in compost to make sense of the recycling of autumn leaves each year reported later of his significant childhood experience in helping a neighbor with composting and his surprise about the disintegration of eggshells. In the later years of the study, students' thoughts about learning were probed, which shed light on the way they had framed their science explanations over the years. Students' epistemological perspectives, relating to the nature of their characterization of science learning and knowing, showed continuity over time while being very individual.

Holgersson and Löfgren (2004) further explored Helldén's ideas and followed students' explanations of three contexts for changes to matter, following a brief intervention in which students were introduced to molecular ideas. They traced the use of the molecular idea over time and found this occurred as a delayed effect and used more for physical science phenomena than for changes in biological material. Eskilsson (1999) used a similar design but interspersed interviews over a number of years with a managed intervention designed to promote molecular understandings and utilized explicit scaffolding in the interviews if students did not spontaneously mention molecules. He also found that different phenomena encouraged different views and that molecular ideas appeared gradually but only after the scaffolding had occurred in interviews in the previous year. Thus, molecular ideas are at first tentative and become established as students learn to use them with greater confidence. The methodological issues for both these studies relate to the complexity of the interaction between the intervention, the contexts, and in the variation in individual students' thinking. The argument for the early introduction of molecular concepts rests on the extent to which the molecular ideas can be convincingly shown to advance children's thinking.

Tytler and Peterson's (2004, 2005) study used twice-yearly interviews to track children's understandings and reasoning over the primary school years in relation to a number of topics and aspects of science. In preparing for the interviews, classroom sequences were planned and run with teachers, without pushing for closure on the ideas to be probed. Tracking student conceptions of evaporation showed children's ideas to develop slowly but with complex, contextually dependent pathways. A closer analysis showed that children's different approaches to explanations over time and their orientations to learning led to different conceptual pathways. A comparison of children's responses to the same investigative task 2 years apart demonstrated a number of ways in which science knowledge and reasoning are interdependent. The evidence in this study for complexity of student learning pathways was the comparison of categories of response over context and over time. The methodological task involved the identification of key categories of conceptual ideas and of reasoning that could convincingly be shown to interrelate in children's explanations.

Shapiro's (2004b) study began with an intensive investigation of student understandings of light as part of an upper primary unit over a few months. Six students were focused on, and their changing understandings and involvement in learning were explored in depth. The study was extended to follow these six participants' science ideas through five interviews conducted over 18 years. The study of one participant (Donnie) showed movement of ideas about light away from the original, scientifically acceptable understanding she held in primary school. A comparison of interview transcripts over time showed a consistency in Donnie's view of learning and the nature of science, and the nature of the way she constructed personal meaning in relation to light. The study, as with Helldén's, provides insight into long-term retention of ideas and the efficacy of school programs in establishing scientific ideas. It does not lead directly to policy outcomes but rather deepens the nature of the questions that might be asked of programs of learning in school science.

These nine studies vary considerably in a number of aspects, not the least being that the Adey and Novak studies followed a quasi-experimental design, whereas the other studies were substantially qualitative in their data collection and analysis and differed in their purpose, from researching specific interventions to more fundamental studies of student learning. The different purposes, logic of the claims, findings, and methodological issues for these nine longitudinal studies are summarized in Table 5.1. In the next section, the logic pertaining to the establishment of knowledge claims for the different types of design is analyzed in more detail.

5.3 Different Models of Intervention and Causal Effect

The purpose of this section is to identify the key differences in the studies and to examine in some detail the logic on which knowledge claims are made and supported by evidence. The intention of the analysis is to establish the sorts of principles relating to reliability and validity that these different designs illustrate. I will argue that (a) the restricted methodology that comprises the Gold Standard (i.e., RCTs) offers but one possible defensible pathway to establish justifiable knowledge claims and (b) a research enterprise based solely on this methodology could not duplicate the important insights yielded by these longitudinal designs.

White and Arzi (2005), in their discussion of longitudinal study methodology, pointed out the different degrees of intervention that these studies employ and the way time appears in the design. They make a distinction between studies that are experimental, leading to *conclusions* (about whether the intervention led to the hoped-for outcomes), and those that are descriptive, leading to *insights* (which emerge during the study and are neither stated beforehand nor explicitly tested as hypotheses). The studies described in this chapter provide examples of a continuum between these notionally pure categories. This section draws on these ideas but extends the analysis to consider in more detail the nature of the logic of causal or other relations that flow from these distinctions and how differences in this logic relate to the different purposes of the studies. The argument being advanced is that there are many facets to the problem space of how to effectively support science learning and that these nonexperimental designs can be seen as either (a) contributing to the necessary theory building that must precede the development and validation of evidence-based teaching and learning intervention, or (b) offering practical but limited trials of interventions that might be regarded as precursors to a more specifically generalizable comparative study. This echoes the point made by Shelley et al. (see Chap. 1) concerning the need to judge studies in the context of

Table 5.1 Profiles of the nine longitudinal studies

Study	Purpose	Logic	Findings	Methodological issues
1. Adey	Evaluation of near-and-far transfer of learning following a Piagetian/Vygotskian-based intervention	Intervention study. Quasi-experimental design with matched control groups, a variety of measurement instruments	A highly significant improvement in student scores, some years after the CASES intervention. Lack of clear evidence in other studies	Maintaining the integrity of the control groups over time. Maintaining cohort size
2. Novak	Evaluation of an elementary school intervention in terms of its influence on subsequent student learning in science	Intervention study. Quasi-experimental design with matched control groups. Use of concept maps to track student understanding	Significant advance in learning going into secondary school, for participants in the Grades 1–2 intervention, compared to a control group	Maintaining consistency of outcome measures over 12 years. Organizing repeated probes
3. Johnson	Design and evaluation of a teaching sequence over 3 years	Intervention study and exploration of learning trajectory. No control group. Changes in student understandings of substance are linked to show key progression points, and these are linked to particular aspects of the teaching sequence	Demonstration of effectiveness of units, recommendations for further refinement of years 7–9 chemistry teaching, and insight into the importance of particle ideas for developing understandings of substance	The complexity of establishing a tight causal link between aspects of the intervention and student learning
4. Hubber	Exploration of detailed effects on student conceptions, of activities in a teaching sequence on light	Intervention study and qualitative exploration of learning trajectory. No control group. Linking, through descriptive observations (using a range of probes), aspects of the intervention with changes in conception. Identification of student mental models as key determinants of conceptual understanding	Demonstration of effectiveness of conceptual change challenge strategies to shift conceptions. Insight into the importance of mental models for underpinning thinking	Complexity of analysis using triangulation of data. Establishment of argument comparing this to other possible interventions

5. Helldén	Exploration of children’s growth in understandings of key biological concepts over the long term	Nonintervention qualitative study. Establishing patterns of continuity and change in participants’ explanations. Exploring participants’ views of their own learning to make sense of their approaches to explanation	Demonstration of the explanatory power of Ausubelian ideas. Establishing the importance of early experience in framing understandings, and continuity in knowledge and beliefs over many years	Dealing with complexity in analysis. Generating a theoretical perspective
6. Holgersson & Löfgren	Exploration of patterns of children’s use of the molecule concept over time, for different phenomena, to understand the difficulties students have with the concept	Minimal intervention. No control group. Establishing links across time in children’s explanations, and differences between types of context	Demonstration of the relative difficulty of using molecular ideas for biological phenomena compared to physical phenomena. Establishment of patterns of use over time	Dealing with complexity of student responses in analysis. Factoring in the effect of the intervention
7. Eskilsson	Exploration of children’s use of the molecule concept with staged teaching interventions	Qualitative intervention study that linked student explanations to particular teaching interventions and traced a pattern of growth in using the molecule idea in different circumstances and with different degrees of scaffolding	Demonstration of the way molecular ideas are used following explicit challenge and scaffolding but with a time lapse. Exploration of patterns of use in different contexts	Dealing with complexity in analysis
8. Tytler & Peterson	Exploration of patterns of growth in conceptual understanding and reasoning over the long term	Activity sequences and interviews used to stimulate student perceptions rather than being evaluated. Qualitative analysis of patterns of change in ideas, and linking reasoning and student “narratives of the self” with conceptual growth	Demonstration of complexity of growth in conceptual understanding and refutation of previous presumptions. Insight into links between conceptual learning, reasoning, and individual student approaches to learning in science	Retention issues. Change in direction of research over time. Dealing with complexity of student responses in analysis
9. Shapiro	Exploration of changing science understandings over long time periods following formal teaching	Qualitative analysis of interviews conducted at repeated intervals over a long time period, following probes during formal elementary schooling	Development of a theoretical construct: “the learning person,” which elucidates the dimensions of learning. Insight into long-term outcomes of school science learning and the construction of personal meaning	Maintaining contact with participants over a long time period. Dealing with complexity in analysis

the wider research agenda and not limit ourselves to evaluation at the rollout stage. Each type of research design is governed by its own particular purpose. Each study considered here has different relations between aspects of the intervention and the intended learning dimensions, the type of evidence generated, and the way this is used to mount a convincing argument leading to conclusions. If we are to judge the quality of the findings in such studies, these relations need to be made explicit.

5.3.1 Contained Intervention with Intended Long-term Effects

One type of intervention and model of causal effect is exemplified by the Adey and Novak studies, where the intervention is targeted and restricted in time and the longitudinal data collection is for the purpose of exploring the ongoing effect of this intervention into the future. Figure 5.1 shows the relationship between the intervention and measures of outcomes for these studies.

The possible effects over time are many, ranging from decay (as in memory studies) to consolidation or building, or reversion as with the case of a study of Gauld (1986) who found that students 3 months after a laboratory intervention on electric circuits generated what seemed to be false memories of their observations to justify reversion to prior conceptions. In Adey's (2005) study, the improved learning over time is held to flow from improved processing ability; while in Novak's (2005) case, the interpretation (following Ausubel) seems to be that the intervention established a conceptual framework on which further experiences could build. In these studies, the guiding knowledge and learning metaphor is psychological with the intervention presumed to cause a lasting change in mental processes or conceptual structure thereby providing a platform for ongoing and increased effective learning.

The primary research question in this type of study relates to the effectiveness of the intervention, which in these examples is specific in nature and conforms to a particular theoretical position/explanation. Hence, the question is: *Did it work to improve understandings of the cohort over time?* The implied logic of the question is that the intervention is compared with the status quo of no intervention, implying the need for a control group and an experimental design. The outcome measures

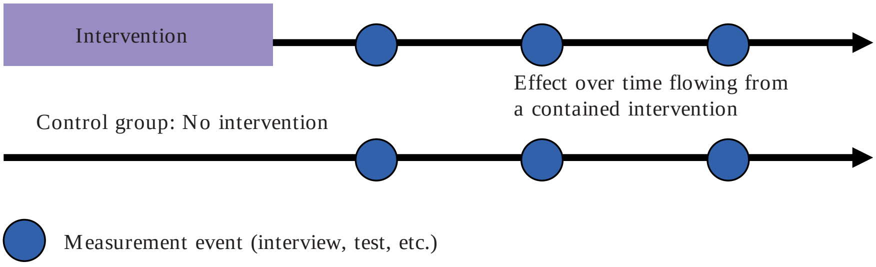

Intervention
Effect over time flowing from a contained intervention
Control group: No intervention
Measurement event (interview, test, etc.)

Fig. 5.1 Studying the effect over time of a contained intervention

are quantitative in nature to allow comparison; although, particularly in the case of the Novak study, these were constructed from interviews using interpretive methods and scoring rubrics.

A major issue for these studies is the selection of control groups that can be assumed to be exposed to the same experiences as the experimental group over subsequent years with the only difference being the intervention compared to a notional status quo. Novak (2005) adopted the device of treating the school class a year following the intervention as the control group. Adey (2005) selected control schools with similar characteristics to the experimental schools and discussed the practical difficulty of maintaining a realistic status quo in the control groups. He described a case where a control school, under an innovative principal, adopted some of the treatment features as a *back-door* arrangement. We found a similar problem with the School Innovation in Science (SIS) project (Tytler, 2005, 2007) in that schools would only agree to participate in the testing program on the promise they join the project in the following year. A number of schools borrowed the materials and initiated changes in the first year. The other problem we had was that, while we matched experimental and control schools on socioeconomic and student performance indicators, there was considerable variation between teachers, making comparison difficult at the class level.

The potential dangers of a Gold Standard research design in relation to setting up control groups are illustrated by a story of Adey (2005). He discussed an experimental study to test cognitive acceleration programs in science and mathematics in Finland, where students were randomly assigned to the experimental or control groups and bused to different locations. The researchers found significant differences immediately after the intervention, but this evaporated over 3 years with the control group lifting their performance above the national norm to that of the experimental groups. Adey argued that:

> A possible cause of this effect is that on average two-thirds of the students in any one class will have been in (an experimental) group, and this majority would have influenced the general level of thinking and intellectual interaction in the class in such a way as to ‘raise the game’ of the control students. (p. 8)

The control group was necessary in Adey’s and Novak’s research designs; otherwise, it would not have been possible to link subsequent developments in the experimental group’s thinking or performance unambiguously with the experience of the intervention. The general nature of both the intervention and the measures means that subsequent experiences and teaching would inevitably muddy the waters of such causal links. Such interference is traditionally referred to as relating to history (White & Arzi, 2005) whereby what happens to individuals subsequent to the intervention, including maturation and subsequent exposure to ideas and experiences, would affect their ongoing performance and make it impossible to unequivocally ascribe any outcomes specifically to the intervention. However, the logic is that details of subsequent variation in individuals’ learning are eschewed by the device of the cohort comparison, which is made using constrained and largely predetermined measures. The design is set up to confirm the success of an intervention and not to explore detailed relations between aspects of the intervention and aspects of subsequent learning. Thus, insights into more complex causal relations can only be inferred through the theoretical underpinning of the intervention. The meaning and interpretation of this theory is inevitably steeped in the history of studies that led to its development, and many of these studies (e.g., Piaget or Vygotsky) were profoundly qualitative in nature. Thus, the meaningfulness of the claims being made in such studies here are dependent on the insights generated by the research history underpinning the development of the treatment.

In fact, the need for a control group in these designs is really only necessary, in the logic of the experiment, because of the fact that the notional comparison group (with no intervention) can be expected to change in uncontrolled ways. Student learning is almost always of this nature given expected changes through maturation and experience. In a situation where we could reasonably expect no change in the status quo, there would be no real need for a control. However, the same is not necessarily true for adults. In a study of teacher change in the SIS project, science teachers' conformity with a framework of effective teaching and learning was tracked using a validated interview protocol and a score assigned for individuals at three points over 2 or, in some cases, 3 years. The scores improved appreciably, leading to a claim of success for the intervention, without a control group being factored into the design (Tytler, 2005).The logic of the design was a comparison of the experimental group with a notional group of teachers continuing to teach in ways consistent with the beginning behaviors of teachers in the study (as illustrated in Fig. 5.2). The benchmark or baseline value prior to the intervention is assumed to be the steady state for teachers based on the literature about teacher change.

The study of Holgersson and Löfgren (2004) is superficially similar in form to the Adey and Novak studies described above, involving the introduction of a molecular model and subsequent exploration of its use by students in three different material change contexts: physical, chemical, and biological. In their study, however, there was no comparison group but the evidential argument relates to the identification of patterns of response over time. They explored the particular ways individual children responded to the contexts and constructed meaning, the nature of the individual pathways, and differences between the children in the way their ideas developed over time. Thus, the logic of the intervention is not that of a carefully

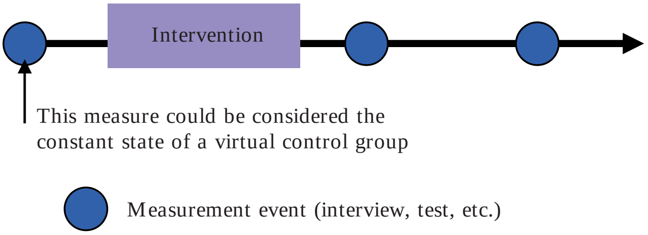

Intervention
This measure could be considered the constant state of a virtual control group
Measurement event (interview, test, etc.)

Fig. 5.2 Contained intervention with the prior measurement acting as a de facto control

structured program to be tested so much as a perturbation in children's experience introduced in order to trace the developmental progression of an idea and its dependence on context and on individual reasoning. The intervention was minimal, and there was no particular care within the design to account for the possibility that the children may have accessed other, uncontrolled experiences of molecular ideas. Thus, the effect of children's developing use of molecular ideas should be considered as resulting from an interaction of the original instruction and ongoing lived experiences rather than exclusively the instruction itself. The purpose of the study was not exclusively to generate conclusions on the efficacy of early introduction of molecular ideas—it was primarily concerned with the generation of insights into the influence of context in supporting children's use of the molecular idea and the differences in patterns of use between individuals. In studies such as this, the longitudinal design generates data on individuals that are rich in interconnections across time. This allows a detailed exploration of conceptual progression for individual children. Such data can provide more insight into growth or learning trajectories than simple comparisons with cohort control groups. The degree to which the evidential links are convincing is determined by the coherence of the patterns of developing understanding, the plausibility of the links to the intervention, and the consistency of the explanatory narrative over a number of individuals.

5.3.2 Long-term Intervention Contiguous with Effects

A second, distinct type of study is one in which the intervention and effects are both long term and contiguous in time, such that the relationship between them is explored. Eskilsson's intervention, which was more sustained than Holgersson and Löfgren's, was interspersed with interviews on a biennial basis. To an extent, it is interventionist in nature because of the researcher scaffolding children's responses as part of the design (Fig. 5.3). In a sense, we could regard these interventions and interviews as essentially a long-running, continuous event with probing occurring alongside it. Insights are generated through analysis of the patterns of change in ideas and the way these might link in time to particular features of the interventions. Again, the quality of the research rests on the construction of a coherent, evidence-based narrative that links patterns of student explanation and specific interventions across time.

This was also the case with the Hubber and Johnson studies. In Johnson's (2005) study, the intervention consisted of four curriculum units over 3 years, and interviews were used to probe students' developing ideas over time. Links were

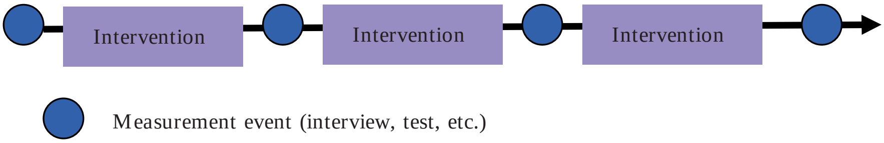
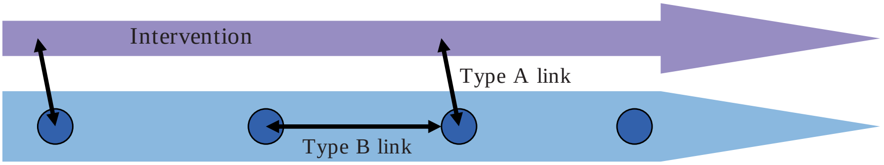


Fig. 5.3 Long-term intervention interspersed with measurement of effects



Type A link: Links between student learning and particular features of the intervention (e.g., linking advances in learning about evaporation with teaching about particles)

Type B link: Links between student's responses over time are made to establish the nature of growth patterns (e.g., tracing the pathways of change in student's thinking about evaporation)

● Measurement event (interview, test, etc.)

Fig. 5.4 Co-development of intervention and effect

made between student conceptual change and particular features of the intervention. Evidence from interviews and written responses was used to evaluate aspects of the curriculum and to generate insights into key points of conceptual growth. These insights provided ongoing indication of productive ways to sequence learning. Johnson also used interpretive analysis of the history of students' conceptual advancement to argue against the orthodoxy of conceptions being transcendent in nature and for a more evolutionary view of alternative conceptions. Thus, the study gave rise to conclusions concerning the effectiveness of the curriculum and insights into learning pathways that had both theoretical and practical implications. Figure 5.4 provides a model of the logic of this type of intervention study, with the unfolding intervention and measurement of learning gains interwoven in time and two types of evidential links being identified between intervention and learning, and between prior learning and current learning.

In the logic of Johnson's study, argument is constructed on the basis of Type A links between aspects of the intervention and particular changes in thinking and also on the basis of Type B links that identify patterns of change across time. In both cases validity is established through the construction of a history of individuals' and the cohort's thinking so as to identify key points of change. While not an experimental design, the study led to both insights into causal relationships and also conclusions about effective ways of sequencing teaching about the concept of substance. This has been the focus of further developmental work. This refined curriculum sequence would be open to testing on a large scale with a quasi-experimental design. Such research would give information about relative advantage of these new approaches to teaching chemistry compared to traditional sequences, but more detailed findings of cause–effect relationships would require a more fine-grained qualitative approach typified by Johnson's research.

Hubber's (2004, 2005) study is similar, but there is a more embedded relationship between students' learning and particular teaching strategies. The intervention is a structured curriculum, as in Johnson's case, but the probing is of different

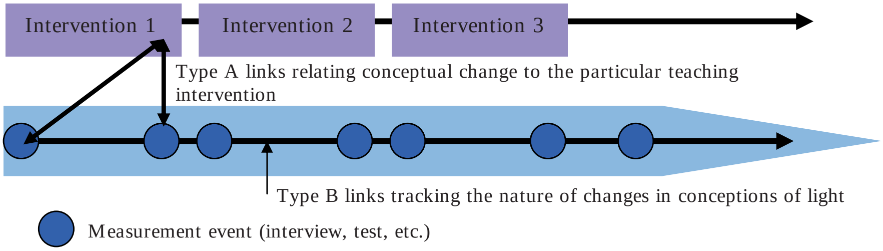

Intervention 1
Intervention 2
Intervention 3
Type A links relating conceptual change to the particular teaching intervention
Type B links tracking the nature of changes in conceptions of light
Measurement event (interview, test, etc.)

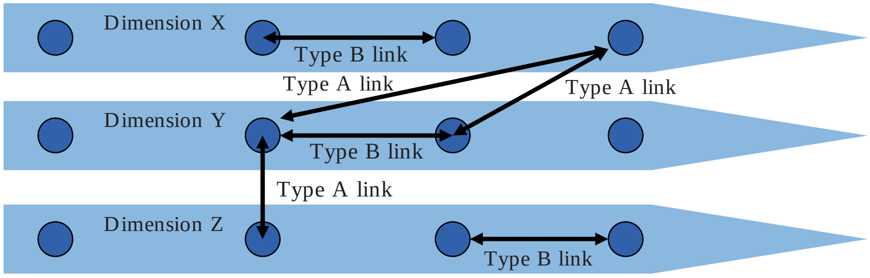
Fig. 5.5 A curriculum intervention dealing with a sequence of concepts of light

conceptions of light following strategies aimed explicitly at alternative conceptions. Figure 5.5 shows this variation.

Students' advances in thinking about a series of concepts concerning light is related to the close attention given in the intervention to challenging prior conceptions and support for conceptual change; the intervention and learning outcomes are more closely interwoven than in the Johnson study. The linking of student ideas across time mainly concerned growth in conceptual sophistication that may not be targeted in the later interventions. In the case of the mental models probing (Hubber, 2006), individual perspectives could be traced and related over time to each other and to conceptual understanding so as to build a theoretical picture of how learning occurs. While the study did not explicitly compare this conceptual change teaching design with other possible methods, it demonstrated that the method supported improvements in learning consistent with expectations based on conceptual change models and also provided insight into the nature of learning about light, including critical points, and the way it links with epistemological perspectives. There is, however, an implied comparison with traditional teaching methods in that many previous studies have shown the difficulty of achieving conceptual change in relation to these ideas about light, so that there is at least a plausible case that this study's methods are superior in achieving substantial advances in learning.

5.3.3 Minimal Intervention

A third type of longitudinal study—of patterns in learning and development—involves no explicit intervention, but changes in student learning or development over time are traced and used to provide insights into the nature of learning and growth generally. This type of study is not necessarily distinct from the two previous types since aspects of this charting of change can be found in studies involving interventions, as we have seen in the cases of Holgersson and Löfgren, and Hubber. Many of these studies involve probing a number of dimensions of learning for individual



Type A link: Links between different dimensions of student learning (e.g., reasoning and knowledge)

Type B link: Links between student's responses over time to establish the nature of growth patterns

● Measurement event (interview, test, etc.)

Fig. 5.6 Studies of learning over time, with no controlled intervention

students, and links can be made between these dimensions to explore learning as a more complex phenomenon. Figure 5.6 illustrates this type of design.

In these cases, Type B links, between subjects' understandings within a specific dimension at different times, are used to make sense of the nature of growth along particular dimensions of thinking (e.g., changes in their conceptual framing of particular topics, in their explanations, or in their reasoning). Type A links are those made between different dimensions of subjects' thinking and feeling (e.g., between explanation and memory of significant episodes or between reasoning and knowledge), and these are used to build insight into the nature of learning and development. These would be difficult to establish through short-term studies. Using this idea, we can divide the qualitative studies described in this chapter broadly into two types:

1. The establishment of general growth patterns or elements of continuity in thinking (Type B links):

   - Helldén's (2005) study confirmed the appropriateness of Ausubel's notion of progressive differentiation to describe the growth in understanding of his subjects, achieved by showing how individuals' ideas about the same phenomena are broadly consistent but become more refined and complex with the introduction over time (as part of schooling or other experiences) of new perspectives.
   - My own study (Tytler & Peterson, 2004) demonstrated more complex conceptual pathways over time than simple conceptual change models predicted. This particularly related to the way students' explanations were sensitive to contextual features of phenomena and differences in the way individuals approached exploration and explanation. These variations were nevertheless against the backdrop of a steady increase in sophistication of children's reasoning in approaching the same tasks repeated over time.

- Holgersson and Löfgren (2004) found relative consistency over time in individuals' use of metaphor in explanation but variation between individuals.
- Hubber (2005, 2006) traced a broad consistency over time in his students' mental models of light, and these were related to their conceptual change pathways.

2. The exploration of links between different dimensions of the participants' thinking (Type A links):

   - Helldén (2004, 2005) found that children's framing of explanations and generation of meaning was related to significant, early-life events. He later explored his subjects' views on their own learning and found this to be related to their conceptual explanations.
   - Hubber (2006) established links between students' mental models of light, their views of the nature of models in science, and their changing conceptual understanding.
   - Shapiro (2004b) traced Donnie's and other subjects' views of science and explored how the different dimensions of the learning person interact to frame learning.
   - My own study (Tytler & Peterson, 2005) linked students' growth in levels of reasoning with their growth in conceptual knowledge. We also established links between children's approaches to explanation with their broader orientation to learning and to school (Tytler & Peterson, 2004).

These studies, through their exploration of both Type A and B links, can be described as leading to insight—in White and Arzi's (2005) terms; and both types can provide evidence for productive ways to support learning that are capable of guiding action and policy. For instance, Helldén's (2005) work provides evidence of the importance of acknowledging and supporting students' anecdotal associations in framing conceptual learning, and my own work (Tytler & Peterson, 2004) demonstrates the importance of using a range of contexts to establish meaningful learning and the importance of linking reasoning with conceptual learning. These types of studies can refine or even transform our views of learning and, as such, are a critical part of that exploration of the problem space of learning science referred to previously; they can also lead to explicit suggestions for productive interventions.

There is a case to be made that the repeated interviews are themselves interventions. In my own study, the children's participation in classroom units prior to the interview, while not intended as an intervention to be evaluated, undoubtedly impacted their thinking. Thus, these studies could be thought of as involving minimal interventions designed to stimulate engagement with and focus on ideas rather than strictly no intervention at all. The intention of these studies is to establish learning pathways and identify individual variations in understanding and approaches to learning rather than establish absolute levels for a particular cohort. The presumption of zero impact of the intervention is, therefore, not a requirement. One must also acknowledge that probes of any sort, and particularly interviews, cannot be seen as neutrally eliciting some well-defined, inner-knowledge state but are importantly co-constructed as an interaction between the subject and the instrument.

Note that in minimal-intervention studies over the longer term, explicit teaching, school experiences, and experience more generally are part of the backdrop to the changes that are being tracked. It is the shape of the learning pathway over time and the interactions of dimensions of participants' conceptual ecology that are of interest; and causal relations are interrogated from within the evidence collected.

The logic of these studies is to identify patterns of change and patterns of interdependence of dimensions of science thinking to build a more complete picture of the nature of learning and development. Validity is established in these studies through relating events over the long term to show continuity in the thinking of individuals. A form of triangulation can be achieved by looking at recurring or similar features of interviews on different occasions, in some cases by prompting students to reflect on aspects of their own history of thinking. This is possible since the historical data are jointly constructed with the researcher. For these as with other types of design, methodological quality relates to the tightness of the evidential links that are made to construct and support the argument.

5.4 Validity and Implications for Method

Campbell and Stanley (1963) distinguished between the internal validity of a study, the extent to which the claims made with respect to the population being studied are unequivocally supported by the evidence collected and the logic of the design, and its external validity, the extent to which the results are generalizable to wider populations. White and Arzi (2005) discussed at some length issues of validity for longitudinal research designs; they argued that while Campbell and Stanley's analysis related at the time to studies using simple test base scores the ideas could be usefully applied to a variety of qualitative studies, such as those described in this chapter. They pointed out that true experimental designs are difficult and expensive to carry out in the field of education and discussed how Adey's (2005) and Novak's (2005) quasi-experimental studies—through their careful choice of control groups (countering validity threats associated with *selection*) and the nature of their repeat measurements—meet the criteria for internal and external validity. Indeed, Adey's story of the Finnish experimental study, described earlier, demonstrates the problem of *history* as a threat to validity (i.e., events over time that the experimental and control groups were exposed to varied in unexpected ways). This compromised the possibility of ascribing differences unequivocally to the nature of the treatment.

There are a number of threats to validity associated with the particular nature of long-term, quasi-experimental studies of interventions in education (these were addressed by the Adey and Novak studies). These are:

- Selection and history: It is difficult to set up a control group that can be defined as status quo over the longer term, given the shifting nature of schools and teachers over time and indeed the ethical problem of denying treatment if it is shown to be effective over the long term.

- Mortality: The threat posed by loss of subjects is particularly a problem over long-term studies because of the possibility that the loss could be selective, for instance, involving lower-scoring individuals thus rendering the score distributions unrepresentative.

For the Johnson (2005), Hubber (2005), and Eskilsson (1999) intervention studies, threats to validity were not a problem—if we assume that the selection of cases was representative of student populations at least in respect to the general patterns of ways they engage with science ideas. For these preexperimental studies not involving a control group, the threats of maturation (i.e., improvements occur simply through maturational development) and history are potentially a problem (White & Arzi, 2005). These threats were not such a problem for the contained teaching interventions of Johnson and Hubber, if time is short and if evidence claims are made on the basis of close association of treatment aspects and learning advances. For Eskilsson's longer term study, which involved no controls for school or outside exposure to particle ideas, both threats are issues. In each case, cross-validation using multiple data sources is essential to establish the strength of the argument.

The validity threat of maturation is not a significant problem for the minimal-intervention studies of Holgersson and Löfgren (2004), Tytler and Peterson (2005), Helldén (2004, 2005), and Shapiro (2004b) since it is the process of learning through both maturation and experience—the shape of the pathway and its dimensions—that is the focus of the research. For these qualitative studies, the external validity is also different than for the quasi-experimental studies since they make no statistical claims in relation to student scores. For qualitative studies such as these, validity is associated with terms like credibility, transferability, dependability, and confirmability (Guba & Lincoln, 1994) rather than generalizability in a formal sense. The validity claims should be judged on the extent to which the descriptions of methods, evidence, and nature of the outcomes is rich, clear, and convincing in the way these are linked together. In practical terms then, qualitative, long-term studies of learning:

- Should involve a variety of subjects that are in some sense representative of the population under consideration (e.g., a range in ability, socioeconomic background, gender).
- Can and should involve data generation that is sufficiently rich to allow cross-validation and linking of a range of dimensions and outcomes.
- Need to be designed so that interviews or other probes are sufficiently consistent across time to allow growth patterns to be unambiguously discerned.
- Need to clearly deal with the potential threat of history, associated with experiences that occur outside the control of the researcher.
- Need to be clear in linking various types of evidence to theoretical constructs.

Two problems with long-term studies are that understandings of the field can shift over the period of the research and insights generated can influence the aims and frameworks being used. Thus, our study (Tytler & Peterson, 2005) started as an exploration of patterns of conceptual growth but quickly moved to a broader framework as the implications of the findings became clear. The need for consistency of measures over time must be weighed against the need to be somewhat flexible and responsive in the research pathway if the value is to be maximized. One answer is to build in rich data-generation methods that partially anticipate what might be of ongoing interest.

White and Arzi (2005) discussed the varied threats to validity implicit in long-term studies and ways of dealing with these. They pointed out that cross-sectional studies, which are often used to establish trends in understanding, have their own validity problems, such as the untested assumption that cohorts separated in age are equivalent to the same cohort separated in time. Also, such studies can misrepresent learning pathways by smoothing out the complexity of individual growth in understanding that only long-term studies can reveal. They pointed out that the Adey and Novak quasi-experimental, long-term studies go well beyond mere numbers and simplistic conclusions and that long-term studies generally can lead to general principles concerning teaching and learning as well as providing insight into the complexities of individual learning pathways. They stated, “If we wish to understand important long-term changes and how they occur, our research must stretch over time and include different styles and methods” (p. 148).

5.5 Conclusion

A comparison of these longitudinal studies shows diversity in methodology, methods, and the logic by which conclusions are drawn or theoretical insights are established. Each type brings with it particular methodological issues. Key features that differentiate the design types are (a) the way intervention occurs and is conceived of—if it occurs at all—and (b) the relation through time between the intervention and various elements being traced in the study. The design types are distinct in the ways in which causal relations are thought of and argued and the type of findings that come out of the study. These different designs, while contravening the Gold Standard presumption, address different aspects of the problem space of student learning in science. It is argued that these methods can be entirely appropriate for exploring the educational effectiveness of interventions that take place over the longer term, for evaluating the longer term effects of limited interventions, or for exploring patterns of student learning over the longer term.

The evident value of the longitudinal nature of the designs rests on two related circumstances: first, that significant learning occurs through complex pathways over long time periods and, second, that the key effects of schooling are inevitably long term. In relation to the first circumstance, longitudinal design is unique in offering the opportunity for in-depth analysis of the nature of learning pathways and influences. In relation to the second circumstance, the design allows investigation of the long-term effects of particular interventions. These are different versions of time in thinking about learning, each illustrated by studies included in this analysis.

As a third circumstance, the longitudinal study offers significant advantages in probing learning processes in that it can chart changing knowledge against a background of continuity for individuals, thus focusing on the relationship between different dimensions of learning, such as knowledge and reasoning, or identity formation. Such studies can identify in some detail the way learning progresses with the confidence that patterns of growth and patterns of relation between aspects of thinking that are coherent over large time periods must be in some sense significant and not ephemera caused by the research intervention.

In terms of purpose, it is clear that (a) information about what might constitute a successful intervention (i.e., a conclusion) can come from nonexperimental studies and (b) there is a continuum between studies focusing on *insight* and those looking for *conclusions*. It has been argued that an experimental design utilizing control groups can provide a conclusion about what works best but cannot directly say why or if this particular intervention is optimal. For that to occur, a theoretical perspective is needed. The identification of detailed cause–effect relations in Adey's (2005) or Novak's (2005) case can only be inferred on the basis that the success of the intervention must verify the theory underpinning it. For more detailed information about cause–effect relations, we must look to the more complex designs aimed at generating insight. Thus, the question—*Compared to what?*—about the success of an intervention cannot be fully answered either by an experimental design (which by its nature restricts the number of possible factors under investigation) or by more complex designs (which do not set up direct comparisons but use an inferential chain of reasoning).

The generation of theory that will underpin planned interventions—in Adey's case, Piagetian and Vygotskian theoretical perspectives (involving probes of higher-order thinking and scaffolding) and, in Novak's case, Ausubelian perspectives (involving the establishment of a conceptual framework through which subsequent experience is interpreted)—involves bodies of work utilizing a range of methodologies—in these cases, qualitative methods. Thus, we can see that even the experimental studies discussed in this chapter indirectly depend on a range of designs to interpret their findings. The fact that the studies span the intervention-evaluation and the theoretical insight-generation ends of the educational improvement spectrum reflects the breadth of the research program that is needed to continuously advance our understandings and support of learning in science.

The quality of research in each of these studies relates to whether the causal logic proceeds in a defensible way. This chapter has attempted to identify the logical relations between intervention and outcome, cause and effect, and relationship patterns to clarify the different ways in which longitudinal studies can contribute to our understanding of how to support student learning in science. It has also attempted to demonstrate that—in terms of designing, implementing, and evaluating teaching and learning interventions—quasi-experimental design is not the only defensible pathway.

References

Adey, P. S. (2005). Issues arising from the long-term evaluation of cognitive acceleration programs. *Research in Science Education*, *35*(1), 3–22.

Arzi, H. J. (1988). From short- to long-term: Studying science education longitudinally. *Studies in Science Education*, *15*(1), 17–53.

Arzi, H. J. (2004). On the time dimension in educational processes and educational research. *Canadian Journal of Science, Mathematics, & Technology Education*, *4*(1), 15–21.

Black, P. J., & Simon, S. (1992). Progression in learning science. *Research in Science Education*, *22*(1), 45–54.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), *Handbook of research in science teaching* (pp. 171–246). Chicago: Rand McNally.

Eskilsson, O. (1999). Longitudinal study on lab work and 10–12-year-olds' development on the concepts of transformation of matter. In J. Leach & A. C. Paulsen (Eds.). *Practical work in science education* (pp. 229–243). Dordrecht, The Netherlands: Kluwer.

Gauld, C. (1986). Models, meters and memory. *Research in Science Education*, *16*(1), 49–54.

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (pp. 105–117). Thousand Oaks, CA: Sage.

Helldén, G. F. (2004). A study of recurring core developmental features in students' conceptions of some key ecological processes. *Canadian Journal of Science, Mathematics, & Technology Education*, *4*(1), 59–76.

Helldén, G. F. (2005). Exploring understandings and responses to science: A program of longitudinal studies. *Research in Science Education*, *35*(1), 99–122.

Holgersson, I., & Löfgren, L. (2004). A long-term study of students' explanations of transformations of matter. *Canadian Journal of Science, Mathematics, & Technology Education*, *4*(1), 77–96.

Hubber, P. (2004, April). *Students' changing conceptions of light over three years of instruction*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Vancouver, British Columbia, Canada.

Hubber, P. (2005). Exploration of Year 10 students' conceptual change during instruction. *Asia-Pacific Forum on Science Learning and Teaching*, *6*(1), Article 1, 1–27. Retrieved from http://www.ied.edu.hk/apfslt/

Hubber, P. (2006). Year 12 students' mental models of the nature of light. *Research in Science Education*, *36*(4), 419–439.

Johnson, P. (2005). The development of children's concept of a substance: A longitudinal study of interaction between curriculum and learning. *Research in Science Education*, *35*(1), 41–61.

Novak, J. D. (2005). Results and implications of a 12-year longitudinal study of science concept learning. *Research in Science Education*, *35*(1), 23–40.

Shapiro, B. (Ed.). (2004a). Developing understanding: Research on science learning and teaching over time [Special Issue]. *Canadian Journal of Science, Mathematics, & Technology Education*, *4*(1), 1–6.

Shapiro, B. (2004b). Studying lifeworlds of science learning: A longitudinal study of changing ideas, contexts, and personal orientations in science learning. *Canadian Journal of Science, Mathematics, & Technology Education*, *4*(1), 127–147.

Shymansky, J. A., Woodworth, G., Norman, O., Dunkhase, J., Matthews, C., & Liu, C. T. (1993). A study of changes in middle school teachers' understanding of selected ideas in science as a function of an in-service program focusing on student preconceptions. *Journal of Research in Science Teaching*, *30*(7), 737–755.

Shymansky, J. A., Yore, L. D., & Anderson, J. O. (2004). Impact of a school district's science reform effort on the achievement and attitudes of third- and fourth-grade students. *Journal of Research in Science Teaching*, *41*(8), 771–790.

Shymansky, J. A., Yore, L. D., Treagust, D. F., Thiele, R. B., Harrison, A., Waldrip, B. G., et al. (1997). Examining the construction process: A study of changes in level 10 students' understanding of classical mechanics. *Journal of Research in Science Teaching*, *34*(6), 571–593.

Tytler, R. (1998a). Children's conceptions of air pressure: Exploring the nature of conceptual change. *International Journal of Science Education*, *20*(8), 929–958.

Tytler, R. (1998b). The nature of students' informal science conceptions. *International Journal of Science Education*, *20*(8), 901–927.

Tytler, R. (2005). School innovation in science: Change, culture, complexity. In K. Boersma, M. Goedhart, O. de Jong, & H. Eijkelhof (Eds.), *Research and the quality of science education* (pp. 89–105). Dordrecht, The Netherlands: Springer.

Tytler, R. (2007). School Innovation in Science: A model for supporting school and teacher development. *Research in Science Education*, *37*(2), 189–216.

Tytler, R., Arzi, H. J., & White, R. T. (2005). Longitudinal studies on student learning in science [Editorial]. *Research in Science Education*, *35*(1), 1–2.

Tytler, R., & Peterson, S. (2004). Young children learning about evaporation: A longitudinal perspective. *Canadian Journal of Science, Mathematics, & Technology Education*, *4*(1), 111–126.

Tytler, R., & Peterson, S. (2005). A longitudinal study of children's developing knowledge and reasoning in science. *Research in Science Education*, *35*(1), 63–98.

United States National Research Council. (1999). *How people learn: Brain, mind, experience, and school*. Committee on Developments in the Science of Learning. J. D. Bransford, A. L. Brown, & R. R. Cocking (Eds.). Commission on Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

United States National Research Council. (2005). *How students learn: Science in the classroom*. Committee on *How people learn,* A Targeted Report for Teachers. M. S. Donovan & J. D. Bransford (Eds.). Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

United States National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Committee on Science Learning, Kindergarten through Eighth Grade. R. A. Duschl, H. A. Schweingruber, & A. W. Shouse (Eds.). Board on Science Education, Center for Education, Division of Behavioral and Social Sciences and Education, Washington, DC: The National Academies Press.

White, R. T. (1987, April). *The future of research on cognitive structure and conceptual change*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

White, R. T. (2001). The revolution in research on science teaching. In V. Richardson (Ed.), *Handbook of research on teaching* (4th edn., pp. 457–471). Washington, DC: American Educational Research Association.

White, R. T., & Arzi, H. J. (2005). Longitudinal studies: Designs, validity, practicality, and value. *Research in Science Education*, *35*(1), 137–149.

Chapter 6
An International Perspective of Monitoring Educational Research Quality: Commonalities and Differences

Richard K. Coll, Wen-Hua Chang, Justin Dillon, Rosária Justi, Eduardo Mortimer, Kim Chwee Daniel Tan, David F. Treagust, and Paul Webb

This chapter considers the notion of educational research quality and evaluation from an international perspective. We consider how and why these approaches differ from the US-based Gold Standard design (i.e., research based on randomized controlled trials [RCTs] mimicking third-stage drug trials; see Shelley, Yore, & Hand, Chap. 1). The Gold Standard is based on the assumption that RCT design alone, regardless of other factors, provides the desired quality.

We suggest here that the notion of quality in research and the mechanisms used to evaluate research quality are highly dependent on the overarching aim of education. To illustrate, the governments of many countries see education, especially science education, as a key component in economic progress and as a means of delivering on social services (Coll & Taylor, 2008). Hence, there are a number of reasons why we need to evaluate research quality in science education. We need to provide evidence that our science education regimes (and vocational education and training) do in fact produce outputs in terms of qualified people needed to

R.K. Coll
University of Waikato

W.-H. Chang
National Taiwan Normal University

J. Dillon
King's College London

R. Justi
Universidade Federal de Minas Gerais

E. Mortimer
Universidade Federal de Minas Gerais

K.C.D. Tan
National Institute of Education

D.F. Treagust
Curtin University of Technology

P. Webb
Nelson Mandela Metropolitan University

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

drive economic success. There is then the notion of accountability; the expenditure of taxpayer monies by government—especially in the area of education—is subject to much public scrutiny and often to criticism. There also is accountability to specific legalization in which governments require the education sector to *deliver* on education aims, such as scientific literacy. In each of these examples, we need to be as sure as we can that the research findings are trustworthy—to use Guba and Lincoln's (1989) term—or believable. New curricula and teaching and learning approaches often prove highly controversial (e.g., Bell, Jones, & Carr, 1995; Coll & Taylor; Matthews, 1994), and education stakeholders—including government—naturally want to see convincing evidence that costly educational interventions actually work.

6.1 Background

In some countries, there is awareness of the Gold Standard; in others, there are localized versions that set benchmarks for educational research quality nationally; in yet others, the situation is rather more idiosyncratic. As Shelley et al. (see Chap. 1) note, research quality and evaluation exercises like the Gold Standard exist in a complex sociopolitical environment, often driven or powerfully linked to legislation, for example, in the United States to the No Child Left Behind Act of 2001 (NCLB, 2002). The purposes and motives appear to be a sincere desire to enhance the quality of educational research in an attempt to equip politicians and bureaucrats to develop more effective public policies and more informed decisions about education and the expenditure of public funds. In the process, some approaches appear to place unreasonable trust in their evaluation criteria and underresourced researchers and institutions.

Even though the US-based Gold Standard in educational research seems to have a relatively low profile internationally, many nations watch how America grapples with issues in educational research with interest. Problematic educational issues are often common across international boundaries and, as in other matters, what goes on in the United States is often eventually exported—for better or worse! If there is little recognition of the Gold Standard for educational research internationally, it is interesting to consider how other nations monitor and facilitate educational research quality. It seems the way of monitoring quality educational research varies substantially, with most countries using some professional organization or an arm of government. In some cases, highly sophisticated, massively bureaucratic processes, such as the Research Assessment Exercise in the United Kingdom, are involved; in others, the process is less complex but not necessarily less rigorous.

Facilitation of research has been even more variable and less well considered. Many countries and universities genuinely want to move towards high-quality and high production of research. But this takes a considerable amount of time and resources, it occurs in an environment of financial constraint, and many academics face competing demands on their time (e.g., balancing teaching, research, and administration commitments). Developing an environment and research culture that supports highly productive, world-class researchers also necessitates considerable support; it means institutions must have appropriate infrastructure that supports and rewards researchers. In science education, there is an added complication. Many university academics began life as teachers and subsequently became involved in colleges of education or teacher training institutions. Professional experience as a teacher was highly valued, and research was of lesser importance—indeed, many academics involved in teacher training did not hold doctoral-level qualifications. However, over the years many teacher training institutions have been subsumed into universities, where research is valued more than professional practice. So in many university departments or schools of education, there are two subcultures: a cohort of staff with considerable teaching experience but limited research experience and frequently no higher research degree, and another more traditional cohort with a strong research culture.

We present here a view of educational research quality for seven nations: the United Kingdom, New Zealand, Australia, Singapore, Taiwan (Republic of China), South Africa, and Brazil. We detail how different nations monitor the quality of educational research and examine how research quality in each measures up against the Gold Standard. An analysis of the sociopolitical contextual factors that influence educational research internationally also is presented. Three central themes emerge from examination of these nations' efforts to evaluate research quality: first, the processes each nation uses to evaluate the quality of educational research; second, policies or governmental views that privilege particular types of research; and, third, the barriers, both internal and external, to quality research.

6.2 The United Kingdom: Monitoring Educational Research with the Research Assessment Exercise

For more than 20 years, the quality of academic research in the United Kingdom has been examined by a process known as the Research Assessment Exercise (RAE). The RAE is conducted jointly by the Higher Education Funding Council for England, the Scottish Funding Council, the Higher Education Funding Council for Wales, and the Department for Employment and Learning of Northern Ireland. RAE 2008 is the sixth in a series that began in 1986 (UK RAE, n.d.-a); its primary purpose is to produce quality profiles for each submission of research activity made by higher education institutions (HEI). The four higher education funding bodies intend to use the quality profiles to determine their research grant to the institution, which they fund with effect from 2009/10 (UK RAE, n.d.-f).

To give some indication of its scale, RAE 2001 involved submissions from 173 HEIs and examined the research output of almost 50,000 researchers. Some 2,598 submissions were reviewed by 69 panels, ranging from clinical laboratory sciences to sports-related subjects. The results directly affected the distribution of around 5 billion pounds (£) sterling of research council funding to HEIs (UK RAE, 2001). It is impossible to quantify just how much time is spent by individuals and institutions on collecting data and writing submissions, but there is no doubt that the RAE is an extremely high-stakes exercise. However, as the *Times Higher Education Supplement* (THES) pointed out, the RAE is a "competition in which Britain's cleverest people fight for hundreds of millions of pounds of public money. So it is bound to be a unsettling experience for many of those involved" ("Excluding the many", 2007, p. 12).

6.2.1 The Organization of the RAE

The system for RAE 2008 has changed from that adopted in 2001. Each academic discipline has been assigned to 1 of 67 units of assessment (UOA). The number of panels went down by 2, and the 67 subpanels work under the guidance of 15 main panels (UK RAE, n.d.-d). The main panels are responsible for (a) reviewing and endorsing the criteria and working methods to be used by the subpanels, (b) deciding on the quality profile to be awarded to each submission, and (c) communication and joint working with the other main panels. The subpanels are responsible for preparing draft statements of relevant criteria and working methods, undertaking the detailed assessment of submissions from HEIs, and making recommendations to the main panels on the quality profiles to be awarded for each submission. Each subpanel is composed of academics from HEIs and other bodies nominated by subject associations and other stakeholders. For RAE 2008, almost 5,000 nominations were received from nearly 1,400 bodies for the 1,000 panel positions (UK RAE, n.d.-c). The RAE 2008 Education Subpanel is composed of 20 people, most of whom are university academics. For each UOA, an institution must submit details of staff, research outputs, research students and studentships, research income, and research environment and esteem (UK RAE, n.d.-b).

6.2.2 What Counts as Educational Research?

The RAE criteria do not privilege any specific research approach; and in its guidance to HEIs, the Education Subpanel specifically noted, "research in education is multidisciplinary and closely related to a range of other disciplines with which it shares blurred boundaries" (UK RAE, 2006, p. 29). The panel provided three "illustrative and non-exhaustive" lists indicating what could be counted as research in education:

> a. Research which focuses on education systems including: pre-school, primary, secondary, … work-based learning, lifelong learning
>
> b. Research which addresses substantive areas such as: assessment, curriculum, teaching, pedagogy, learning, … social exclusion/inclusion and equity issues
>
> c. Research which employs qualitative and quantitative methodologies drawn from a variety of disciplinary traditions (including but not limited to applied linguistics, economics, … evaluation, literature review, critical theory, documentary analysis, analytic work) (p. 29)

6.2.3 Judging Quality: Originality, Significance, and Rigor

Each subpanel assesses the research submitted by the HEIs against a set of criteria. The judgments about quality are made on the institutional research outputs, and the panels are instructed not to "rate or score individual researchers" (UK RAE, 2006, p. 30). All outputs—books, chapters, papers, reports, and so on—are judged using three key criteria: originality, significance, and rigor (detailed description of these terms is provided in UK RAE, 2006).

Original work is expected to be work that does more than replicate other work or apply well-used methods to straightforward problems. Significant work is judged in several ways; for example, it breaks new theoretical ground, tackles important current problems, or creates an observable impact in its field. Rigor also may be judged in a number of ways but includes traditional qualities, such as reliability and validity; rigorous work demonstrates a sound background of scholarship and familiarity and engagement with the literature.

6.2.4 Quality Levels

Most academic staff entered for RAE 2008 have nominated four research outputs. The work must have been published between January 1, 2001, and October 31, 2007. Panel members examine each output, which is made available either electronically or by submission of hard copies. Each output is rated on a 4-point scale broadly described as: 1 = nationally recognized, 2 = recognized internationally, 3 = internationally excellent, and 4 = world-leading. To aid the panels in their assessments, each individual must provide a ≤150-word description of each research output indicating the contribution's theoretical perspectives, analytical and/or empirical methods used, original contribution to theory/methodology/policy/practice, and so on. An example of what would constitute such a summary of how the work met the originality, significance, and rigor criteria was provided to HEIs (UK RAE, 2006).

6.2.5 The Final Judgment

The subpanels report their findings for each UOA as a quality profile. The profile shows the proportion of the research outputs that has been judged to meet each of four quality levels or is unclassified. The panels also take account of "three overarching components of the submission – research outputs, research environment and indicators of esteem" (UK RAE, n.d.-e, para. 1). Two examples representing the quality profile of two institutions (universities X and Y) were published in the guidance for HEIs.

As might be expected, one had a higher proportion of 4* and 3* (40%) than the other (5%) and less 2* and 1* (55% and 85%, respectively; see UK RAE, n.d.-e).

The results of RAE 2008 will be published as a graded profile by the UOA for each submission. The actual evaluations are public information but typically are reduced to simplified tables in media reports.

6.2.6 Issues in the RAE Process

The impact of the RAE on institutions and their staff is impossible to quantify. However, the THES noted that "Unhappiness is set to peak, as people who have been performing and publishing research for years find out that they are not to be entered in the RAE" ("Excluding the many", 2007, p. 12). Some researchers were not entered into the RAE by their own institutions for fear of damaging the institution's score. The challenge facing HEI managers is to interpret the rules of the RAE game to their best advantage. As the THES explained, some decided to minimize the number of staff they submitted to get the best rating—although it was also noted that a department that gets a high RAE score on the basis of a small number of people is likely to be seen as lacking depth.

Following RAE 2001, the funding councils commissioned Sir Gareth Roberts to review the research assessment procedure. Roberts (2003) opened his report by noting that the recommendations of the report "constitute a radical overhaul of the Research Assessment Exercise" (p. 1). He mollified this somewhat by going on to say:

> They do not however represent a wholesale rejection of the RAE … [and that] the impact of the RAE upon UK research and its international reputation … has made us more focused, more self-critical and more respected across the world … in large part, by encouraging universities and colleges to think more strategically about their research priorities. (p. 1)

He noted, however, a series of concerns that had been expressed about the RAE process. These concerns were to do with the effect of the RAE upon the financial sustainability of research, an increased risk that games-playing would undermine the exercise, the burden of administration, the need to properly recognize collaborations and partnerships across institutions and organizations outside the higher education sector, a need to recognize all aspects of excellence in research, some concern over the disciplinary basis of the RAE and its effects upon interdisciplinary and multidisciplinary work, and a lack of discrimination in the current rating system—with the creation of a ceiling effect.

The results of the latest exercise will not be known until December 2008. Already the signs are that the RAE will be replaced by a metrics-based system (a research assessment formula), which might involve taking account of universities' research income and numbers of completed Ph.D.s. What is clear, though, is that the RAE, whatever its strengths, has caused "personal and professional pain to academics" (Lipsett, 2007, p. 6).

Are there other external and internal barriers to quality research in the United Kingdom? The answer to this question depends on who you ask! Presubmission results for the RAE indicate that several of the more ambitious universities are bringing in well-published faculty in an attempt to bolster their scores. In summary, UK researchers are evaluated by a number of national funding councils' assessment of four nominated research outputs. Evaluations are done via peer review, with no particular research methodology privileged, and the principal driving force being accountability of public funds. The main constraints on research quality within the RAE environment are financial sustainability of research, manipulation of the system by institutions, the administrative burden, and inhibiting collaborative, interdisciplinary, and multidisciplinary work.

6.3 New Zealand: The Performance-Based Research Fund

New Zealand is a small nation with 8 universities, 20 polytechnic institutions, and a very modest-sized science education research community. Much educational research in New Zealand is interpretive in nature; that is, exploratory case studies seeking to understand educational issues or to explore educational issues identified in surveys to greater depth and/or action research seeking to improve practices, often conducted by teachers doing postgraduate study and research projects as part of career or professional development. Such research projects are typically small scale and interventionist in nature with very few genuine experimental studies; almost no large-scale, national studies are conducted.

Educational research, including science education research, is principally conducted by faculty within university schools of education. Traditionally, teacher training was provided outside the university sector by colleges of education. From about 20 years ago, these colleges have gradually been subsumed into universities' schools of education, a process that is nearing completion. Nowadays, university faculty members represent a mixture of academics from a variety of backgrounds—as compared with a science department, which would be more homogeneous in terms of staff qualifications and experience. Relative to other disciplines (e.g., science), education faculty are underqualified (i.e., many without Ph.D.s), although there is a strong push to change this. There is virtually no knowledge of the US-based Gold Standard in New Zealand. Monitoring of research quality in the higher education sector is driven by the Performance-Based Research Fund (PBRF), a system instigated in 2004 that was essentially modeled off the UK's RAE and is managed by the New Zealand Tertiary Education Commission (NZ TEC, 2007).

6.3.1 Monitoring the Quality of Educational Research in New Zealand: The PBRF

The TEC (NZ TEC, 2007, para. 1) states that the "primary goal of the Performance-Based Research Fund (PBRF) is to ensure that excellent research in the tertiary education sector is encouraged and rewarded" with funding directly tied to what is purported to be excellence in research as determined by the PBRF peer review. However, no additional monies were allocated via PBRF; instead, the funding pool established a clawback of funds from the teaching component of higher education government funding. The PBRF model has three elements: reward and encourage the quality of researchers—allocated 60% of the fund; reflect research degree completions—allocated 25%; and reflect external research income—allocated 15% (2007). The major element, known as the quality evaluation, is held periodically. The first round was completed in 2003 (NZ TEC, 2004), and a partial second round was held in 2006 (NZ TEC, n.d.-b). The next full round is scheduled for 2012, unless there is a change of government and policy.

For the quality evaluation of excellence in research, faculty members prepare portfolios that show evidence of their performance across three elements: overall performance (characteristics such as their total output), peer esteem factors (such as awards, fellowships), and contributions to the development of new researchers and/or a vital, high-quality research environment. The evaluation consists of peer review assessment of the evidence portfolios submitted by higher education institutions, but prepared by individual faculty, to the TEC. Each evidence portfolio is then assigned one of six quality categories: A, B, C, C(NE), R, and R(NE). Faculty ratings are combined to produce departmental rankings, and departmental rankings are combined to produce institutional rankings—based on the proportion of active research staff during the assessment period in question.

To be assigned an A, it would normally be expected that the faculty member has produced research outputs of a world-class standard, established a high level of peer recognition and esteem within the relevant subject area, and made a significant contribution to the New Zealand and/or international research environments. To be assigned a B, it would normally be expected that the faculty member has produced research outputs of a high quality, acquired recognition by peers for research at least at a national level, and made a contribution to the research environment beyond the home institution and/or a significant contribution within the institution. To be assigned a C, the faculty member would have produced a reasonable quantity of quality-assured research outputs, acquired some peer recognition for research, and made a contribution to the research environment in the home institution. To be assigned a C(NE), a new or emerging (NE) researcher would normally be expected to have either completed a doctorate or equivalent qualification and produced at least two quality-assured research outputs or produced research outputs equivalent to a doctorate and at least two quality-assured research outputs—this category is only open to newly emerging researchers. Finally, to be assigned an R or R(NE) means the faculty member has produced no quality-assured research outputs, that is, no active research.

PBRF has been phased in over about 5 years with a proportion of monies originally allocated to student tuition subsidies being clawed back and reallocated to a competitive pool, then subsequently awarded based on an institution's PBRF ranking. PBRF set out to separate teaching and research components in order to better target funding to research, and the competitive element was intended to enhance research quality.

So is PBRF a measure of research quality and has it achieved one of its stated aims, that is, to enhance the quality of research (including education and science education research) in New Zealand? The truth is no one knows. Whilst it arguably has achieved its objectives in terms of transparency of funding for higher education, there is no evidence that it has enhanced research quality; indeed, this has never been evaluated. There is a vague statement from TEC suggesting it has been evaluated, but closer examination shows this had to do with the process of implementing PBRF and not about the quality of research (NZ TEC, n.d.-a).

Notwithstanding the lack of logic in ranking a modest number of institutions (8 universities and 20 polytechnics), PBRF, we argue here, is fundamentally flawed because of the size of the New Zealand research community and the competitive nature of the process. The community is so small that conflicts of interest are impossible to avoid; identification of research groups is simple (in many cases, only two research groups exist for specialty areas), and blind review hardly features (faculty have to identify their five best publications!).

In summary, New Zealand uses a variation of the UK's RAE in which the researcher is evaluated by means of national peer review. There are two principal driving forces: accountability of public funds and a desire to improve economic performance. The PBRF does not appear to privilege any particular research methodology and none is explicitly precluded. Constraints imposed are similar to that for the United Kingdom, with institutional game-playing rife, and a huge increase in administration, which takes time and money away from research. An additional constraint in New Zealand is the presence of two subcultures in educational research: one based on professional teacher educators who are somewhat less qualified and experienced in educational research, and traditional academics more experienced in research but with less professional experience.

6.4 Australia: The Australian Research Council and the Importance of Monitoring Quality

There are around 40 universities in Australia; many have faculty members teaching courses in education though not all faculties conduct and publish educational research. However, over the past decade or more, there has been increased pressure for faculty members to be seen as research-active, with the result that more researchers are seeking research funds from a fairly constant government source, the Australian Research Council (ARC). Research productivity across the universities—as measured by successful grant applications and publications in refereed journals—is uneven although there are many successful individuals and groups within those universities.

6.4.1 Educational Research Quality in Australia

The ARC is the major funding body and distributes grants for conducting research twice yearly. Two types of grants are available: Discovery Grants typically for pure research and Linkage Grants typically for applied research. The latter grants include funding by industrial partners that may be local, national, or international companies or education systems of various flavors; the ARC contributes an amount depending on the cash component from industry. An equivalent body, the National Health and Medical Research Council (NHMRC), provides funding for medical research and related disciplines.

The quality of successful applicants is maintained by peer review. In each discipline, there is an overview group, known as the College of Experts, in which membership is for 3 years with one-third renewed each year. Nominations to the college are sought from hundreds of organizations and advertised broadly. There are two kinds of peer reviewers: *Ozreaders* who are internationally recognized Australian researchers and *Intreaders* who are internationally based experts with international reputations about their research. A very large proportion of the latter are based overseas and, hence, not eligible for ARC support.

The ARC reserves around 15% of its funding in the large Discovery Projects scheme for early-career researchers. Applications for cross-disciplinary research constitute about a third of all submissions. The success rate of applications seeking support for cross-disciplinary research is little different from—and in the case of Linkage Projects, typically higher than—that obtained for single-discipline applications.

6.4.2 Monitoring the Quality of Educational Research in Australia

The quality of research is monitored by the ARC and NHMRC peer review process, and the success rate for grant applications in 2007 was around 20%. For monitoring the quality, to date there has been no consideration at all of the US-based Gold Standard. While colleagues attending US conferences may have heard of this term, it is otherwise unknown and unrecognized. However, on the horizon is a government initiative similar to that being conducted in the United Kingdom and New Zealand called the Research Quality Framework (RQF). The RQF will be implemented to determine whether or not the government's investment in research is paying off and to identify the best research in a particular field. This is intended to lead to better use of research funds and help potential Ph.D. students decide where to study a particular topic.

In September 2007, the Minister for Education, Science, and Training released the final specifications for the 2008 RQF, which was designed to identify the best research being conducted in terms of two principal dimensions—quality and impact—measured on a five-point scale. Research quality refers to the quality of original research including its intrinsic merit and academic impact (i.e., recognition by peers and effect on related discipline areas). Assessment of research quality will be based on aspects of an evidence portfolio, which is to include a context statement, a review of the best four research outputs (typically publications) submitted by each researcher within a research group, the body of work for all researchers in the group, and metrics associated with the body of work.

Research impact refers to the impact or use of original research outside the academic community and the extent to which it has led to social, economic, environmental, and cultural benefit for the wider community. Assessment of research impact will be based on an impact statement by the research group; the impact statement will include up to four case studies and also journal impact factors, a measure used to determine the importance of a journal to its field. Journal impact factors are included in the Journal Citation Reports (JCR) developed by Thomson Institute for Scientific Research (ISI). Its principal aim is to measure the frequency with which the average journal article has been cited in a particular year or period. The annual JCR impact factor is a ratio between citations and recent citable items published. Thus, the impact factor of a journal is calculated by dividing the number of current-year citations to the source items published in that journal during the previous 2 years; it can be considered to be the average number of times published papers are cited up to 2 years after publication. In August 2007, *The Australian* (Lane, 2007) presented a report of the top 3 Australian institutions in 21 research fields from 2002 to 2006. In education research performance, the top 3 universities were Curtin University followed by the University of Queensland and the University of Sydney, with impact factors of 1.86, 1.38, and 1.35, respectively.

As well as research outputs (publications), there are other ways to determine research performance, such as peer esteem, research inputs (research grants), and the research environment of the institution. The basic idea (rational flexibility) is that institutions with different approaches to research will suggest different contributions of these dimensions as being most relevant to them. As might be expected, views differ on how an RQF best can be introduced and executed to ensure that the maximum benefit flows back to the community (e.g., the RQF will be expensive to implement, be disruptive to researchers required to provide documentation, and reduce the time available for research and writing). Nevertheless, if the RQF can demonstrate the quality and impact of Australian research, there is an argument for it to drive the allocation of increased funding—new, not reallocated—for research.

Since these decisions were made (by the Howard Liberal government), there has been a change of government in Australia; and the Rudd Labor government, which had reservations in opposition, has categorically decided not to go ahead with the RQF. The government anticipates that, after a 12-month consultation period, a monitoring of quality of educational research will be put in place based on metrics in discipline-specific areas that will be less of an administrative burden for universities than the planned RQF. The government is nonetheless concerned about monitoring the outcomes of research and is looking to standardize processes using metrics that go beyond impact.

6.4.3 *Factors Impacting upon the Quality of Educational Research in Australia*

The major factor impacting upon the quality of educational research is the lack of funds with the implementation of the RQF resulting in an increasing number of researchers applying for funds from a constant dollar allocation (Mather, 2001). Furthermore, evaluations of ARC proposals are assessed against criteria that seem unreasonably high and place unreasonable demands on researchers and universities without associated levels of funding and support. Two other factors that are beginning to impact on the quality of educational research are (a) the way that resources are allocated within universities as each focuses on a best research output and (b) the decline of full-time research students from a decade ago. However, the majority of research in education is conducted by graduate students—most of who study part-time while earning their income as teachers, curriculum specialists, resource persons, or environmental educators—so there are low costs to the federal government for student scholarship living allowances in addition to those for the research training scheme, which is essentially a fees scholarship.

In summary, research quality in Australia is currently driven by a competitive funding model administered by the ARC. The principal driving force is accountability of public funds. An RAE/PBRF-type exercise had been proposed but was abandoned because of the change of government—instead, a discipline-specific, metrics-based system looks likely. Both applied and basic research is supported; and although fierce competition makes success rates low, a modest proportion is earmarked for emerging researchers, and interdisciplinary research has been encouraged in the past. Evaluation of research proposals for funding is based on both internal and external peer review. The principal constraint to research quality in Australia is the highly competitive nature of the funding system that inevitably means numerous good projects and potential new research spaces fail to gain funding. Subsidies for postgraduate student tuition research provide an alternate, low-cost way of funding quality research.

6.5 Singapore: A Broad-brush Approach

In Singapore, the National Institute of Education (NIE), an institute of Nanyang Technological University, is the sole teacher education and educational research institution; thus, educational research quality is essentially monitored by the NIE. As Singapore has no natural resources apart from its people and deep-water harbor, great importance is placed on developing the talents and capacities of the people to their fullest extent and inculcating lifelong learning to meet the challenges of the future. Therefore, the educational research conducted by the NIE is focused on maximizing the potential of students, teachers, and school leaders through (a) the improvement of teaching, learning, leadership, and organizational practices in schools; (b) leveling up of competencies and opportunities for all; and (c) development of educational innovations that are effective, sustainable, and scalable.

The NIE has two research centers: the Centre for Research in Pedagogy and Practice (CRPP) and the Learning Sciences Lab (LSL). The bulk of the funds given by the government to the NIE for educational research are administered by the CRPP and the LSL, so these two research centers oversee most of the educational research projects conducted by NIE faculty.

The CRPP was established in 2002 to drive educational research to inform educational policy and decision making as well as classroom practices. The CRPP has initiated more than 100 research projects over 5 years including six large-scale projects or panels on the secondary analysis of student database, cross-sectional study of pedagogical practices and student outcomes, classroom observations, classroom interactional analysis, classroom-based assessment and student performance, and a longitudinal study of students' institutional experiences, attainments, goals and choices, and pathways. These panels involve survey work, ethnographic studies, statistical analyses of demographic and achievement data, qualitative observation and discourse analysis (Singapore [SG], CRPP, n.d.).

The main research focus of the LSL is to determine how information and communication technologies can be used to promote engaged learning in classroom practices (SG LSL, 2006). The research conducted by the LSL is focused on the areas of new literacies, science as systems, mathematics and problem solving, knowledge-building community, and emerging research and pedagogies. None of the CRPP and LSL research projects, past and present, involve RCT. There is some local awareness of the US-based Gold Standard for educational research as many faculty members are active in research and are in touch with developments in the research arena; as well, they have links with fellow researchers and research centers in the United States and around the world. However, RCTs are not the quality standard for educational research in Singapore. In general, no research approach is favored more than others; the approach used must match the goals of the research.

6.5.1 Educational Research Quality in Singapore

Quality of research in Singapore has multifaceted meanings; it can mean that the research is important and useful to the stakeholders, for example, when the data obtained can be used in policy making by the Singapore Ministry of Education (MoE) or to improve the learning of students in the classroom. It can also indicate the use of sound research methodology or development of new ones, or generate new theories and insights into, for example, the teaching–learning processes and educational administration. Even the composition of the research team and the reputation of the team members can give some indication of the quality.

Measures that are generally used to judge the quality of educational research are the success of research grant applications, the quality of the publications resulting from research, and recognition and accolades afforded the research, publications, and researchers by international educational associations. As research grant applications are reviewed by local and overseas peers, reviewers act as gatekeepers to determine if the proposed research is well conceptualized and worth funding and, hence, whether it will be conducted in the first instance. Calls for proposal for education research in the NIE are sent via email together with the guidelines and application forms. The general evaluation criteria are:

- Significance of the Research
  - Does the project address an important problem?
  - Is there potential for new knowledge or value creation?
  - Will the concepts and methods proposed drive future research in this field?
- Approach
  - Are the conceptual framework, design, and methods appropriate to the aims of the project?
  - Are potential difficulties, limitations, and problem areas adequately identified and alternative courses of action considered?
- Investigator
  - Is the investigator qualified and competent to carry out this work?
  - What is the investigator's (and collaborator(s), if applicable) past record and present level of activity in this area?
  - What is the investigator's track record pertaining to previous grant(s)?
- Environment
  - Will the environment contribute to the probability of success of the project?
  - Are there any factors that might impede the progress of the project?
- Resources
  - Is the project cost-effective, and are the resources requested justified and appropriate?

Thus, ill-conceived and poor-quality research proposals are likely to be rejected at this stage. One may argue that the quantum of research grant awarded indicates the quality of the research, but there may not be conclusive evidence to prove the larger the sum of money awarded, the better the research is thought to be.

The conferences at which the research was presented, the journals and books in which it was published, and the citation indices of the publication are taken to indicate the quality of the research. Papers presented at prestigious conferences and published in first-tier journals have been carefully and stringently scrutinized by peer reviewers and judged to be of high standard. The reception by the international research community of the publication and presentations, for example, winning the best paper or best researcher award, is also tangible evidence of research quality. In addition, past performance is taken as a quality indicator of future success. Therefore, the quality of the publication output affects subsequent research grant applications as the grant evaluation committees use these indicators to determine if the researcher has a track record of producing quality publications and, hence, likely to execute the proposed research successfully.

6.5.2 Monitoring the Quality of Educational Research in Singapore

There is no single formal committee or organization that drives or monitors the quality of educational research in Singapore. However, the various research grant evaluation committees play a major role in monitoring the quality of the educational research before it actually starts; and these data and similar evaluations are used internally for personnel-related decisions. The quality of a faculty member's research is also considered during contract renewal and the promotion and tenure application processes. Publications and research grants are examined by NIE committees, and these documents are also sent to international referees for scrutiny to ensure that the evaluation conducted by the NIE is comparable to other educational institutions' standards. The majority of the committee members and international referees are able to provide dependable feedback on the quality of the publication and research; most value their professionalism and, as such, will highlight both the strengths and shortcomings of the faculty member's work. Finally, in the annual performance appraisals, the heads of the NIE academic groups (i.e., departments) assess the faculty member's work and highlight its strengths as well as areas for improvement in research.

6.5.3 Factors Impacting upon the Quality of Educational Research in Singapore

The unique integrated system within the MoE impacts and facilitates focus, funding, and utility of educational research. Positive factors that have an impact upon the quality of education research in the NIE are the importance placed on education in Singapore, the support and availability of adequate funding for research from the MoE, the extensive links the NIE has with established educational institutions and research institutes in the world, a centralized educational system, and the promotion of action research in schools by the MoE. The annual expenditure on education is about $7 billion SGD, a remarkable 11% of the total government expenditure (SG Ministry of Finance, 2007), to provide education for about 531,000 students in 354 primary, middle, and secondary schools, and about 125,000 students in 8 postsecondary institutions (3 institutes of technical education and 5 polytechnics) and 3 publicly funded universities (Koh, Tan, & Cheah, 2008; SG MoE, 2006). Adequate funding for educational research is made available to the NIE, especially to the CRPP and LSL.

In addition, Singapore's National Research Foundation (NRF) has allocated $50 million for research on interactive and digital media (IDM) in education from 2006 to 2010. This is to develop educational models and tools to equip students with the right skills and competencies to support Singapore's long-term vision of growing into a global IDM capital (SG NRF, 2007). Thus, funding for educational research is not an issue for researchers in Singapore. The NIE also collaborates with established institutions and research institutes around the world, for example, the SRI International, Northwestern University, University of Washington, Harvard University, and Futurelab. Such collaborations help to increase the quantity of research done in Singapore and improve the quality of research and the expertise of NIE researchers, especially junior faculty members.

Singapore has a centralized system of education and, with the support of the MoE, large-scale research projects across many or all schools are possible; researchers are able to access schools to work with the administrators, teachers, and students. Schools are also encouraged by the MoE to conduct action research into classroom practices and often welcome collaboration with researchers. Teachers are encouraged to enroll in postgraduate degree programs in education, which are heavily subsidized by the MoE. They acquire relevant research skills in their courses of study, which are important when teachers collaborate with NIE researchers in school-based projects. Being collaborators, they have a sense of ownership in the project and, hence, the motivation to contribute and see the project through. Teachers also can be seconded to the CRPP and LSL for 2–4 years to participate in research projects; they bring valuable knowledge of school context to the research.

Having a sole teacher education institution is advantageous as education research resources are concentrated in one institution. It also has its disadvantages in that NIE has to supply almost all the preservice, inservice, and postgraduate courses required by stakeholders. This translates to a heavy teaching workload for NIE staff and reduces the time available for research. A current shortage of academic staff in the NIE does not help matters for both teaching and research; and recruitment drives are unable to fill the vacancies, especially in science and mathematics education. The NIE also does not have a large number of eminent researchers with sufficient research expertise and experience to conceptualize and direct high-impact or large-scale research projects. Such researchers are considered to be able to attract other researchers and postgraduate students to the NIE and to contribute to the development of local researchers. Unlike many other countries, funding is not an issue in Singapore; lack of expertise and human resources for education research is the main constraint.

6.5.4 Trends in Quality of Educational Research in Singapore

Currently, the CRPP has six panel studies that provide baseline data on the status of teaching and learning in Singapore schools. Intervention studies (large and smaller scale) are being designed to enhance the learning and classroom practices based on the gaps and needs identified in the panel studies. It is uncertain if there will be RCTs as schools are generally reluctant to allow their students to be randomly allocated to classes. Student allocation to classes is usually based on subject combinations and performance in the previous end-of-year examination. Collaboration with local and overseas tertiary institutions and research institutes will become increasingly common as the NIE does not have all the necessary expertise and manpower to conduct research in some areas, for instance, in the development of IDM for education.

In summary, research quality in Singapore is evaluated by a national body, the NIE, with the principal driving force being a desire to remain economically competitive, with a strong focus on using modern technologies to improve learning. No particular methodology or type of research (i.e., applied or basic) is precluded; and evaluation is based on national peer review of competitive, publicly funded grant proposals, and publishing success in internationally renowned periodicals, conferences, and awards. The latter success factors also form part of the peer review of grant proposals.

6.6 Taiwan: Cooperation between the National Science Council and the Ministry of Education

In Taiwan (Republic of China), the few educational researchers who regularly attend international conferences are familiar with the US-based Gold Standard for educational research, but it has had no apparent impact on quality assurance. Educational research in Taiwan is monitored by the National Science Council (NSC). The NSC uses the following measures to judge quality of educational research outcomes: publishing academic journal articles, granting of patents, and technology licensing (TW NSC, n.d.-a).

The NSC is a cabinet-level organization within the Executive Yuan of Taiwan (Yuan is the highest administrative organization in Taiwan). The NSC was established in 1959 and assumed the role of promoting the development of science and technology as well as support for researchers. Its highest governing body is the Council Meeting, which comprises 8–14 members led by a minister and 3 deputy ministers. The NSC consists of eight departments, four offices, and three affiliated organizations. The eight departments are: Science Education, Humanities & Social Sciences, Natural Science, Engineering & Applied Science, Life Science, International Cooperation, Central Processing, and Planning & Evaluation (TW NSC, n.d.-b).

The departments of Science Education (DSE) and Humanities & Social Sciences are responsible for the promotion of educational research (TW DSE, 2006). Both departments are headed by a director and supported by a number of program directors and support staff. The program director is a professor who has been honored as an excellent research awardee. The program directors and their panels work toward several missions. The objectives of Science Education are to promote research in science education, elevate the quality of science education by enhancing the effectiveness and efficiency of science instruction, foster scientific literacy, and prepare students for future careers in science and technology.

Currently, the research programs in Science Education include mathematics education, science education teaching and teacher education, science education learning and assessment, information science education, and applied science education.

Each year, the DSE announces research topics for funding under three main categories: free research projects, mission-oriented research projects, and Ministry of Education-National Science Council Co-operation (MOE-NSC CO-OP) mission-oriented research projects. Researchers can submit single-topic research projects or cooperate with a group of colleagues to submit integrated research projects. The applicants apply to receive financial support according to their project objectives and tasks for human resources (including full-time and part-time research assistants, graduate students, postdoctoral researchers, etc.), equipment and consumables, hardware and software, local and international symposia and related expenses, and travel expenses (for conducting research overseas and attending international conferences).

The research project review process is illustrated in Fig. 6.1. The DSE provides a set of general guidelines to be followed by peer reviewers in determining the quality of the submitted proposals: qualification and competence of principal investigators; significance and uniqueness of the research; overall design and quality of the proposal and potentiality, originality, and universality of the research; project management and implementation (TW NSC, 2006). The weightings of these different dimensions are 40%, 20%, 30%, and 10%, respectively. After any necessary revisions, financial support is allocated by the panel.

A monitoring mechanism is embedded in the review process as well as through the whole research timeline. In addition to the general review guidelines, the DSE holds workshops regularly for the directors and program directors to promote and facilitate quality research. Funding policy and targeted research areas are announced, and examples of quality research proposals are presented. In the call for project proposals, the rationale of each targeted research area is explained along with specific requirements for each research area. These specific requirements usually identify critical components that must be provided in proposals. For example, research projects might aim to develop instruments; if so, the validation process must be described. For projects that aim to develop curriculum materials or programs, enactment experiments must be proposed to provide evidence of the effectiveness of those outputs; and the criteria applied to evaluate the outputs must be listed in the proposals.

In addition, conferences for NSC-supported research projects are held regularly to monitor progress and facilitate success. For completed projects, time is allocated for the principal investigators to present their work and discuss the research outcomes. For partially completed projects, the researchers report their progress in poster sessions. Special mentoring events are organized for new researchers, such as roundtable sessions to discuss research findings and paper drafts with experienced researchers. Although the monitoring mechanism was enacted early, only recently did the DSE require that evaluation criteria be included in group proposals.

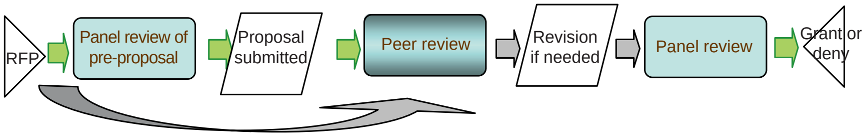


Fig. 6.1 Review process for research projects in Taiwan, as monitored by the National Research Council

Specific monitoring mechanisms are also in place for special research projects. Each year the DSE visits research sites of its mission-oriented research projects and the MOE-NSC CO-OP mission-oriented research projects so as to monitor their progress and quality. Conferences and workshops are held regularly for research project principal investigators to present their research outcomes. The directors, program directors, and their advisory panels provide formative comments and suggestions to the principal investigators based on the listed specific requirements and research quality in general. The main goals are to help the principal investigators by providing feedback, exchanging research experiences, and improving research competence.

6.6.1 Factors Impacting upon the Quality of Educational Research in Taiwan

The requirements in the call for research proposals and review guidelines provide a clear picture of quality research regarding topic and design and also pragmatics of procedures and practice. But some weaknesses regarding procedures and practices were identified in official documents: insufficient basic research funding support, insufficient technical and administrative support, rigid accounting and personnel regulations, rigid rules for vertical integration, lack of cross-discipline collaboration, and lack of team work among education researchers (Cheng, 2006; TW NSC, n.d.-a). From these obvious weaknesses, some actions were seen as necessary and are listed as future goals; for example, finding a balance between theoretical inquiry and practical research, funding, long-term and cross-discipline research, supporting emerging researchers, and international collaboration. Future trends in monitoring research quality by the DSE also were identified; examples include conducting SWOT analyses (a business model used in strategic planning involving an analysis of strengths, weaknesses, opportunities, and threats) to identify priorities, encouraging collaborative groups to work on educational research priorities, trying to link research to improved practice, and encouraging researchers to publish in international journals.

In summary, educational research in Taiwan is driven by a desire to promote science and technology for the purposes of economic development, with research quality evaluated by the NSC and MOE. Quality is driven by a competitive funding regime, with evaluation of proposals done by national peer review; both applied and basic research is supported. Lack of financial resources and balancing applied versus long-term, pure research are the main constraints to research quality in Taiwan.

6.7 South Africa: The National Research Foundation

A telephonic and email survey with a dozen senior academics and research administrators from seven institutions across South Africa revealed that only one respondent was aware of the US-based Gold Standard approach to educational research. This result is notable in that, apart from university-based academics in education faculties, this sample included directors of research and research development departments in universities and senior personnel in the quasi-governmental National Research Foundation (SA NRF, n.d.-b) and the Human Sciences Research Council (SA HSRC, n.d.) of South Africa responsible for education research funding. Although the data generated by this convenience sample are in no way generalizable (and in no way can be said to begin to meet the requirements of the Gold Standard!), the responses suggest that the notion of the Gold Standard in educational research has not penetrated academia in South Africa as of 2008.

However, there are other local measures used to judge and influence the quality of educational research in the South African context. These include the grant-awarding procedures of the NRF and HSRC, the rating of academic researchers and the accrediting of research journals for government subsidy purposes by the NRF, the peer review processes adopted by South African educational journals, and inputs from Higher Education South Africa (HESA) and the South African Department of Education (DoE).

6.7.1 Monitoring the Quality of Educational Research in South Africa

The NRF is the main body that monitors research quality in South Africa. NRF's strategic role is to support and promote research through funding, human resource development, and the provision of the necessary research facilities in order to facilitate the creation of knowledge, innovation, and development in all fields of science and technology, including indigenous knowledge and, thereby, contribute to the improvement of the quality of life of all people of the Republic of South Africa (SA NRF, n.d.-d).

Currently, the NRF has identified nine areas that provide the landscape for most research support activities of the organization and that constitute the Focus Area Programme (FAP) within which researchers may apply for funding, either as individuals or as teams. These focus areas currently are:

- Challenge of globalization: Perspectives from the global south
- Conservation and management of ecosystems and biodiversity
- Distinct South African research opportunities
- Economic growth and international competitiveness
- Education and the challenges for change
- Indigenous knowledge systems
- Information and Communication Technology (ICT) and the Information Society in South Africa
- Sustainable livelihoods and the eradication of poverty
- Unlocking the future: Advancing and strengthening strategic knowledge (SA NRF, n.d.-c, para. 4)

The purpose of the FAP is to build the knowledge base by directing research into areas of strategic importance to South Africa. The organization offers an open invitation to all rated and unrated researchers (assessment process described later) in the country to submit project proposals. The size of the grant depends on the proposal and the funding available. Successful proposals from unrated researchers are funded for up to 2 years, after which successful applicants are eligible for a further two cycles of up to 2 years. These cycles do not have to be consecutive. Rated researchers are funded for up to 5 years.

An important measure of educational research is NRF's peer-rating process of individual researchers. Reviewers are approached to assess an applicant's research by considering the applicant's submission and by appraising the quality of the applicant's research outputs over the previous 7 years. Reviewers are requested to write a concise appraisal by giving their opinion on the applicant's standing as a researcher, both broadly in the discipline and specifically in the research field. Applicants are evaluated as researchers in their own right, independent of research proposals; these evaluations are based on critical comments (both positive and negative) on research achievements and outputs (e.g., impact factor of the journals in which they publish, coherence of their work, and development of high-level teams).

Reviewers are asked whether they know the applicant personally; have previously encountered the applicant's work, for example, having heard aspects of their research presented at a conference; have read any of the applicant's work before being asked to undertake the appraisal or subsequently; or have cited any of the applicant's work in their reports. They are asked to discuss the impact, if any, they think the applicant's work has had on its specific research field and whether it has impacted on other fields. Their opinion is also solicited on the standing and appropriateness of the journals, books, conference proceedings, and other forms of research outputs that the applicant may have listed; and they are asked to estimate the applicant's current standing as a researcher.

Categories of researchers include leading international researchers, internationally acclaimed researchers, and established researchers. Leading international researchers are those who are unequivocally recognized by their peers as leading international scholars in their field for the high quality and impact of their recent research outputs. Internationally acclaimed researchers are those who enjoy considerable international recognition by their peers for the high quality of their recent research outputs. Established researchers are those with a sustained recent record of productivity in the field who are recognized by their peers as having produced a body of quality work—the core of which has coherence and attests to ongoing engagement with the field—and as having demonstrated the ability to conceptualize problems and apply research methods to investigating them.

There is also recognition of young researchers (normally younger than 35 years), who have held a doctorate or equivalent qualification for less than 5 years and who, on the basis of exceptional potential demonstrated in their published doctoral work, are considered likely to become future leaders in their field. A category also exists for persons (normally younger than 55 years) who were previously established as researchers and who are considered capable of fully establishing or reestablishing themselves as researchers within a 5-year period after evaluation. Candidates who are eligible in this category include black researchers, female researchers, those employed in a higher education institution that lacked a research environment, and those who were previously established as researchers and who have returned to a research environment from administrative postings or a nonresearch-oriented institution or assignment.

Accredited journals are peer-reviewed periodical publications that constitute recognized research output by meeting specified criteria, which allow authors to qualify for subsidy by the DoE. This subsidy is normally paid to the researcher's institution 2 years after publication of an article in an accredited journal. The development of a policy on accredited journals was driven by the imperatives for transformation of the higher education system, as contained in the 2001 National Plan for Higher Education (South African Government Gazette, 2003).

Only journals that appear in the Sciences Citation Index of the ISI, the Social Sciences Citation Index of the ISI, the Arts and Humanities Citation Index of the ISI, and the International Biography of the Social Sciences qualify as accredited journals for subsidy purposes. South African journals not appearing in the above indices but meeting a specified list of minimum criteria are also included in the accredited journals list. One of the minimum requirements for South African journals on the list is that they are peer reviewed; as such, this process plays an important role in maintaining the level of educational quality in published research. The DoE periodically samples South African journals and evaluates them, removing journals from the list that do not continue to meet the minimum criteria.

The purpose of the Policy and Procedures for Measurement of Research Output of Public Higher Education Institutions Act, from which the policy of accredited journals originated, is to encourage research productivity by rewarding quality research output at public higher education institutions (South African Government Gazette, 2003). The subsidy implications of the act have resulted in universities implementing their own systems to reward research productivity, which include varying levels of professional and personal financial rewards based on the subsidy amounts received from government.

6.7.2 Factors Impacting upon the Quality of Educational Research in South Africa

As in most countries, there are a number of factors that impact the quality of educational research in South Africa. One factor that appears to be an issue of growing magnitude is that of *creeping managerialism* and a move away from collegiality in universities, particularly in newly merged and historically black institutions. Stewart (2007) noted that, as South African universities have changed, a sense of malaise has emerged among many academic staff. The changes in higher education that have occurred around the world in an era of informational capitalism (i.e., practices tied to key policy debates about computerization and privacy) also have impacted South African institutions. Corporate-style managements have instilled a business-style culture and ethos with the expectation that academics must *pay their way* while, at the same time, South African academics face issues of deracialization, diversity, and a low skills base in an economy demanding high levels of skills (Stewart). All of the above have resulted in cost-cutting and profit-centered approaches, a rapid rise in student numbers, an increased percentage of part-time and informal faculty members, and a large number of demoralized employees—factors that cannot do anything other than impact on research output and quality in several ways.

Although a stated core component of academic life at most South African universities is knowledge creation, management's imperative to balance the financial books means that often only lip service is paid to this espoused value, resulting in impoverished infrastructure and fewer services that are needed to facilitate quality research as a sustainable endeavor (Stewart, 2007). This may sometimes result in tensions between the policies developed by university research and development offices, and their execution by deans and other administrators who are seeking to be more managerially and financially efficient, for example, by their academics teaching large numbers of students (Johnson & Cross, 2004). Nevertheless, recent measures taken by some universities to reward research output and raise the profile of research does appear to have impacted positively on research output (SA NRF, n.d.-a). However, while the NRF believes that the implementation of the accredited journal policy may have had an impact on the quality of research output, perfunctory discussions with some South African journal editors suggest that the need for academics to perform may have resulted in the temptation to publish (or attempt to publish) data and claims that could be better served by further investigation.

6.7.3 Trends in Quality of Educational Research in South Africa

In 2007, the NRF commissioned a research project entitled *An Audit and Interpretative Analysis of Education Research in South Africa: What have we learnt?* via email to the research offices of all South African universities. The project's terms of reference state that there is no clear consensus in terms of how educational research has developed in South Africa over the past few decades. This particular strategy was established as part of the NRF's commitment to advance education and includes an exercise that looks both retrospectively and forward to determine current gaps and strengths based on completed research and future research. The road map exercise is divided into two phases. The first phase is an audit and interpretative analysis of research undertaken over the last 10 years followed by meta-analyses and metasyntheses of groupings as suggested by the data generated. All types of research are considered—trying to find out what has been done; thus, it is something of a fact-finding exercise. Phase two is aimed at developing a research agenda through the identification of current research gaps and priorities and then brainstorming on future research needs and priorities for the next decade.

The NRF call for expressions of interest in this audit states that a variety of methods must be employed (e.g., a Delphi survey, horizon scanning and brainstorming workshops), while the scope includes the provision of an inventory of research projects (i.e., title, research questions–aims, and research findings) undertaken on and about education research in South Africa. This is to be done by interrogating, among others: peer-reviewed literature; grey literature, technical reports (consultancies), masters and doctoral theses; indigenous knowledge; lists of research projects by institutions; and the proceedings of national conferences. The research is also to be conducted using a multidimensional matrix including scale (large, case studies, etc.), type (epistemological, methodological), level (systemic, institutional, classroom, out of school, etc.), and area (discipline and major themes). Expected deliverables of the project are an inventory database, an interpretive analysis with conclusions, lessons learnt and recommendations, and possible identified areas of inquiry for further meta-analyses and metasyntheses. The audit is expected to be completed before the end of 2008 and, hopefully, will provide a much-needed resource in terms of educational research trends and set the scene for South African research in education in the 21st century.

In summary, research quality in South Africa is driven by NRF and HSRC grant awards and rating of academic researchers and accreditation of journals. The main purpose of educational research in South Africa is to improve the quality of life and raise standards of living via improvements to economic prosperity. To gain NRF funding, proposals must be of strategic importance consistent with these aims; but no particular methodology is favored. The main constraint to educational research in South Africa is the development of a corporate or business model for running higher educational institutions, along with cultural and racial issues specific to South Africa related to the Apartheid regime. There also are an overall low-skill base for education research and considerable financial constraints that are exacerbated by the cost-cutting corporate model of institutional management.

6.8 Brazil: Educational Research Quality Driven by Assessment of Postgraduate Qualifications

Educational research is conducted at universities in Brazil and, in particular, those that offer postbaccalaureate degrees in education. In 2008, there were 78 universities in Brazil offering masters and doctoral degrees in education and an additional 30 universities offering a masters degree in science and mathematics education. The quality of educational research is in effect regulated by an agency of the Brazil Ministry of Education that assesses the quality of postgraduate courses and degrees. Interestingly, the most important criteria used to assess the courses are the quality of the research outputs of the teaching professors. The main criteria used to assess the quality of postgraduate courses—and thereby educational research—indirectly connected to the faculty members involved are:

- Scope, coherence, and relevance of current research projects.
- Resources available for the development of the research projects.
- Academic level of the researchers.
- Participation of students as authors in papers and conference papers.

- Quality of the journals and conferences where the papers are published or presented.

There are also some additional criteria used to classify national and international journals and conferences—mainly the scope (international, national, or regional) and existence of a qualified board of reviewers. A more direct assessment of the quality of educational research occurs at two other levels: proposal of the project and publication of the findings arising from the project.

When a research project is proposed, the Brazilian researcher applies for a grant. Funding for research grants is provided by a federal government agency (Brazil Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, [National Council of Scientific and Technological Development], n.d.) and by some state agencies. In either case, the research project is analyzed and evaluated with respect to particular criteria. The main criteria used to judge research projects for funding are: (a) relevance of the theme; (b) coherence between the theoretical framework, methodology, and research question(s); (c) possible implications of the research; (d) adequacy of the research team; and (e) development of the methodology. These criteria address importance and utility, alignment of problem, question and approach, and procedural rigor (see Shelley et al., Chap. 1; Yore & Boscolo, Chap. 2).

When a paper is submitted for publication in a Brazilian journal, it is generally analyzed in the same way as for a good international journal (i.e., blind peer review) according to the same criteria typically used in the main international journals in education, such as the *International Journal of Science Education*, the *Journal of Research in Science Teaching*, and *Science Education*.

6.8.1 Monitoring the Quality of Educational Research in Brazil

The quality of educational research in Brazil is regulated by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (BR CAPES [Coordination for Improvement of University Level Staff], 2006), an agency of the Ministry of Education that has assessed all postgraduate programs since 1976. This is done in a triannual assessment process that is conducted by researchers invited from different universities who are recognized as excellent researchers.

Courses are classified on a scale of 1–7 based on criteria that are analyzed qualitatively and quantitatively. The main criteria are: (a) clarity and coherence of the proposed program, (b) quality of the researchers, (c) time required for students to conclude their studies, (d) quality of the master's theses and doctoral dissertations as well as of the papers published from them, and (e) social impacts of the proposed program. Any Ph.D. course that reaches level 6 or 7 is considered excellent and similar to the best such course internationally. One of the main criteria used to judge a course as 6 or 7 is the proportion of research papers published in international journals by the researchers who work in that particular program.

In the last assessment (2004/06), five programs attained level 6 (Universidade Federal de Minas Gerais, Pontifícia Universidade Católica do Rio de Janeiro, Universidade do Estado do Rio de Janeiro, Universidade Federal do Rio Grande do Sul, and Universidade do Vale do Rio dos Sinos e), 13 programs attained level 5, 41 attained level 4, and 19 attained level 3. (More details are available from http://www.capes.gov.br/avaliacao/resultados [in Portuguese].)

6.8.2 Factors Impacting upon the Quality of Educational Research in Brazil

As elsewhere, a number of local factors impact educational research quality in Brazil—some positive, some negative. One positive factor is the high academic level of researchers, many of whom completed a Ph.D. abroad or in one of the main Brazilian universities, which means that many researchers are well trained in educational research. A second factor is the interaction between Brazilian researchers and important international research groups in education. This has led to a significant number of publications in renowned international research journals. A third factor is the existence of the regulatory system built by CAPES, which drives research quality across the country. A fourth factor is that Brazil has experienced and capable researchers across different areas of education. A final factor is that competition for entry into postgraduate degrees in education in Brazil is stiff, meaning only very able students apply to masters or doctoral programs.

A negative factor is the difficulty in getting a grant to conduct research projects, mostly because there are so many researchers and so little money available for research. A second factor is the view held by some primary and secondary school directors about educational research, which they see as something that disturbs the students and so they are reluctant to allow educational research to be conducted in their schools. As elsewhere, another factor is the demands on researchers' time as a result of the weight and diversity of demands with university professors required to do research, teach (at least 8 hours per week), be involved in university administration, and coordinate extension projects in collaboration with schools and school systems in which the research results are disseminated.

6.8.3 Trends in Quality of Educational Research in Brazil

It is difficult to say what trends there are in educational research in Brazil in terms of research quality. It does seem that with the creation of the postgraduate assessment process there has been an improvement in research quality for a number of reasons: (a) more researchers now try to publish their work in international journals, meaning research must be of a standard acceptable to such journals; and (b) Brazilian journals have become more rigorous in refereeing papers submitted for publication, which has resulted in high-quality papers available in Portuguese as well as in English. Overall, it is probably fair to say that the quality of education research in Brazil is improving.

There are substantial differences between universities' levels and the research conducted in different parts of Brazil. In the southeast and southern regions (the richest states), some universities are excellent. Also, the quality of the educational research is comparable to good research conducted internationally because, for instance, such researchers publish internationally and have their Ph.D. courses classified as of international standard by CAPES. These researchers frequently gain grants for developing their projects and have developed joint research projects with international researchers.

In summary, in Brazil educational research quality is monitored by the assessment of postgraduate qualifications and particularly the stature of the academics in the postgraduate programs as judged via their research output in blind-peer-reviewed journals. No particular type of research is precluded, and the main driver of research quality is a desire for Brazil's educational institutions to be seen as comparable to others internationally. Enabling factors for quality research are a high proportion of academics with Ph.D.-level qualifications and significant international collaboration. Constraints include the competitive nature of research funding, the reluctance by schools to be involved in educational research, and the demands on time.

6.9 An International Perspective of Education Research Quality: Commonalities and Differences

As we noted at the beginning of this chapter, the demand for quality is consistent but the processes used to evaluate educational research quality seem to vary worldwide. The United States has focused on design (RCT) to influence quality while other countries have focused on importance, alignment, and rigor. It is impossible for us to know for sure whether or not the processes used in these seven countries are sound; the best we can do is examine the nature of the processes and make suppositions about the integrity of those processes by comparing them with the Gold Standard.

Whilst the processes do vary, they share some similarities. First, a key component of all evaluation mechanisms is the universal use of *peer review* in one form or another—in the form of expert panels, peer review of conference proceedings and presentations, and of journal articles or book chapters—and *peer recognition* in terms of awards. The assumption here is partially based on past performance as the single best indicator of future performance while the argument here is that such scrutiny eliminates weak studies. In some cases, this peer review is mostly internal (i.e., institutional or national, as in Taiwan); in others, it is mostly external (i.e., international, as in Brazil); and in most, it is a combination of the two (e.g., the United Kingdom and Singapore). Such processes, or at least some of them, may be compromised somewhat if the peer review is not blind. In many cases, the peer review is rather indirect. By this, we mean the research is evaluated as whole—rather than just the methodological soundness or rigor of the research. To illustrate, whether or not a given study is published in an international journal (a criteria in some places) is dependent on more than research rigor.

A highly topical issue (e.g., scientific literacy, inquiry-based learning, etc.) may get a kinder review than a more methodologically rigorous study that is deemed of less interest. Likewise in government-run funding regimes, what is deemed of national (e.g., economic) importance may be looked upon more favorably than something seen as obscure or of little practical worth. For example, in many evaluation systems (e.g., the RAE in the United Kingdom) the national significance or perceived value of the work is one criteria used in evaluation of quality. Here, such regimes may predispose funding bodies to value applied, pragmatic—indeed useful—research over more creative, blue skies research. The argument here is that if some research project is seen as in the national interest or political priority then research quality may be compromised.

A second key feature is the strong *link between research and funding*. What is deemed quality is funded and what is not is unlikely to be funded. At first sight, this seems eminently sensible and indeed forms part of the rationale for the Gold Standard. Why would we fund work that is not quality? However, as Ioannidis (2005) argued, there is strong potential for bias under such regimes: “The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true” (Corollary 5). It is worthwhile to note Ioannidis here is actually writing about biomedical research that would appear to meet all the requirements of the Gold Standard! In support of this notion, Antman, Lau, Kupelnick, Mosteller, and Chalmers (1992) argued that expert opinion can often be biased in favor of prevailing dogma. Such a stance and tunnel vision could thereby inhibit creative research and potentially reduce the likelihood of innovative, groundbreaking research that genuinely moves the educational agenda forward. An additional issue here is such evidence claims may predispose funding bodies toward particular research approaches such as experimentalist, quantitative studies that might appear to provide more convincing, numerical evidence.

A third point is that it seems for a variety of reasons relatively few countries are in a position to conduct, or support, *large-scale national projects* that might lend themselves to a research design consistent with the Gold Standard (except perhaps Singapore). One might well argue that Gold Standard research consistent with the most rigorous scientific research is impossible in education anyway (e.g., it is not obvious how double-blind evaluation of a teaching intervention is possible); but if we are not in a position to do such research, then the Gold Standard well may remain a dream internationally. Does this then mean educational research that does not meet the Gold Standard is not rigorous or credible? Without getting into a paradigm debate, we doubt it. In our view, the principle purpose of all education research is (or certainly should be) to improve teaching and learning. To claim this can only be achieved in a certain way, we suggest, is naive.

However, even if one accepts our stance here, we are left with a thorny problem, one that is key to the notion of the Gold Standard. How can we be sure that what works in a given study will equally apply in other educational contexts? We would argue you probably cannot—either in Gold Standard-type research or in other studies that do not meet the Gold Standard—in terms of research design at least (see Millar & Osborne, Chap. 3). Teachers will always have to adopt and adapt new pedagogies; even those based on the most rigorous research findings will require adaption to a given educational setting. We would further argue that good scientists and educational researchers constantly reassess their research agenda, striving to make each study better, reflective of other studies, and more credible and trustworthy. Shulman (1997) argued that no study, no matter how sophisticated, no matter how well funded and resourced, is without limitations; what really matters is to provide an audit trail of such limitations and to factor such limitations into the interpretation of our findings. The processes described earlier in this chapter suggest that internationally there are strong, ongoing efforts to do just that. We likely will never achieve perfection in educational research, but we must constantly improve and stretch to meet Gold Standards of our own—ones that reflect the complexity of educational contexts, recognize reality in terms of resources and funding, and incrementally enhance the credibility and trustworthiness of our research outcomes.

References

Antman, E. M., Lau, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: Treatments for myocardial infarction. *Journal of the American Medical Association, 268*(2), 240–248.

Bell, B., Jones, A., & Carr, M. (1995). The development of the recent national New Zealand science curriculum. *Studies in Science Education, 26*(1), 73–105.

Brazil Conselho Nacional de Desenvolvimento Científico e Tecnológico [National Council of Scientific and Technological Development]. (n.d.). *Homepage*. Retrieved July 4, 2008, from http://www.cnpq.br/english/cnpq/index.htm

Brazil Coordenação de Aperfeiçoamento de Pessoal de Nível Superior [Coordination for Improvement of University Level Staff]. (2006). *Homepage*. Retrieved April 17, 2008, from http://www.capes.gov.br/ [in Portuguese]

Cheng, Y.-J. (2006, Fall). *Enhancing the quality of group proposals in science education*. Paper presented at the annual workshops for Writing NSC Project Proposals held in several universities in Taiwan.

Coll, R. K., & Taylor, N. (Eds.) (2008). *Science education in context: An international examination of the influence of context on science curricula development and implementation*. Rotterdam, The Netherlands: Sense.

Excluding the many for the sake of the few. (2007, February 2). *The Times Higher Education Supplement*, p. 12. Available from http://www.timeshighereducation.co.uk/story.asp?sectioncode=26&storycode=207676)

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *Public Library of Science Medicine, 2*(8), e124. Retrieved from http://dx.doi.org/10.1371%2Fjournal.pmed.0020124

Johnson, B., & Cross, M. (2004). Academic leadership under siege: Possibilities and limits of executive deanship. *South African Journal of Higher Education, 18*(2), 34–58.

Koh, T. S., Tan, K. C. D., & Cheah, H. M. (2008). Science education in Singapore: Meeting the challenges ahead. In R. K. Coll & N. Taylor (Eds.), *Science education in context: An international examination of the influence of context on science curricula development and implementation* (pp. 283–290). Rotterdam, The Netherlands: Sense.

Lane, B. (2007, August 1). Our papers have influence. *The Australian Higher Education Section*, p. 22.

Lipsett, A. (2007, February 2). Casualties rise even before the RAE starts. *The Times Higher Education Supplement*, p. 6. Available from http://www.timeshighereducation.co.uk/story.asp?sectioncode=26&storycode=207665)

Mather, J. (2001). RQF blamed for spike in discovery applications. *Campus Review, 17*(40), 1–14.

Matthews, M. R. (Ed.). (1994). *Science teaching: The role of history and philosophy of science*. New York: Routledge.

New Zealand Tertiary Education Commission. (2004). *Performance-based research fund (PBRF) — Evaluating research excellence — The 2003 assessment*. Retrieved April 17, 2008, from http://www.tec.govt.nz/templates/standard.aspx?id=589

New Zealand Tertiary Education Commission. (2007). *Development of the performance-based research fund*. Retrieved April 29, 2008, from http://www.minedu.govt.nz/index.cfm?layout=document&documentid=7503

New Zealand Tertiary Education Commission. (n.d.-a). *PBRF Quality evaluation 2006 — Questions and answers?* Retrieved April 29, 2008, from http://www.tec.govt.nz/templates/StandardSummary.aspx?id=1206

New Zealand Tertiary Education Commission. (n.d.-b). *Performance-Based Research Fund — Evaluating research excellence — The 2006 assessment*. Retrieved April 29, 2008, from http://www.tec.govt.nz/upload/downloads/pbrf-full-report-2006.pdf

No Child Left Behind Act of 2001. Pub. L. No. 107–110, 115 Stat. 1425. (2002).

Roberts, G. (2003). *Review of research assessment. Report by Sir Gareth Roberts to the UK funding bodies*. Bristol, Avon, UK: Higher Education Funding Council for England.

Shulman, L. S. (1997). Disciplines of inquiry in education: A new overview. In R. M. Jaeger (Ed.), *Complementary methods for research in education* (2nd edn., pp. 3–31). Washington, DC: American Educational Research Association.

Singapore Centre for Research in Pedagogy and Practice. (n.d.). *Core research program*. Retrieved April 17, 2008, from http://www.crpp.nie.edu.sg/course/view.php?id=306

Singapore Learning Sciences Lab. (2006). *Research overview*. Retrieved April 17, 2008, from http://lsl.nie.edu.sg/overview.htm

Singapore Ministry of Education. (2006). *Education in Singapore*. Retrieved April 18, 2008, from http://www.moe.gov.sg/about/files/moe_corporate_brochure_2006.pdf

Singapore Ministry of Finance. (2007). *Budget highlights financial year 2007: Ready for the future, ready for the world*. Retrieved April 17, 2008, from http://www.mof.gov.sg/budget_2007/budget_speech/downloads/FY2007_Budget_Highlights.pdf

Singapore National Research Foundation. (2007). *Call for proposals*. Retrieved April 17, 2008, from https://rita.nrf.gov.sg/IDM/default.aspx

South Africa Human Sciences Research Council. (n.d.). *Homepage*. Retrieved July 4, 2008, from http://www.hsrc.ac.za/index.phtml

South Africa National Research Foundation. (n.d.-a). *Evaluation and rating facts and figures*. Retrieved April 20, 2008, from http://evaluation.nrf.ac.za/Content/Facts/factsfigures.htm

South Africa National Research Foundation. (n.d.-b). *Homepage*. Retrieved April 17, 2008, from http://www.nrf.ac.za/

South Africa National Research Foundation. (n.d.-c). *The NRF focus area programme*. Retrieved April 20, 2008, from http://www.nrf.ac.za/focusareas/

South Africa National Research Foundation. (n.d.-d). *NRF key performance indicator report 2003/04*. Retrieved April 17, 2008, from http:www.nrf.ac.za/publications/annrep/04/intro.htm

South African Government Gazette. (2003). *Policy and procedures for measurement of research output of public higher education institutions*. Retrieved April 17, 2008, from http://www.uct.ac.za/downloads/uct.ac.za/research/office/pubcount/307.pdf

Stewart, P. (2007). Re-envisioning the academic profession in the shadow of corporate managerialism. *Journal of Higher Education in Africa*, *5*(1), 131–147.

Taiwan Department of Science Education. (2006). *Mission*. Retrieved April 17, 2008, from http://www.nsc.gov.tw/sci/mp.asp?mp=1

Taiwan National Science Council. (2006). *Request for proposals*. Retrieved April 17, 2008, from http://www.nsc.gov.tw/sci/lp.asp?ctNode=1622&CtUnit=1224&BaseDSD=7 [in Chinese]

Taiwan National Science Council. (n.d.-a). *Homepage*. Retrieved April 17, 2008, from http://web.nsc.gov.tw/default.asp?mp=7

Taiwan National Science Council. (n.d.-b). *Organization*. Retrieved June 20, 2008, from http://web.nsc.gov.tw/ct.asp?xItem=14932&CtNode=3417

United Kingdom Research Assessment Exercise. (2001). *What is RAE 2001?* Retrieved April 17, 2008, from http://www.hero.ac.uk/rae/AboutUs/

United Kingdom Research Assessment Exercise. (2006). *Panel criteria and working methods: Panel K*. Retrieved April 17, 2008, from http://www.rae.ac.uk/pubs/2006/01/docs/k45.pdf

United Kingdom Research Assessment Exercise. (n.d.-a). *History of the RAE*. Retrieved April 17, 2008, from http://www.rae.ac.uk/aboutus/history.asp

United Kingdom Research Assessment Exercise. (n.d.-b). *Making a submission to the RAE 2008*. Retrieved April 17, 2008, from http://www.rae.ac.uk/aboutus/subs.asp

United Kingdom Research Assessment Exercise. (n.d.-c). *Panel members for the 2008 RAE*. Retrieved April 17, 2008, from http://www.rae.ac.uk/panels/members/

United Kingdom Research Assessment Exercise. (n.d.-d). *Panels*. Retrieved April 17, 2008, from http://www.rae.ac.uk/panels/

United Kingdom Research Assessment Exercise. (n.d.-e). *Quality profiles*. Retrieved April 14, 2008, from http://www.rae.ac.uk/aboutus/quality.asp

United Kingdom Research Assessment Exercise. (n.d.-f). *RAE 2008: Research assessment exercise*. Retrieved April 17, 2008, from http://www.rae.ac.uk/

Chapter 7
Considering Research Quality and Applicability Through the Eyes of Stakeholders

Denyse V. Hayward and Linda M. Phillips

Quality in educational research and practice has come under considerable scrutiny by policy makers in the United States. This scrutiny is due, in part, to a desire to develop and implement efficient and effective interventions based on scientific evidence and, in part, by concerns that investment in practices that lack adequate empirical support may drain limited resources. Consequently, there has been a move toward the adoption of the evidence-based practice (EBP) model and accompanying evidence hierarchies from medicine by policy makers and funding agencies as a means to evaluate the quality of education research and to allocate research funding. It is imperative for any discussion of the EBP model in education to know the model as it was conceptualized and implemented in medicine. Sackett, Rosenberg, Gray, Haynes, and Richardson (1996) described EBP in medicine as "the conscientious, explicit and judicious use of best current evidence in making decisions about the care of individual patients" (p. 71). Implementation of an EBP model in medicine involves five essential steps:

1. Convert information needs into answerable questions (formulate the problem).
2. Track down, with maximum efficiency, the best evidence with which to answer these questions—evidence may come from clinical examination, the diagnostic laboratory, published literature, or other sources.
3. Appraise the evidence critically (weigh up) to assess its validity (closeness to the truth) and usefulness (clinical applicability).
4. Implement the results of the appraisal in clinical practice.
5. Evaluate performance. (Greenhalgh, 2006, p. 2)

Adopting and implementing EBP requires that practitioners not only read research but also read the research at the right time and alter their clinical behaviors and the behavior of others in light of what they have found (Greenhalgh). Hierarchies have been developed to support practitioners' critical appraisal and trustworthiness of the research evidence. In evidence hierarchies that evaluate quantitative research designs, studies that conduct systematic reviews of randomized controlled trials (RCTs) and studies utilizing RCTs are at the pinnacle (Greenhalgh). Thus, the EBP model is appealing because it appears to offer objective criteria to determine

D.V. Hayward and L.M. Phillips
University of Alberta

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

best practice (Horner et al., 2005) since it allows for types and strengths of *evidence* to be differentiated.

There is considerable debate regarding the appropriateness and applicability of adopting EBP and the accompanying evidence hierarchies from medicine to education. Participants at the 2nd Island Conference discussed many of these issues, and the authors of Part I of this book discuss the implementation of EBP from a variety of perspectives. Our goal in this chapter is to highlight and discuss important concepts and issues raised by these authors as they relate to various stakeholders.

7.1 Evidence-based Practice—What Counts as Evidence?

Yore and Boscolo (see Chap. 2) began by situating the issues that are discussed in each chapter within the broader context of the shift toward EBP and legislation—Gold Standards in Education Research (Bush, 2002); No Child Left Behind Act of 2001 (NCLB, 2002)—for education research. This shift is described by the authors as a result of (a) ideological and political agendas to improve educational outcomes for all students and (b) skepticism regarding the quality, rigor, and effects of research effectiveness on student outcomes. Yore and Boscolo discuss the challenges that have resulted from misunderstandings or misinterpretations in the translation of legislation that has privileged quantitative methodologies and evidence hierarchies, in particular RCTs, rather than focusing on research designs (quantitative, qualitative, mixed methods) that are appropriate to answer particular research questions. Stakeholders at different levels of the implementation process will have differing but important perspectives regarding EBP that other stakeholders need to consider, address, and incorporate.

7.1.1 Educators, Employers, and Professional Bodies

Many of these stakeholders rightfully question whether EBP, like so many other practices of the past, is just the latest fad. Upon hearing that EBP challenges them to consider questions such as *How do you know that what you do works?* many teachers indicate that they regularly ask such questions because it is part of what constitutes good teaching practice. However, proponents of EBP state that what sets EBP apart is the emphasis on using scientific evidence to answer such questions rather than relying on expert opinion or past practice (Greenhalgh, 2006; Reilly, 2004). Proponents argue that by adopting an EBP model educators will be more able to critically appraise the benefits and risks associated with particular instructional methods, interventions in classrooms, and individual student contexts.

Problematic to the claims made by EBP proponents is the lack of consensus for the EBP model across any discipline, including medicine (see Beecham, 2004;

Greenhalgh, 2006; Odom, Brantlinger, Gersten, Horner, Thompson, & Harris, 2005). For many nonmedical disciplines, the conceptualization and underlying assumptions of the evidence-based medicine model are at odds with the conceptualization and reality of their practitioner–patient or teacher–student relationships. Beecham spoke to this issue as it relates to the discipline of speech–language pathology. She argued that speech–language pathologists (SLPs) understand their practice differently from that of medical practitioners. For SLPs, the establishment of equitable and collaborative practitioner–patient relationships is viewed as central to, and an important component of, the success of therapeutic goals. Thus, the EBP model adopted from medicine, where evidence focuses only on external, measurable variables, is problematic. Many of the variables that support success in a collaborative treatment context are neither external nor easily measured. Given that a large proportion of speech–language pathology practice occurs within educational contexts, Beecham's arguments are informative and insightful for educators.

Recommendations made by EBP proponents, however, are often presented as though there is consensus as to what counts as evidence and what sorts of evidence are better than other evidence (Johnston, 2005). Johnston noted that amidst the enthusiasm for EBP it is easy to lose sight of the fact that these assumptions are virtually untested when adopted by other disciplines, often left unstated, and most definitely arguable, as shown by Beecham (2004). With the existence of considerable and substantial debate within and across disciplines, it is reasonable for educators, employers, and professional bodies to be confused about why EBP should be adopted—given that the costs of such change are substantial for this particular set of stakeholders.

7.1.2 Policy Makers and Funding Agencies

In their zeal to be fiscally responsible, policy makers and funding agencies' stakeholders need to carefully weigh the available evidence that exists in the research literature across a number of disciplines that have attempted to adopt EBP from medicine. Legislation of a practice model that will have substantive human and financial costs requires a priori knowledge of known problems in the conceptualization of the particular model. It is clear from a variety of publications (see Graham, 2005), however, that conceptual clarity has not been achieved; unfortunately, practitioners and researchers with the least power to affect change in ill-conceived and poorly articulated policies are left to face the consequences.

7.2 Uptake of Research Evidence

Millar and Osborne (see Chap. 3) begin by citing comments made by Hargreaves (1996) that educational research has offered little to inform teaching practice over the past 50 years because research studies are noncumulative, produce inconclusive and contestable findings, and are of little practical relevance. This position appears to have some support amongst practitioners (e.g., Lijnse, 2000) who have expressed dissatisfaction with the lack of research evidence to support teaching. Millar and Osborne devote the remainder of their chapter to examining this research-to-practice issue within the context of EBP. Three actual examples of instructional approaches in the teaching of science that are cumulative and conclusive—and have substantive practical relevance—are presented. Although all three studies had significant impact for the schools in which the research was conducted, broader application in science teaching has not occurred for at least two of these approaches. EBP proponents would argue that the lack of broad impact relates to the weakness of the evidence these studies offer because none were conducted using RCT designs. However, Millar and Osborne examined such a claim and concluded that it would be difficult to justify the expense in human and material resources to achieve the same findings using a RCT methodology for the three examples cited.

The reluctance to engage in, indifference toward, or ignorance of research evidence for purposes of uptake is of considerable importance for all stakeholders. Although Millar and Osborne demonstrate that it is clearly not simply a void in the availability and accumulation of quality research evidence, as suggested by Hargreaves (1996), there is limited expectation on the part of practitioners and policy makers that relevant research exists and an even lower expectation that research is to inform policy and practice. Such perceptions persist at all stakeholder levels and must be addressed if we are to make advances.

7.2.1 Educators

Sweeping statements, such as those made by Hargreaves (1996), denigrating the relevance of education research have serious consequences. First, such comments permit educators and others to dismiss relevant research findings out of hand. Second, such comments diminish the significant advances made in literacy and science education research. Finally, once such disregard is permissible, it becomes even more difficult to convince educators that any model, including EBP, will improve circumstances. Many authors throughout this book have reported on, referred to, and mentioned relevant and important research in literacy and science education that has left each of us with a greater appreciation of how our individual research fits within the larger picture of education—a picture that differs little from other areas in the social sciences and humanities.

If, as proponents suggest, the EBP model holds promise in bridging the gap between research evidence and practitioner uptake for the field of education, then the question remains as to how educators are to develop the skills necessary to implement an EBP model in classrooms in order to take advantage of research-based evidence to teach particular content, grade, and developmental levels. Many articles, chapters, and books (e.g., Greenhalgh, 2006; Johnston, 2005; Reilly, 2004; Silagy & Haines, 2001) are devoted to outlining the skills practitioners across a variety of disciplines need to develop in order to implement EBP. For example, the following skills are offered by Reilly: (1) completing a course or online tutorial on EBP, (2) developing critical appraisal skills when reading research papers, (3) becoming skilled users of research to enable the application of scientific information in their day-to-day practice, and (4) developing questions related to day-to-day practice that can be answered using evidence-based research. Unfortunately, educators often find themselves having to undertake learning skill sets such as those described with minimal or no support from employers, professional bodies, or the government agencies mandating practice changes. Many educators question whether the time needed to learn new skills, often at their own expense, is worth it, if EBP will likely be replaced in an ever-changing political agenda.

7.2.2 Employers, Professional Bodies, Preservice Education Programs, Funding Agencies, and Policy Makers

EBP proponents advocate and purport that research conducted using RCT will improve research uptake in education practice; however, evidence from medicine and other health professions does not support this contention. Many examples exist where evidence from RCTs demonstrated that particular interventions are not beneficial and may even be detrimental, yet these interventions continue to be widely used (see Gillam, Crofford, Gale, & Hoffman, 2001, for *Fast ForWord* language intervention; Greenhalgh, 2006, for back pain; Phillips, Norris, & Steffler, 2007, for *Meaningful Applied Phonics* reading instruction). Odom et al. (2005) suggested that EBP proponents have ignored the issue of whether or not results from RCTs are positive.

Further, there is evidence showing that, while health care practitioners consider research to be important, research findings have little impact on their day-to-day practice (Brener, Vallino-Napoli, Reid, & Reilly, 2003; Metcalfe et al., 2001). Reilly (2004) found that practitioners tend to read the abstract, introduction, and discussion sections of research articles but feel much less confident about understanding methods and results sections. Yet, to conduct critical appraisals of the research literature, these are the very sections that educators need to understand. If such is the modus operandi amongst the health profession that have implemented EBP for a much longer period of time, then we must question whether we realistically can expect a different outcome in education.

Logemann (2004) pointed to yet another issue that impacts uptake of research evidence, that is, the focus on productivity in health care and educational institutions. A productivity model is at odds with EBP, which requires time to develop expert skills, acquire new knowledge, and read and apply evidence. Currently, the cost of developing expert skills is not included in funding models in health care (Reilly, 2004) or education, but is an important issue for these stakeholders to consider if the EBP model is to be adopted in education consistently and successfully.

7.2.3 Researchers

Uptake of research evidence by educators is a significant concern for researchers. Researchers can support not only practitioners but also audiences across all levels if, according to Johnston (2005), there is a concerted effort to (a) situate the research within the larger context of the problem being studied, (b) provide clear indications for educational practice, and (c) clearly explain the extent of any limitations or generalizability issues. Logemann (2004) also suggested that researchers take the lead by conducting systematic reviews of assessment and intervention strategies as a means to critically appraise and synthesize the research literature for specific issues. Such syntheses, according to Logemann, would be helpful to practitioners who have limited time and resources to access and examine the available research. However, this recommendation would mean examining studies across a much broader range of methodologies than is the current practice (Johnston). We would add that, unless issues of why practitioners do not use research in practice contexts are addressed by all stakeholders, no improvement in uptake of research information is likely to occur no matter how exhaustive or clearly written the information.

7.3 Misinterpretation of Evidence Hierarchies

Two chapters in Part II focus on demonstrating the limitations of the wholesale adoption of evidence hierarchies developed for medicine to determine strengths of evidence in educational research and the allocation of research funds. Alvermann and Mallozzi (see Chap. 4) highlight the contributions of qualitative and quantitative research perspectives to teaching and learning, while Tytler (see Chap. 5) presents evidence from longitudinal studies showing that RCTs can neither duplicate nor supplant important insights yielded by these designs. The important issue raised by these authors relates to policy implementation, where misinterpretations of particular research methodologies are sanctioned whilst others are discouraged and denied funding for research programs. The consequence of misinterpretation narrows not only the range of questions that can be researched but the type of information that will be available to educators to support teaching and learning.

7.3.1 Policy Makers and Funding Agencies

The appeal of RCT design is that it reduces bias and increases generalizability of results because treatment groups are equivalent and representative of the larger group with the exception of the intervention received. Even in medicine, where RCTs are considered the Gold Standard, problems exist in optimal implementation. Due to the expensive, time-consuming nature of RCTs, many studies are conducted with inadequate numbers of participants or too short a time frame (Greenhalgh, 2006).

She added that there are often hidden biases in RCTs that result from imperfect randomization, failure to randomize all applicable individuals, and failure to blind examiners to the randomization status of study participants. Exclusion and inclusion biases also limit generalizability of RCT findings. In education, individuals with learning or reading disabilities, low socioeconomic status, behavioral or attention difficulties, or from minority populations are often excluded. The *normal* participants in many RCT study samples will likely differ in important ways from students within a particular school or community thus confounding results and limiting generalizability (Montgomery & Turkstra, 2003). The heterogeneity of participant characteristics and individuals with low-prevalence disorders and disabilities—as is common in educational contexts—poses a significant challenge to RCT research designs, which are based on establishing equivalent groups and where relatively large numbers of participants are needed to achieve analytical power (Greenhalgh).

These are all important considerations that have been overlooked in the shift of emphasis to RCT designs to the exclusion of other designs in education. However, by far the most significant problem overlooked by the RCT shift in funding allocation is that RCT designs are *only* applicable to questions regarding intervention. RCTs are not appropriate to answer questions related to diagnosis, prognosis, motivation, preferences, or beliefs; examination of these important issues requires quantitative, qualitative, and mixed-method designs (Greenhalgh, 2006). Excluding or limiting the pursuit of these critically important issues goes directly against the purpose of the legislation.

7.4 High-quality Research Requires Adequate Funding Support

The penultimate chapter in Part II (see Chap. 6) offers a review of mechanisms used to evaluate quality in education research across seven nations: Australia (AU), Brazil (BR), New Zealand (NZ), Singapore (SG), South Africa (ZA), Taiwan (TW), and the United Kingdom (UK). The authors found that mechanisms were dependent on the overarching aim of education for each nation; these included: (a) accountability of public funds (AU, NZ, UK), (b) improvement in economic performance and quality of life (NZ, SG, ZA, TW), and (c) making educational institutions comparable to institutions internationally (BR). Aims across nations were similar to those in the USA; however, no particular research methodology was privileged by any of the seven nations.

All countries identified constraints in developing and conducting high-quality research programs. The range of constraints included: (a) lack of government-level financial support resulting in numerous high-quality projects failing to be funded, administrative burden, and legislative demands (AU, BR, NZ, ZA, TW, UK); (b) lack of expertise and human resources to conduct research (SG); (c) cultural and racial issues related to the apartheid regime (ZA); and (d) reluctance by schools to be involved in educational research (BR).

The international survey revealed a clear commitment to quality in educational research but a consistent lack of funding to support high-quality research programs. These issues require the attention of policy makers and funding agencies.

7.4.1 Policy Makers and Funding Agencies

Chapter 6 by Coll and colleagues speaks to an international commitment to the application of quality indicators that represents rigorous application of research methodologies appropriate to answer the particular questions. Such indicators serve as guidelines for (a) researchers designing and conducting research, (b) policy makers and funding agencies evaluating the believability of research findings, and (c) educators determining the usability of research findings (Horner et al., 2005).

All seven nations achieve high-quality research without an emphasis on particular methodologies. In fact, Coll and colleagues show that relatively few countries are even in a position to conduct large-scale projects that might lend themselves to RCT research designs. Additionally, the expense of such studies would be problematic for the majority of nations. It is clear across all nations that the lack of financial resources available from government and funding agencies impacts both development and implementation of high-quality research programs. If policy makers and funding agencies are serious about committing to improving educational outcomes for all students, then increased financial support for research programs, including RCTs, is needed.

7.5 Conclusions and Implications

Berliner (2002) proposed that scientific research in education is not a hard science—such as medicine, chemistry, and biology—but it is the hardest-to-do science. Educational researchers conduct scientific research under conditions that physical scientists would find intolerable. They face particular problems and must deal with local conditions that limit generalizations and theory building—problems that are different from those faced by the easier-to-do sciences of chemistry, biology, and medicine. Mandating EBP has a significant impact on stakeholders at all levels. When there is less than optimum understanding and acceptance of new practice models, consistent and successful implementation is seriously challenged.

One of the two prominent issues raised by the authors in Part II is the appropriateness of the wholesale adoption of an EBP model and accompanying evidence hierarchies developed for medical practice to educational practice. The assumptions of the EBP model are virtually untested when adopted by other disciplines, frequently left unstated, and most definitely arguable (Johnston, 2005). The potential danger of focusing more or less solely on EBP is that it leads to disproportionate emphasis on the tools of experimental design rather than the specific questions that need to be answered (Montgomery & Turkstra, 2003). Greenhalgh (2006) concurred, stating:

> [W]hen applied in a vacuum (that is, in the absence of common sense and without regard to the individual circumstances and priorities of the person being offered treatment) the evidence-based approach to patient care is a reductionist process with a real potential for harm. (p. xiii)

A unidimensional focus on funding RCTs in intervention research in education is misinformed. By adopting such a position, the implication is that only intervention studies are needed to support teaching and learning. Studies engaged in diagnosis, screening, prognosis, and motivation—all of which most stakeholders consider imperative to the success of both teaching and learning—could not be conducted since RCT is an inappropriate methodological choice. We propose that if policy makers and funding agencies had enacted the five essential steps to implement evidence-based practice then many of the problems in adopting the model in education may have been preempted.

The other prominent issue concerns lack of uptake of research evidence in educational practice. This is a complex issue with a variety of reasons posited, including: (a) practitioners claim that there is a lack of any research to support practice, (b) research participants or treatments do not represent the reality in everyday practice, and (c) lack of time to access research evidence. The acknowledgment that educational practice functions primarily as a productivity model, which is at odds with the EBP model, is a significant consideration for all stakeholders since the development of these EBP skills is not included in funding models. We suggest that government policy is also more closely aligned to a productivity model, which is also at odds with the mandated legislation.

Despite the initial difficulties, we strongly believe that stakeholders in education have the opportunity to be leaders in developing an evidence model and accompanying hierarchies. Such developments within education that adequately address the types of research that best take account of the complexities of conducting educational research and the numerous challenges faced by educators in the uptake of research evidence are necessary for and fundamental to the education of our nations' children.

References

Beecham, R. (2004). Power and practice: A critique of evidence-based practice for the profession of speech-language pathology. *Advances in Speech Language Pathology*, *6*(2), 131–133.

Berliner, D. C. (2002). Educational research: The hardest science of all [Comment]. *Educational Researcher*, *31*(8), 18–20.

Brener, L., Vallino-Napoli, L. D., Reid, J. A., & Reilly, S. (2003). Accessing the evidence to treat the dysphagic patient: Can we get it? Is there time? *Asia Pacific Journal of Speech, Language and Hearing*, *8*(1), 36–43.

Bush, G. W. (2002, November 5). *Statement on signing legislation to provide for improvement of federal education research, statistics, evaluation, information, and dissemination, and for* *other purposes*. Retrieved February 14, 2008, from http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname = 2002_presidential_documents&docid = pd11no02_txt-21.pdf

Gillam, R. B., Crofford, J. A., Gale, M. A., & Hoffman, L. M. (2001). Language change following computer-assisted language instruction with *Fast ForWord* or *Laureate Learning Systems* software. *American Journal of Speech-Language Pathology*, *10*(3), 231–247.

Graham, S. (Ed.). (2005). [Special Issue]. *Exceptional Children*, *71*(2).

Greenhalgh, T. (2006). *How to read a paper: The basics of evidence-based medicine* (3rd edn.). London: Blackwell BMJ Books.

Hargreaves, D. H. (1996). *Teaching as a research based profession: Possibilities and prospects* [The Teacher Training Agency Annual Lecture]. London: The Teacher Training Agency.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*(2), 165–179.

Johnston, J. R. (2005). Re: Law, Garrett, & Nye (2004a). The efficacy of treatment for children with developmental speech and language delay/disorder: A meta-analysis. *Journal of Speech, Language & Hearing Research*, *48*(5), 1114–1117.

Lijnse, P. L. (2000). Didactics of science: The forgotten dimension in science education research. In R. Millar, J. Leach, & J. Osborne (Eds.), *Improving science education: The contribution of research* (pp. 308–326). Buckingham, UK: Open University Press.

Logemann, J. (2004). Evidence-based practice. *Advances in Speech Language Pathology*, *6*(2), 134–135.

Metcalfe, C., Lewin, R., Wisher, S., Perry, S., Bannigan, K., & Moffett, J. K. (2001). Barriers to implementing the evidence base in four NHS therapies: Dietitians, occupational therapists, physiotherapists, speech and language therapists. *Physiotherapy*, *87*(8), 433–441.

Montgomery, E. B., Jr., & Turkstra, L. S. (2003). Evidence-based practice: Let's be reasonable. *Journal of Medical Speech Language Pathology*, *11*(2), ix–xii.

No Child Left Behind Act of 2001. Pub. L. No. 107–110, 115 Stat. 1425. (2002).

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children*, *71*(2), 137–148.

Phillips, L. M., Norris, S. P., & Steffler, D. J. (2007). Potential risks to reading posed by high-dose phonics. *Journal of Applied Research on Learning*, *1*(1), Article 2, 1–18.

Reilly, S. (2004). The challenges in making speech pathology practice evidence based. *Advances in Speech Language Pathology*, *6*(2), 113–124.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *British Medical Journal*, *312*(7023), 71–72.

Silagy, C., & Haines, A. (Eds.). (2001). *Evidence-based practice in primary care*. London: Blackwell BMJ Books.

Part III
Curriculum and Pedagogy

Chapter 8
Researching Effective Pedagogies for Developing the Literacies of Science: Some Theoretical and Practical Considerations

Vaughan Prain

Science education researchers now broadly agree about the fundamental role of the literacies of science in learning in elementary and secondary school (Gee, 2004; Lemke, 1998, 2003; Moje, 2007; Norris & Phillips, 2003; Yore, 2004). These literacies include all the signifying language practices of science discourse, including verbal, visual, and mathematical languages, as well as understanding the purposes and rationale for these literacies in representing scientific thinking and practices. For example, verbal language refers not just to technical science vocabulary and knowledge of functional features of particular science text types but also to verbal reasoning capacities evident in scientific explanations (Osborne, Erduran, & Simon, 2004). There is now broad consensus that students need to learn what Moje has characterized aptly as "disciplinary literacy" (p. 1). In the case of science, this means that students need to (a) learn how, why, and when they should interpret and construct models, graphs, tables, and diagrams and then (b) integrate these representations with the written language of science as part of the broader process of becoming scientifically literate.

Researchers in this field are united in seeking to characterize and explain current or possible future effective classroom practices that promote, or could promote, this disciplinary learning. However, as with all key curricular areas in school, researchers are now also more aware of (a) the marked diversity of learners' needs, cultural resources, and representational capacities; (b) the impact of new technologies on how science is conducted and represented in the science community, and possible or desirable parallel teaching and learning tasks in school; and (c) the complex challenges entailed in students learning the meaning-making and knowledge-production practices of this subject. In the science education research community, this has led to a fitting diversity of research orientations and foci for study.

Recent research has been guided by different theories of how this learning might be characterized and promoted, drawing variously on constructivist, semiotic,

V. Prain
La Trobe University

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

genrist, conceptual change, systemic linguistics, sociocultural, postmodern, and cognitive science theories of meaning-making. Researchers have also used multiple interpretive frameworks to guide their data collection and analyses and have focused on different areas, such as the needs of particular student cohorts, the role of teacher-provided and student-constructed representations, the effectiveness of different student task types and varied technological resources, teacher and student roles, videotaping of classroom interactions with artifacts, and comparative studies of student academic performance following contrasting teaching programs. There has also been recognition of the need for a mix of quantitative and qualitative methods in much of this research so as to measure change and to identify (and explain) participant perceptions and attitudes.

Given the rich diversity and emergent nature of this research, this chapter provides only a broad outline of major developments in this field, noting areas of consensus about effective classroom practices, complementary possibilities across different research methods and foci, and future potential research areas. I focus mainly on student text production in science learning, while acknowledging that this learning entails both constructing and interpreting texts and that many researchers are working predominantly in the area of designing effective texts for students to view, manipulate, and interpret (Ainsworth, 2006; Schnotz & Bannert, 2003). There is general agreement that interpretation of texts, including manipulation of multimodal texts, needs to be part of future research on effective pedagogies for science learning. As noted by Alvermann (2004), Lemke (2004), and others, science findings and explanations are now represented in videos, CD-ROMs, hypertext, and hypermedia as well as traditional print materials; thus, student interpretation and construction of these different text types pose a range of new literacy learning challenges for teachers and students beyond traditional conceptions of reading and writing in science. However, in order to achieve some useful specificity of focus in this chapter, I will consider mainly research on effective pedagogical approaches to student text production.

To frame discussion about the kinds of research undertaken in this area, I briefly review the assumptions and implications of the current mandated version of research excellence, the Gold Standard (Boruch & Mosteller, 2002), noting how this standard does not align easily with the current state of research on learning the literacies of science.

8.1 The Gold Standard

The US Department of Education's (US ED, 2003) assertion that educational research should be strongly evidence-based and use large randomized controlled trials (RCTs) with intervention and control groups could seem on face value to be appropriate for establishing a Gold Standard for learning the literacies of science. Similarly, the What Works Clearinghouse's (US Institute of Education Sciences, n.d.) additional guidelines for Gold Standard educational research could seem appropriate to this field, in that, according to the set criteria, high-quality educational research:

- Employs systematic, empirical methods that draw on observation or experiment.
- Involves data analyses that are adequate to support the general findings.
- Relies on measurements or observational methods that provide reliable data.
- Makes claims of causal relationships only in random-assignment experiments or other designs (to the extent such designs substantially eliminate plausible competing explanations for the obtained results).
- Ensures that studies and methods are presented in sufficient detail and clarity to allow for replication or, at a minimum, to offer the opportunity to build systematically on the findings of the research.
- Obtains acceptance by a peer-reviewed journal or approval by a panel of independent experts through a comparably rigorous, objective, and scientific review.
- Uses research designs and methods appropriate to the research question posed.

While these criteria might be suited to a highly resolved agenda characterized by consensus about research goals, appropriate units of analyses, established causal relationships, and proven instruments to measure appropriate learning outcomes, such conditions are not evident in current research on effective teaching of the literacies of science. This research instead seeks to investigate the complexities and diversities of different classroom procedures, sequences, emphases, technological resources, contexts, and effects on different student groups. This research has also tended to avoid a medical model of assessing binary options for identifying more effective treatments of large populations although some studies of contrasting methods have been undertaken.

As noted by Berliner (2002), the significant complexities of educational contexts and interactions as well as the importance of local knowledge require the use of diverse research methods, such as “case studies, survey research, time series, design experiments, action research, and other means to collect reliable evidence” (p. 20). In agreeing with this viewpoint, Shavelson, Phillips, Towne, and Feuer (2003) noted that “the wide range of questions posed in educational research calls for a healthy diversity of scientific methods [ranging from] pre-science exploration to well-warranted descriptive, causal, and mechanism-driven studies” (p. 28). Berliner also asserted that educational research is shaped by cultural contexts and by researchers’ changing perceptions, assumptions, and goals, thus resulting in unpredictable shelf lives for findings. Agendas, contexts, and priorities necessarily change, creating new imperatives for research foci. Certainly the use of new technologies to conduct and report scientific findings in the science community has led to new challenges in teaching and learning science using these new forms of representation.

Phillips (2005) claimed RCTs did not provide an effective method for research in science education in that they often failed to identify causal mechanisms for treatment outcomes. He argued that evidence, at best, constrains decisions and that educational research should focus much more sharply on how causes produce effects.

Various researchers have also questioned the idea of a single version of research excellence, while others have argued for a more sophisticated mix of qualitative and quantitative methods to capture the complexities and effects of interactions in classrooms (Gee, 2003; Lincoln, 2004; Ryan & Hood, 2004; Teddlie & Tashakkori, 2003). Newman and Cole (2004) claimed there was "ecological invalidity" (p. 261) when laboratory-based research methods, including controlled experiments, were used to investigate effects in everyday classroom environments.

In summary, current research methods into effective pedagogies for developing the literacies of science only partially address the criteria of a proposed Gold Standard. While diverse, systematic, empirical studies have been undertaken to quantify learning effects in different settings, the field remains emergent and complex with much still to research and integrate; therefore, a singular research approach is inappropriate. As noted by Hand and Prain (2006), a totalizing convergence of different research methods and diverse content foci remains, at best, a future challenge and an open question. The complexities and diversities of different classroom environments, resources, and teacher strategies against a backdrop of constant technological and cultural change pose significant challenges for identifying and calculating the effects of different interventions. Indicative of this context of constant change is the lack of a research-based progress map that charts what students might be expected to learn of the literacies of science incrementally over the course of elementary and secondary schooling. At the same time, there is a growing body of research evidence about effective classroom practice for some topics and contexts and also some signs of emerging overlap in researchers' orientation and methods. The next section explores current key research agendas, their theoretical justifications, and their gains in more detail.

8.2 Learning Theories and Learning the Literacies of Science

As Moje (2007) has noted, research into disciplinary learning in science over the last 10 years has tended to draw on three broad perspectives—with considerable overlap across each area and with some researchers showing multiple allegiances. These perspectives are (a) a formal focus on linguistic and semiotic practices and processes, and an advocacy of explicit classroom teaching of these aspects of meaning-making in science; (b) a focus on effective pedagogical strategies drawing on cognitive theories of knowledge production in science, and an advocacy of diverse representational opportunities for learners; and (c) a main focus on sociocultural theories of meaning-making and practice in science, and an advocacy of inclusive cultural border-crossing teaching and learning strategies. The first perspective focuses predominantly on the nature of learning tasks in science, the second on conditions that maximize this learning, and the third on cultural contexts and their relationship to learner diversity and learning opportunities. Various linkages have been attempted across these semiotic, cognitive, and sociocultural orientations.

Each set of perspectives has provided generative insights into both the demands and complexity of learning in science and the classroom processes that are likely to engage different groups of learners. Differences between these perspectives have centered on such issues as the role of everyday language in learning the literacies of science; the value and role of explicit formal knowledge in this learning; the value or place of nonstandard representations in learning these literacies; and the extent to which representational practices are fixed or change (or should change) to accommodate new technologies, new contexts, and different learners' needs. In more recent studies, there has been some degree of perspectival convergence, with increased recognition of the need for teachers to establish representation-rich learning environments that enable students to build bridges between the subject-specific practices of science and their everyday communicative competencies. There is also an increased focus in current research on (a) the implications of recent theories of factors affecting cognition for understanding this learning, (b) implementing and evaluating explicit framing strategies to support this literacy learning, and (c) investigating the role of different sequences of multimodal integration of meaning in this learning. The next section provides an overview of each of the three broad perspectives and their rationale.

8.2.1 Linguistic and Semiotic Perspectives

These perspectives represent a diverse range of approaches to formal analyses of meaning-making processes and practices in science discourse and activity. They include genrist approaches focusing on textual features that affect interpretation, such as semantic density, function–form conventions in multimodal texts, and degrees of abstraction in different texts (Halliday & Martin, 1993; Parkinson & Adendorff, 2004; Scheppegrell, 1998). Other approaches include taxonomic structuralist accounts of visual language (Kress & van Leeuwen, 2006), poststructural multimedia semiotics and discourse analysis (Lemke, 2003, 2004), and sociocultural perspectives on science discourse (Alvermann, 2004; Gee, 2004; Moje, 2007) that seek to foreground the effects of situational factors on different learner cohorts' engagement with science. These perspectives are broadly united by the view that students must learn primarily to understand and reproduce the meaning-making practices of the science community if they are to become scientifically literate (Bazerman, 2007; Gee; Halliday & Martin; Kelly, 2004; Kelly & Chen, 1999; Martin, 2000; Martin & Veel, 1998; Unsworth, 2001; Veel, 1997). Some, like Gee, perceive these practices to be at odds with everyday discourse and language, where the use of "patterns and associations, and repetitions and parallelisms [and] making integrative connections across domains" in everyday discourse is unsuited to thinking scientifically or learning science; he claimed that everyday language tended to "obscure the details of causal or otherwise systematic relations" that are crucial to sense-making in science (p. 27). While acknowledging the key role of the vernacular in anchoring student understanding, Gee is making the reasonable point that science loses its epistemic distinctiveness and its purpose-built forms of knowledge production and representation if attempts are made to simply recast its domain-specific representations in the vernacular. Other researchers, such as Moje, have argued that a socially just approach to teaching the literacies of science requires that teachers cater to students' contrasting needs and capacities; therefore, such teaching is not reducible to a singular pedagogical agenda. For some students, extensive work in vernacular translation may be an essential ingredient for successful engagement.

Within this broad semiotic perspective, the genrist orientation assumes that explicit knowledge of generic rules enables students to "process information deeply [as they] construct relationships among ideas" (Klein, 1999, p. 230). For Bazerman (2007), generic knowledge, once internalized, "provides the basis for a new disciplined way of seeing and thinking" (p. 8). This viewpoint assumes that the languages of science should be understood as a stable, denotative, representational system that must be learnt in order to manipulate its symbols, understand its ways of developing propositional knowledge, and draw appropriate logical inferences. For others, students will learn effectively the rules and meanings of the particular language practices of science through the following teaching strategies: detailed analysis of linguistic features of textual examples, joint construction of genres with their teacher, and through an explicit, extensive, teacher focus on key textual function–form relationships and their rationale (Martin, 2000; Veel, 1997). In other words, researchers within this orientation favor a highly directed, explicit, teacher-focused pedagogy that emphasizes the functional aspects of language features of this discourse. However, as noted by Bazerman, "direct grammatical instruction" tends to be replaced by more subtle teacher "modeling" in practice (p. 7).

Empirical research to support and justify the effectiveness of this range of pedagogical strategies has largely taken the form of case studies of reputed desirable or exemplary implementation (Martin, 2000; Martin & Rothery, 1986; Scheppegrell, 1998; Unsworth, 2001). This research generally fits what Shavelson and colleagues (2003) have labeled "design studies" (p. 28), which they further subcategorize into three possible stages with different questions driving each stage. In the first stage, "in the so-called context of discovery, open-ended exploration is common to design studies, just as it is in any other branch of science" (p. 28). This stage is concerned with a descriptive account of what is happening, identifying a student population with a statistical sample, using ethnographic or case-study research to identify participant perspectives, establishing warranted knowledge claims through qualitative and quantitative data, and linking these to insights into the context and motivation of participants in order to establish a basis for refashioning learning environments. Their second stage is concerned with identifying causal effects and often relied on RCTs. They claimed that (a) these "quasi-experiments and casual models" have a place in educational research, especially in determining which of alternative methods or designs might produce a better outcome; and (b) that RCTs are useful at the "scaling up stage" (p. 28), to see whether a local success is repeatable in different settings or with a larger sample. In this way, such studies can be used to test the generalizability or limits of particular effects. Their third stage focuses on why something occurs by identifying causal mechanisms. They note that "through iterative tryout-redesign-tryout, claims for understanding the mechanism are advanced, and the question of replicability and generalizability then comes into play" (p. 28).

Empirical research based on genrist orientations has tended to focus on the first and third stages of design while ignoring a focus on assessing contrasting treatments. This research has generally assumed that formal knowledge of generic structures of science discourse is the essential mechanism that enables students to learn the literacies of science; it has, therefore, focused on practices that enact this theory rather than undertake comparative studies with other pedagogies. Alternative approaches—such as progressive pedagogies where students are expected to pose problems, ask questions, and negotiate the focus of topics—were criticized as favoring middle-class, knowledgeable, confident, motivated learners and, therefore, failing to provide a successful science learning environment for disadvantaged students (Martin & Veel, 1998). More recently, Gee (2004) has espoused mechanisms more aligned with current research in cognitive science on strategies and practices that enable learning (see Klein, 2006), such as the role of perception, motor actions, feelings, embodiment, analogy, metaphor, and student ability to identify and complete patterns in experiences or texts. This perspective views knowledge as more implicit, perceptual, concrete, and variable across contexts rather than as propositional, abstract, and decontextualized. However, Gee has still asserted that students need to be "scaffolded overtly in how they use and think about scientific social languages, interpretations, and arguments" (p. 31).

8.2.2 Pedagogical Perspectives on Knowledge Production in Science

The second perspective seeks to identify cognitive and communicative conditions that support knowledge building in science and advocates that students construct a diverse range of representations to enable learning (Boscolo & Mason, 2001; diSessa, 2004; Greeno & Hall, 1997; Hand, 2007; Hodson, 1998; Levin & Wagner, 2006; Prain, 2006; Prain & Hand, 1996; Stadler, Benke, & Duit, 2001). This approach, with a predominant pedagogical focus, asserts for various reasons that students should use a more diversified range of writing types to acquire science literacy as well as knowledge of and particular attitudes towards scientific inquiry. This perspective assumes that mobilizing students' use of their community language is crucial to achieving effective engagement with and learning of the literacies of science. In advocating text diversification, these researchers accept that students need to demonstrate a capacity to use accurately the vocabulary and multimodal representations of science discourse. However, they argue that there are motivational gains and enhanced learning opportunities when students engage in a cycle of planning and guided revision of different text types where there is a strong emphasis on clarification of meanings for both self and others.

Researchers within this pedagogical perspective draw on a diverse range of pedagogical and educational theorists, including Bereiter and Scardamalia (1987), Galbraith (1999), and Klein (1999), to advocate the value of expanding the purposes, writing types, and readerships for science texts in science beyond induction into traditional school genres. Writing in science is viewed as a resource to enable learners to understand science concepts, scientific method, and practices beyond the classroom. Educators have asserted that students, in striving to clarify networks of concepts, should be encouraged to write in diverse forms for different purposes (Hand & Prain, 1995; Levin & Wagner, 2006; Rivard & Straw, 2000; Rowell, 1997; Stadler et al., 2001; Wallace, Hand, & Prain, 2004; Wallace, Hand, & Yang, 2004).

This perspective assumes that learning is enabled when students are required to rerepresent or translate an understanding by drawing on their current or emerging linguistic, rhetorical, and conceptual capacities. Agreeing with Greeno and Hall's orientation (1997), diSessa (2004) claimed there was value in students constructing nonstandard representations. He argued that producing these self-designed representations of science topics enabled students to come to understand the logic and aptness of current conventions in scientific representation. He also claimed that students already bring to learning in science some understanding of the need for "conciseness, completeness and precision [in representing ideas, and that] good students manage to learn scientific representations in school partly because they can almost reinvent them for themselves" (p. 299). Bereiter and Scardamalia (1987) claimed that learning is strengthened when students have to transform or reshape knowledge through writing. Galbraith (1999) proposed that writing is a "knowledge-constituting process" (p. 138) where student-writers negotiate a tacit network of linked subject-matter, linguistic, rhetorical, and dispositional understandings. Wallace, Hand, and Yang (2004) claimed there were strong learning gains when students built explicit links between science language (and discourse) and their community language. Researchers from the genrist orientation claim a similar mechanism operates to promote learning when students address generic demands of science writing tasks, but they insist that such writing must focus only on the generic conventions of this discourse.

Empirical research based on a diversified writing-task orientation generally aligns with stages one and two of the educational study design stages proposed by Shavelson and colleagues (2003). Descriptive studies where diversified science writing tasks have been used have reported positive effects on students' attitudes toward, and engagement with, the subject (Hand, Lawrence, & Yore, 1999; Hand & Prain, 1995; Hildebrand, 1998; Prain & Hand, 1996). Students reported that a focus on making sense of scientific ideas and justifying their views was valuable to their learning and also promoted a positive attitude toward the subject when contrasted with undertaking only traditional writing tasks. Hildebrand reported that diversified writing tasks, including more imaginative writing, assisted students' learning processes, had strong motivating effects, and improved learning outcomes.

More extensive comparative studies of contrasting treatments have been conducted by Hand and colleagues around diversified writing types, including the use of a framework called the Science Writing Heuristic (SWH, Hand, 2007). This framework of a modified laboratory report structure leads students through a reiterative process of knowledge construction in science through a focus on making and justifying claims, gathering and representing evidence, and reflecting on the progression of ideas. As noted by Moje (2007), this framework provides an exemplar of “disciplinary text production” (p. 21) in that the development of scientific argumentation is embedded within representations of inquiry processes. Using the SWH in control–treatment group-designed research, Hand and colleagues claimed some significant learning gains when it is used effectively in science classrooms. In reporting on a range of comparative studies, Gunel, Hand, and Prain (2007) noted that using writing-to-learn strategies was advantageous for students compared to those students working with more traditional science writing approaches. Using diversified types of writing enabled students in treatment groups to score significantly better on conceptual questions and total test scores than those in comparison groups. When the cognitive demand of questions increased from extended recall to a design-type question, there were significant performance differences between comparison and treatment groups in favor of treatment. The researchers argued that writing-to-learn strategies required students to rerepresent their knowledge in different forms thereby enabling greater learning opportunities. In another study Gunel, Akkus, Hohenshell, and Hand (2004) reported that students’ performance in answering higher-order cognitive questions was enhanced when they used a modified writing genre contrasted with the traditional laboratory report—although the teacher’s implementation strategies were viewed as a major factor in this outcome. Incorporating a mixed-method approach using qualitative and quantitative data to assess learning outcomes, Hand, Hohenshell, and Prain (2004) found that students in a treatment group focused on a range of writing-to-learn strategies performed better on conceptual questions than a control group. Students’ comments provided support for using nontraditional writing tasks as a means to assist learning, particularly when the audience was different from the teacher. The researchers claimed that writing serves learning when (a) writing tasks are designed to require students to focus on conceptual understanding and also require students to elaborate and justify these understandings of the topic, (b) the target readership is meaningful for the students, (c) students are provided with sufficient planning support, and (d) planning activities engage students in purposeful backward and forward search of their emerging texts.

8.2.3 Cross-cultural Perspectives

Researchers within this orientation seek to identify and build effective pedagogical bridges between the values, interests, discursive practices, and representational resources of different student cohorts and science disciplinary literacy learning (Alvermann, 2004; Ford & Forman, 2006; Gee, 2004; Lee, Luykx, Buxton, & Shaver, 2007; Lee & Roth, 2003; Moje, Collazo, Carrillo, & Marx, 2001; Moje, Peek-Brown, Sutherland, Marx, Blumenfeld, & Krajcik, 2004; Wallace, 2004). These researchers assume that this learning is enabled when teachers work with students to (a) negotiate effectively between everyday discourse, culture, and values and those of the science community; (b) develop explicit understanding of the rationale for the norms of science knowledge production and communication; and (c) sustain connections between expression and values in both cultures. However, Moje et al. (2001) noted that maintaining meaningful links is a significant challenge and that this approach has not been evaluated in any large-scale study of learning effects. Given the significant demands on students' cross-discursive understandings implied by this approach and the challenge of an appropriate standard for measuring learning, this research gap is perhaps not surprising. Working within this broad perspective, Lee et al. (2007) reported on a professional development program for teachers that aimed to link home language and instructional practices in elementary science learning. They explained the limited success of the intervention in terms of various factors, including the teachers' views of "static cultural attributes" (p. 1287) blocking a more emergent view of student identity formation through negotiating multiple meanings around science activities.

Moje (2007), while a strong contributor to this field, pointed out further weaknesses to what she terms the "cultural navigation perspective" (p. 30) on science disciplinary learning. She noted that researchers in this area have tended to take up global, interdisciplinary viewpoints rather than focus on specifics, such as functional linguistic features of textual practices in science, and often fail to suggest practical ways in which everyday text production can be linked precisely to the literacies of this subject. Nevertheless, most researchers in this orientation take for granted the broad stability of science as a disciplinary literacy. By contrast, some postmodern researchers, from cross-cultural perspectives, critique mainstream science knowledge and methods in terms of the selectivity of their governing logic and propose alternative epistemological assumptions, agendas, procedures, and outcomes.

8.3 Implications of this Research Review

This very brief review of pedagogical research into student text production in learning the literacies of science suggests strengths and weaknesses to current research in this field as well as several implications for the focus and kind of future research. The strengths include a broad acceptance across these different perspectives of the pedagogical value of (a) an explicit focus on interpreting and constructing science texts; (b) providing students with effective cognitive strategies (e.g., planning, reviewing, and responding to feedback) to enable successful text production; (c) teaching students the function and form of textual features to show how reasoning, language practices, and meaning-making are interconnected in doing and learning science; and (d) constantly linking learning the disciplinary literacy of science to students' everyday discourses, values, and representational capacities. While there is general agreement that this linkage can only be nurtured through extensive student use of their community or vernacular language, how this linkage is enacted for different learners remains to be clarified through research. Similarly, the consensus about the value of explicit formal knowledge of text structures leaves the questions: What level of understanding is useful? What knowledge is beneficial for which age groups? How might this learning best be achieved?

For Moje (2007), the overriding weakness in current research into these broad pedagogical principles is a general lack of focus on standardized measurement of learning gains arising from particular interventions. While some research agendas have addressed this issue, many studies either explore new practices in a descriptive way or assume and advocate the benefits of a particular approach or sequence, and provide prescriptions rather than empirical evidence for enhanced learning effects. This is not to imply that large-scale RCTs could be introduced to establish a Gold Standard for instruction; the complexities of diverse contexts, contrasting student populations, variable teacher capacities, and different underpinning theories of learning mean that more research is needed in how to enact and link these principles. However, as Moje argued persuasively, "the field needs more studies that report effects in precise and systematic ways" (p. 35).

Lemke (2004), Alvermann (2004), Unsworth (2001), and others have proposed the need for more causal, mechanism-driven studies that identify (a) the changing demands students face in constructing and interpreting multimodal science texts as part of learning science literacy, (b) theories of learning to enable student learning of these new literacies, and (c) pedagogical processes that enhance and maximize this learning. There is also growing recognition that traditional school science writing genres do not match how science is conducted or reported in the science community. On the question of which kind of tasks students should tackle, commentators have noted new complexities, particularly in relation to the need for students to integrate multiple media simultaneously to reinterpret and recontextualize information in one "channel in relation to that in the other channels" (Lemke, p. 41). For Lemke, "the meaning of a text is not expressed by using only one mode" (p. 41); and students have to be able to translate, integrate, and reinterpret meanings across verbal, visual, and mathematical expressions as well as connect these modes to earlier experiences of science activity. However, this complexity, while intensified by new technologies, exists when students have to make sense of traditional printed science texts across different representational modes. This is evident when students interpret the individual and relational meanings between a diagram, an accompanying text, and its referents in the world. Equally, students participate in similar processes when they construct their own text to clarify or elaborate on the meaning of a graph, photograph, or diagram.

There is also the issue of how a developmental curriculum of these literacies might be designed, implemented, evaluated, and refined to promote effective learning throughout the years of schooling but especially in elementary school. Elementary school teachers normally use a mix of generic and science-specific representations for learning in science. Generic representations used in the community and classroom include students' everyday language, cooperative small-group work, whole-class guided discussion, posters, word walls, PowerPoint® presentations, charts, verbal reports, role plays, debates, and narratives. Science-specific representations include three-dimensional models, tables, graphs, diagrams, science journals, multimodal reports, and appropriate vocabulary and measurement for specific topics. This raises the issue of what might count as an appropriate learning sequence for understanding the purpose and structural features of diagrams, graphs, and tables through elementary school. When, why, and how should students learn about bar, column, and line graphs? How might a developmental sequence of increasing complexity be designed, and on what basis? When, why, and how should students learn about the conventions used in diagrams (such as arrows, cross sections, cutaways, and scale drawings) and in tables?

A recent Australian national professional learning program (Primary Connections—Linking science with literacy, n.d.) was designed to support elementary teachers in teaching science with an explicit focus on student learning of the literacies of science. This program focuses on key textual representations, but these representations were chosen on the basis of their appropriate fit with the content of a range of topics across elementary school year levels rather than on research into their developmental appropriateness. There is a need for systematic mapping of current science curricula to identify precisely the developmental demands for students in interpreting and constructing science literacies as they progress through elementary school. Such a map needs to take into account the challenges elementary students face as they learn not only the appropriate verbal language and discourse of science but also learn to "draw, tabulate, graph, geometrize and algebrize science in all possible combinations" (Lemke, 2004, p. 41), incorporating new and old technologies. In other words, there is a strong need for descriptive, ongoing research on the demands new and old technologies make of students as they develop knowledge of how to understand and construct these new literacies.

8.4 New Theories of Learning, New Agendas for Research

Recent accounts by cognitive scientists of how learners learn provide new insights into how students might be supported in learning the literacies of science and also provide further leads on how this acquisition should be researched. Barsalou (1999) and others asserted the situated nature of cognition, including the fundamental role of context, perception, student identity, feelings, storytelling, embodiment, and pattern completion in learning. As noted by Klein (2006), the reasoning of learners is now understood as perceptual processing and analogical mapping, where concepts and language are "perceptually-based, fuzzy and contextual" (p. 151). According to Schwartz and Heiser (2006), perception links strongly to motor actions, where learners habitually draw on their understanding of past actions to coordinate perception. Researchers within this orientation claim that, when students are constructing spatial–visual representations of their understanding of a science topic, they are likely to be using perceptual mapping of features of the phenomena, making expressive personal links with past experiences and associated values, and embedding new understandings in a narrative of themselves as learners of this topic. Rather than using language in a literal, referential way to denote a preexistent propositional model of the world, students are using words and other representations as discursive tools to construct new personalized understandings based on metaphoric reasoning and pattern recognition. This means that students will (a) intertwine conceptual, aesthetic, and emotional aspects of learning experiences; (b) rely more on associative thinking through metaphor, analogy, and pattern searching than on logical manipulation of symbols; (c) develop perceptually based concepts where language and thought intertwine rather than where language operates as a by-product of thought; and (d) depend heavily on the interplay of artifacts, representations, and context to develop understandings.

These accounts of how learners learn have a variety of implications for learning the literacies of science. They foreground both the degree of individual differences in learner responses and thinking and the importance of affective dimensions in learning; they also point to the complex coordination of thought, memory, language, perception, reasoning, and interaction with artifacts in learning in this domain. These accounts imply that this disciplinary literacy learning will be enhanced when students (a) participate in activity sequences that have a strong perceptual context to allow students to use perceptual clues to make connections between aspects of objects and their representation; (b) engage in a sequence of representational challenges that elicit their ideas, enable them to explore and explain these ideas, extend them to new situations, and integrate different representations meaningfully; and (c) focus on topics that take into account their interests, values, aesthetic preferences, and personal histories. Various studies have sought to identify student learning and attitudinal effects when classroom programs were guided by these conditions and positive learning gains were claimed (diSessa, 2004; Hackling & Prain, 2005; Waldrip & Prain, 2006).

In responding to increased recognition of the complex factors influencing classroom science, researchers in various countries are now working to apply new media and multiple analytical frameworks to interpretation of routine and exemplary classroom lesson sequences in elementary and secondary school settings. Clarke and colleagues at the International Centre for Classroom Research (ICCR, n.d.) are investigating patterns of interactions in science classrooms in terms of the distribution of responsibility between teacher and students for knowledge generation in the classroom and the function of classroom artifacts in this process. ICCR uses Clarke's (2001) video-stimulated, postlesson-interview technique to identify student and teacher responses to *meaningful* lesson moments. Bruckmann et al. (2007) have undertaken quantitative analyses of micro features of classroom interactions in teaching physics in a large sample of junior secondary classes in Germany and Switzerland; their work has scope for further identification of key factors in students' learning of the literacies of science, including literacy task demands, teacher scaffolding of these tasks, and effective mechanisms for enabling this learning based on analysis of student performance.

Researchers are focusing on how students engage with the multimodal demands of process and knowledge representation in the science classroom (Ainsworth & Iacovdies, 2005; Danish & Enyedy, 2007; Jewitt, 2007; Parnafes, 2005; Tytler, Prain, & Peterson, 2007; Waldrip, Prain, & Carolan, 2006). This research seeks to take account not only of planning, drafting, and feedback opportunities in student text production as students engage with the functional micro and macro features of genres but also analyzes the effects of rerepresentational work in clarifying students' conceptual understanding. The challenge remains to identify which tasks, task sequences, and kinds of interactions inside and outside the classroom engage learners and optimize learning opportunities. There is growing interest in the role of multiple representations in learning science, with researchers investigating various dimensions, including teacher and student co-construction of diagrams and other representations to identify learning affordances from these processes.

8.5 Concluding Remarks

In this chapter, I have argued that research into learning the literacies of science has been guided by three main orientations and that there has been some valuable convergence between these semiotic, pedagogical, and sociocultural perspectives on how to conceptualize the field of study. Researchers broadly agree that this disciplinary learning needs to be guided by persuasive accounts of the nature of the task, by effective teaching and learning strategies that enable this learning, and by a nuanced sense of how the learning needs of students—especially those currently underrepresented as successful in science—can be met effectively.

However, to develop some broad, generalizable claims about effective teaching and learning frameworks and practices for science literacy learning for different learner cohorts, various research challenges still remain to be addressed. These challenges relate to (a) appropriate interpretive frameworks to guide data collection and analysis, (b) appropriate methods for measuring and explaining change in learner understandings and attitudes, and (c) conceptualizing the impact and opportunities of new learning contexts. There is a need for case studies that seek to blend, or synthesize further, various dimensions of semiotic, pedagogical, and sociocultural perspectives. Current claims made for particular orientations need to be assessed through comparative studies of the effectiveness of contrasting interventions in terms of learning gains. The current widespread recognition of the need for mixed-method approaches in this research needs to lead to a further refinement of which quantitative and qualitative aspects of this disciplinary literacy learning should be the object of investigation—as well as which methods will enable appropriate data collection. There is also a need for descriptive research on the demands and opportunities of new technologies that students use to construct and interpret multimodal science texts within and outside classroom contexts.

References

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, *16*(3), 183–198.

Ainsworth, S., & Iacovdies, I. (2005, August). *Learning by constructing self-explanation diagrams*. Paper presented at the 11th conference of the European Association for Research in Learning and Instruction, Nicosia, Cyprus.

Alvermann, D. E. (2004). Multiliteracies and self-questioning in the service of science learning. In E. W. Saul (Ed.), *Crossing borders in literacy and science instruction: Perspectives in theory and practice* (pp. 226–238). Newark, DE: International Reading Association & National Science Teachers Association.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–600.

Bazerman, C. (2007, August 15–18). *Genre and cognitive development: Beyond writing to learn*. Retrieved May 11, 2008, from http://www3.unisul.br/paginas/ensino/pos/linguagem/cd/English/5i.pdf

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum.

Berliner, D. C. (2002). Educational research: The hardest science of all [Comment]. *Educational Researcher*, *31*(8), 18–20.

Boruch, R., & Mosteller, F. (2002). Overview and new directions. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 1–14). Washington, DC: Brookings Institution Press.

Boscolo, P., & Mason, L. (2001). Writing to learn, writing to transfer. In G. Rijlaarsdam (Series Ed.) & P. Tynjälä, L. Mason, & K. Lonka (Eds.), *Writing as a learning tool: Integrating theory into practice* (Vol. 7 of Studies in Writing, pp. 83–104). Dordrecht, The Netherlands: Kluwer.

Bruckmann, M., Duit, R., Tesch, M., Fischer, H., Kauertz, A., Reyer, T., et al. (2007). The potential of video studies in research on teaching and learning science. In R. Pintó & D. Couso (Eds.), *Contributions from science education research* (pp. 77–89). Dordrecht, The Netherlands: Springer.

Clarke, D. (Ed.). (2001). *Perspectives on practice and meaning in mathematics and science classrooms* (Vol. 25, Mathematics Education Library Series). Dordrecht, The Netherlands: Kluwer.

Danish, J. A., & Enyedy, N. (2007). Negotiated representational mediators: How young children decide what to include in their science representations. *Science Education*, *91*(1), 1–35.

diSessa, A. A. (2004). Metarepresentation: Native competence and targets for instruction. *Cognition and Instruction*, *22*(3), 293–331.

Ford, M. J., & Forman, E. A. (2006). Redefining disciplinary learning in classroom contexts. Review of Research in Education, *30*(1), 1–32.

Galbraith, D. (1999). Writing as a knowledge-constituting process. In G. Rijlaarsdam (Series Ed.) & M. Torrance & D. Galbraith (Eds.), *Knowing what to write: Conceptual processes in text production* (Vol. 4 in Studies in Writing, pp. 139–164). Amsterdam: Amsterdam University Press.

Gee, J. P. (2003, April). *It's theories all the way down: A response to scientific research in education*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Gee, J. P. (2004). Language in the science classroom: Academic social languages as the heart of school-based literacy. In E. W. Saul (Ed.), *Crossing borders in literacy and science instruction: Perspectives in theory and practice* (pp. 13–32). Newark, DE: International Reading Association & National Science Teachers Association.

Greeno, J. G., & Hall, R. P. (1997). Practicing representation: Learning with and about representational forms. *Phi Delta Kappan*, *78*(5), 361–368.

Gunel, M., Akkus, R., Hohenshell, L., & Hand, B. (2004, April). *Improving student performance on higher order cognitive questions through the use of the science writing heuristic*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Vancouver, British Columbia, Canada.

Gunel, M., Hand, B., & Prain, V. (2007). Writing for learning in science: A secondary analysis of six studies. *International Journal of Science and Mathematics Education*, *5*(4), 615–637.

Hackling, M., & Prain, V. (2005). *Primary connections. Stage 2 trial: Research report*. Canberra: Australian Academy of Science.

Halliday, M. A. K., & Martin, J. R. (1993). *Writing science: Literacy and discursive power*. London: Falmer.

Hand, B. (Ed.). (2007). *Science inquiry, argument and language: A case for the science writing Heuristic*. Rotterdam, The Netherlands: Sense.

Hand, B., Hohenshell, L., & Prain, V. (2004). Exploring students' responses to conceptual questions when engaged with planned writing experiences: A study with year 10 science students. *Journal of Research in Science Teaching*, *41*(2), 186–210.

Hand, B., Lawrence, C., & Yore, L. D. (1999). A writing in science framework designed to enhance scientific literacy. *International Journal of Science Education*, *21*(10), 1021–1035.

Hand, B., & Prain, V. (2006). Moving from border crossing to convergence of perspectives in language and science literacy research and practice. *International Journal of Science Education*, *28*(2/3), 101–107.

Hand, B., & Prain, V. (Eds.). (1995). *Teaching and learning in science: The constructivist classroom*. Sydney, Australia: Harcourt Brace.

Hildebrand, G. M. (1998). Disrupting hegemonic writing practices in school science: Contesting the right way to write. *Journal of Research in Science Teaching*, *35*(4), 345–362.

Hodson, D. (1998). *Teaching and learning science: Towards a personalized approach*. Buckingham, UK: Open University Press.

International Centre for Classroom Research. (n.d.). *Homepage*. Retrieved July 10, 2008, from http://www.edfac.unimelb.edu.au/ict/iccr/index.html

Jewitt, C. (2007). A multimodal perspective on textuality and contexts. *Pedagogy, Culture & Society*, *15*(3), 275–289.

Kelly, G. J. (2004, April). *Epistemological dimensions of science literacy*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Vancouver, British Columbia, Canada.

Kelly, G. J., & Chen, C. (1999). The sound of music: Constructing science as sociocultural practices through oral and written discourse. *Journal of Research in Science Teaching*, *36*(8), 883–915.

Klein, P. D. (1999). Reopening inquiry into cognitive processes in writing-to-learn. *Educational Psychology Review*, *11*(3), 203–270.

Klein, P. D. (2006). The challenges of scientific literacy: From the viewpoint of second-generation cognitive science. *International Journal of Science Education*, *28*(2/3), 143–178.

Kress, G. R., & Leeuwen, T., van. (2006). *Reading images: The grammar of visual design* (2nd edn.). London: Routledge.

Lee, O., Luykx, A., Buxton, C., & Shaver, A. (2007). The challenge of altering elementary school teachers' beliefs and practices regarding linguistic and cultural diversity in science instruction. *Journal of Research in Science Teaching*, *44*(9), 1269–1291.

Lee, S., & Roth, W.-M. (2003). Science and the "good citizen": Community-based scientific literacy. *Science, Technology & Human Values*, *28*(3), 403–424.

Lemke, J. L. (1998). Multiplying meaning: Visual and verbal semiotics in scientific text. In J. R. Martin & R. Veel (Eds.), *Reading science: Critical and functional perspectives on discourses of science* (pp. 87–113). London: Routledge.

Lemke, J. L. (2003). Mathematics in the middle: Measure, picture, gesture, sign, and word. In M. Anderson, A. Sàenz-Ludlow, S. Zellweger, & V. V. Cifarelli (Eds.), *Educational perspectives on mathematics as semiosis: From thinking to interpreting to knowing* (pp. 215–234). Ottawa, Ontario, Canada: Legas.

Lemke, J. L. (2004). The literacies of science. In E. W. Saul (Ed.), *Crossing borders in literacy and science instruction: Perspectives on theory and practice* (pp. 33–47). Newark, DE: International Reading Association & National Science Teachers Association.

Levin, T., & Wagner, T. (2006). In their own words: Understanding student conceptions of writing through their spontaneous metaphors in the science classroom. *Instructional Science*, *34*(3), 227–278.

Lincoln, Y. S. (2004). Dual review of the books: *Scientific research in education* & *Evidence matters. Academe, 90*, 110–115.

Martin, J. R. (2000). Design and practice: Enacting functional linguistics. *Annual Review of Applied Linguistics*, *20*(1), 116–126.

Martin, J. R., & Rothery, J. (1986). What a functional approach to the writing task can show teachers about ‘good writing’. In B. Couture (Ed.), *Functional approaches to writing: Research perspectives* (pp. 241–262). London: Frances Pinter.

Martin, J. R., & Veel, R. (Eds.). (1998). *Reading science: Critical and functional perspectives on discourses of science*. London: Routledge.

Moje, E. B. (2007). Chapter 1: Developing socially just subject-matter instruction: A review of the literature on disciplinary literacy teaching. *Review of Research in Education*, *31*(1), 1–44.

Moje, E. B., Collazo, T., Carrillo, R., & Marx, R. W. (2001). “Maestro, what is ‘quality’?”: Language, literacy, and discourse in project-based science. *Journal of Research in Science Teaching*, *38*(4), 469–498.

Moje, E. B., Peek-Brown, D., Sutherland, L. M., Marx, R. W., Blumenfeld, P. C., & Krajcik, J. S. (2004). Explaining explanations. In D. S. Strickland & D. E. Alvermann (Eds.), *Bridging the literacy achievement gap, grades 4–12* (pp. 227–251). New York: Teachers College Press.

Newman, D., & Cole, M. (2004). Can scientific research from the laboratory be of any use to teachers? *Theory into Practice*, *43*(4), 260–267.

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, *87*(2), 224–240.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching*, *41*(10), 994–1020.

Parkinson, J., & Adendorff, R. (2004). The use of popular science articles in teaching scientific literacy. *English for Specific Purposes*, *23*(4), 379–396.

Parnafes, O. (2005, August). *Constructing coherent understanding of physical concepts through the interpretations of multiple representations*. Paper presented at the 11th conference of the European Association for Research in Learning and Instruction, Nicosia, Cyprus.

Phillips, D. C. (2005). The contested nature of empirical educational research (and why philosophy of education offers little help). *Journal of Philosophy of Education*, *39*(4), 577–597.

Prain, V. (2006). Learning from writing in secondary science: Some theoretical and practical implications. *International Journal of Science Education*, *28*(2/3), 179–201.

Prain, V., & Hand, B. (1996). Writing for learning in secondary science: Rethinking practices. *Teaching and Teacher Education*, *12*(6), 609–626.

Primary Connections – Linking science with literacy (n.d.). *Homepage*. Retrieved June 19, 2008, from http://www.science.org.au/primaryconnections/

Rivard, L. P., & Straw, S. B. (2000). The effect of talk and writing on learning science: An exploratory study. *Science Education*, *84*(5), 566–593.

Rowell, P. M. (1997). Learning in school science: The promises and practices of writing. *Studies in Science Education*, *30*(1), 19–56.

Ryan, K. E., & Hood, L. K. (2004). Guarding the castle and opening the gates. *Qualitative Inquiry*, *10*(1), 79–95.

Scheppegrell, M. J. (1998). Grammar as resource: Writing a description. *Research in the Teaching of English*, *32*(2), 67–96.

Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction*, *13*(2), 141–156.

Schwartz, D. L., & Heiser, J. (2006). Spatial representations and imagery in learning. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 283–298). New York: Cambridge University Press.

Shavelson, R. J., Phillips, D. C., Towne, L., & Feuer, M. J. (2003). On the science of education design studies. *Educational Researcher*, *32*(1), 25–28.

Stadler, H., Benke, G., & Duit, R. (2001). How do boys and girls use language in physics classes? In H. Behrendt, H. Dahncke, R. Duit, W. Gräber, M. Komorek, A. Kross & P. Reiska (Eds.), *Research in science education – Past, present, and future* (pp. 283–286). Dordrecht, The Netherlands: Kluwer.

Teddlie, C. B., & Tashakkori, A. (2003). Major issues and controversies in the use of mixed methods in the social and behavioral sciences. In C. B. Teddlie & A. Tashakkori (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 3–50). Thousand Oaks, CA: Sage.

Tytler, R., Prain, V., & Peterson, S. (2007). Representational issues in students learning about evaporation. *Research in Science Education*, *37*(3), 313–331.

United States Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Washington, DC: Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Available from http://ies.ed.gov/ncee/pubs/evidence_based/evidence_based.asp

United States Institute of Education Sciences. (n.d.). *What Works Clearinghouse overview: Standards.* Retrieved May 6, 2008, from http://ies.ed.gov/ncee/wwc/overview/review.asp?ag = pi

Unsworth, L. (2001). *Teaching multiliteracies across the curriculum: Changing contexts of text and image in classroom practice.* Buckingham, UK: Open University Press.

Veel, R. (1997). Learning how to mean – scientifically speaking. In F. Christie & J. R. Martin (Eds.), *Genre and institutions: Social processes in the workplace and school* (pp. 161–195). London: Cassell.

Waldrip, B., & Prain, V. (2006). Changing representations to learn primary science concepts. *Teaching Science*, *52*(4), 17–21.

Waldrip, B., Prain, V., & Carolan, J. (2006). Learning junior secondary science through multimodal representation. *Electronic Journal of Science Education*, *11*(1), 86–105. Retrieved from http://ejse.southwestern.edu/volumes/v11n1/articles/art06_waldrip.pdf

Wallace, C. S. (2004). Framing new research in science literacy and language use: Authenticity, multiple discourses, and the “Third Space”. *Science Education*, *88*(6), 901–914.

Wallace, C. S., Hand, B., & Prain, V. (2004). *Writing and learning in the science classroom*. Dordrecht, The Netherlands: Kluwer.

Wallace, C. S., Hand, B., & Yang, E.-M. (2004). The science writing heuristic: Using writing as a tool for learning in the laboratory. In E. W. Saul (Ed.), *Crossing borders in literacy and science instruction: Perspectives on theory and practice* (pp. 355–368). Newark, DE: International Reading Association & National Science Teachers Association.

Yore, L. D. (2004). Why do future scientists need to study the language arts? In E. W. Saul (Ed.), *Crossing borders in literacy and science instruction: Perspectives in theory and practice* (pp. 71–94). Newark, DE: International Reading Association & National Science Teachers Association.

Chapter 9
Pedagogy, Implementation, and Professional Development for Teaching Science Literacy: How Students and Teachers Know and Learn

Lori Norton-Meier, Brian Hand, Andy Cavagnetto, Recai Akkus, and Murat Gunel

> *If you want to know what I have learned from this unit, it's that you have to have a good question. A good question is one you really don't know the answer to but you can do stuff like experiments to figure it out. But you can't do any old fun stuff. It has to answer your question.* (Amelia, a Grade 3 student)

As this young learner points out, researchers and scientists alike must plan carefully to ask good questions and then collect data in such a way to answer the research question. This is the focus of this chapter: to explore how the very methodology selected to investigate teacher practices and student learning helped to answer our broad, overarching research question: What is the impact on student learning when teachers are supported as learners during professional development and through the process of implementing an innovative approach to science and literacy?

Amelia, who is quoted above, was a participant in a 3-year research project conducted in a US Midwestern state that investigated teacher implementation of an innovative approach to integrating science and literacy called the Science Writing Heuristic (SWH, Hand & Keys, 1999). Teachers need help in implementing writing strategies within their classrooms that ultimately have an impact on helping students learn science. The No Child Left Behind legislation (NCLB, 2002) emphasized reading and mathematics to the exclusion of other subjects, particularly in the elementary grades. Therefore, there is a growing need for integrating language and science (Bybee, 1995; Hand & Prain, 2006; Saul, 2004).

L. Norton-Meier
Iowa State University

B. Hand
University of Iowa

A. Cavagnetto
State University of New York-Binghamton

R. Akkus
Abant Izzet Baysal University

M. Gunel
Ataturk University

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

Unlike the two following chapters (see Nieswandt & McEneaney, Chap. 10; Levin & Wagner, Chap. 11), where mixed-methods research as applied to individual studies is described, we have adopted this methodological approach at the project level. That is, in order to answer a broad question that has framed the project, we as researchers believed it was necessary to adopt qualitative and quantitative studies at the substudy level that collectively could address the research question. We were guided in part by Howe (2003) who pointed out that qualitative and quantitative methodologies are not distinct but are part of a whole thinking process, and thus mixing methodologies is simply what researchers already do—mixing both qualitative and quantitative knowing. Smith (2006) stated that in mixed methodology "researchers must examine the phenomenon from all angles; describe its manifestation in its context; describe in detail the local conditions and suggest how they might be similar or different from those in related cases" (p. 471).

Our project is framed around the second of Prain's focal areas (see Prain, Chap. 8), that is, how to improve student learning through scaffolding the implementation of science argument and language practices as embedded components of elementary science classrooms. In the following paragraphs, we describe the project on two levels: first, a macroanalysis including a broad overview of the study and findings and, second, a microanalysis of two Grade 5 teachers' implementation and learning processes that provides a more descriptive perspective in answering the research question. However, the overarching key question to our work is: What are the essential elements of the SWH approach and what then is their influence on student learning? First, we provide a brief overview of the SWH approach and the supporting research base related to its use in classrooms.

9.1 About the SWH Approach

This project is based on incorporating an emphasis on language as a critical component of science inquiry as undertaken through the SWH approach. Building on previous success with implementing the SWH approach within middle school (Hand, Wallace, & Yang, 2004), secondary school (Akkus, Gunel, & Hand, 2007; Gunel, 2006; Hand, Hohenshell, & Prain, 2004; Hohenshell & Hand, 2006; Keys, Hand, Prain, & Collins, 1999), and tertiary environments (Greenbowe & Hand, 2005; Rudd, Greenbowe, & Hand, 2001), the researchers were keen to explore its use with younger children (Gunel, Akkus, Hand, & Norton-Meier, 2006; Hand, Norton-Meier, Gunel, & Akkus, 2006). Thus, this chapter discusses the implementation of a 3-year, mixed-methods research project involving 32 prekindergarten–Grade 6 teachers who used the SWH approach for teaching science and embedding language practices within their classrooms. This study explored the use of the SWH approach as a means to promote scientific reasoning, linking embedded language practices within science lessons, and to determine the impact on students' understanding of science concepts.

The SWH approach consists of a framework that guides students through activities and serves as a metacognitive support to prompt student reasoning about data (Hand & Keys, 1999). Similar to Gowin's Vee heuristic (1981), the SWH approach provides (a) learners with a heuristic template to guide science activity and reasoning in writing and (b) teachers with a template of suggested strategies to enhance learning from laboratory activities (Table 9.1). The SWH approach is a bridge between informal, expressive writing modes that foster personally constructed science understandings and more formal, public modes that focus on canonical forms of reasoning in science. In addition, the template for student thinking prompts learners to generate questions, claims, and evidence for claims and making an argument based on valid reasoning.

Argumentation is a fundamental tradition of science communities. Every science community employs argumentation (knowledge, plausible reasoning, patterns of argumentation, variation in evidence) to establish or justify knowledge claims. Arguments have three generally recognizable forms—analytical, dialectical, and rhetorical—that can be used very effectively by teachers to increase students' science understanding (Driver, Newton, & Osborne, 2000; Duschl & Ellenbogen, 1999). Argument can be both an individual activity done through thinking and writing or a negotiated social act (Vygotsky, 1978). Translating these activities into the classroom so that students can build an understanding of and be able to practice scientific argument requires a sequence of instruction that provides opportunities for student growth (Duschl & Ellenbogen, 2002). Traditional discourse patterns used by teachers do not encourage, or even allow, the type of discourse that scientists undertake as they build arguments for scientific claims. Building on this position, Wallace and Narayan (2002) suggested that for students to engage in science where argumentation is a core component they need to be involved in "learning to

Table 9.1 Teacher and student template for the science writing heuristic (SWH) approach

Teacher template	Student template
• Exploration of preinstruction understanding • Prelaboratory activities • Laboratory activity • Negotiation I: Individual writing • Negotiation II: Group discussion • Negotiation III: Textbook and other resources • Negotiation IV: Individual writing • Exploration of postinstruction understanding	Beginning questions or ideas • What are my questions about this experiment? Tests and procedures • What will I do to help answer my questions? Observations • What did I see when I completed my tests and procedure? Claims • What can I claim? Evidence • What evidence do I have to support my claim? How do I know? Why am I making these claims? Reading • How do my ideas compare with others? Reflection • How have my ideas changed?

use language, think and act in ways that enable one to be identified as a member of the scientific literate community and participate in the activities of that community" (p. 4). This requires teachers to create situations in which students can talk science in real science contexts (Lemke, 1990).

By engaging in this process, students experience first-hand science situations where there is more than one plausible solution or answer. This requires thinking critically through the value of each answer or solution and implementing reasoning strategies to argue for a solution to the problem. In addition, the SWH approach can be understood as an alternative format for laboratory reports as well as an enhancement of learning possibilities of this science genre. Instead of responding to the five traditional sections—purpose, methods, observations, results, conclusions—students are expected to (a) respond to prompts eliciting questioning, knowledge claims, evidence, description of data and observations, and methods; and (b) reflect on changes to their own thinking. The SWH approach is structured so that students engage in reasoning that parallels scientists' reasoning and writing.

9.2 Overall Research Design for the SWH Project

In addressing the overarching research question of the project, it became apparent that two distinct research methods would be required. The first issue was to determine the essential elements of the SWH approach. This question was best addressed through a qualitative approach where engagement with teachers was required. Research tools, such as observations, interviews, and reflective evaluations, were used to gather information on all the participating teachers. The intent of this research process was to (a) construct a profile of implementation for each teacher in order to generate both an understanding of the essential pedagogical features of the SWH approach and (b) generate a form of stratification related to the quality of implementation. This stratification was important because we knew from the outset that there would be differences between teachers in terms of their implementation, and we wanted to determine the impact on students' learning as a consequence of this difference. This phase of the research was critical not only to generate some sense of the differences between teachers but also to understand the commonalities between teachers who were approximately equal in terms of the quality of their implementation.

Understanding and stratifying teacher implementation was an essential component of the project because it enabled us to address the second element of the research question: How does implementation impact student performance? For us, the only way to deal with this second component was through the use of statistical analysis. That is, to examine if the quality of teacher implementation impacts student performance on standardized test questions, we needed to establish the levels of teacher implementation, collect student performance scores for science, language, and mathematics, and to have a baseline test that could be used as a covariate if needed. As such, the method for this component of the research project was distinctly different from the first phase.

9.3 Microanalysis of the SWH Project: Focus on Two Grade 5 Teachers

Analysis of student achievement as measured by standardized examinations does provide evidence that the teaching approach utilized during the project had positive influences on the students; however, more specific components of the overall SWH approach are not considered within the macroanalysis. Specifically, we were curious to begin investigating how teachers implement greater dialogical interaction within their science classrooms. While research has increased since the 1970s around language and learning, verbal interaction (both talk and listening) has not been emphasized in elementary classrooms to the extent of reading and writing. Thus, the question guiding the microanalysis was: What essential elements did teachers have to engage with as part of their change process toward greater dialogical interaction within their science classrooms?

9.3.1 Research Setting and Participants

9.3.1.1 School and Students

The two teacher participants for the microanalysis were enrolled in the 3-year SWH project. Both teachers taught Grade 5 at a rural Midwestern elementary school in the United States. During the study, the school district served just under 2,500 students K-12 (per year) with 33% of the students characterized as having low socioeconomic status. Each teacher (Jenny and Lisa) instructed two sections of science as they rotated classes with one of their colleagues; that is, as they taught science, their colleagues taught mathematics and then they exchanged students. Students were predominantly Caucasian (greater than 95% in both years) with 49% female and 51% male in Lisa's classes ($n = 50$) and 45% female and 55% male in Jenny's classes ($n = 53$).

9.3.1.2 Teacher Participants

Lisa and Jenny were chosen for this study as they worked in the same school at the same grade level—yet had very different personalities, backgrounds, and perspectives of the ideal classroom. Upon entering the professional development project, both teachers could be considered traditional with regard to science instruction with each taking on the role of knowledge provider. They were heavily involved in utilizing kit-based science instruction, so they were using a number of hands-on activities, yet the activities were heavily structured so student voice was limited.

At the beginning of the project, Lisa ran a very structured, controlled classroom; although similar, Jenny's tolerance for student talk and general classroom noise was greater than was Lisa's at the beginning of the project. During Year 1, both teachers would be considered in the development stage *inexperienced user* (see Loucks-Horsley & Steigelbauer, 1991) as they were concerned with issues of personal ability to effectively implement the SWH approach and issues of managing the SWH approach. Specifically, both teachers struggled with promoting student voice as they tended to place themselves in the middle of classroom conversations and often provided the right answer to students. Importantly, Jenny did provide more opportunity for student voice than Lisa as she began to ask more explanatory questions, such as *What do you think about that?* In general, however, dialogical interaction was limited as the conversation generated was commonly in the form of an initiate–response–evaluate pattern (Mehan, 2001). The trend of attempting to manage was also found in students' journals of both teachers' classrooms as many of the entries were discussed by the class, worded by the teacher, and then copied verbatim into the student journals.

Lisa. Lisa is an experienced teacher having taught at the elementary level for 11 years—all within the same school district. She taught Grade 1 for 2 years and then moved to Grade 5 for 9 years prior to participation in the microanalysis. Lisa holds a Bachelor of Science in elementary education from a large university in the Midwestern United States and a Master of Science in elementary school counseling from an established private university located within the same state. Although Lisa's background is in education with 12 science credits, she is quite confident in her science knowledge, which may not only be attributed to her experience teaching Grade 5 science but also to her family life (she enjoys animals, the outdoors, and is married to a medical doctor). Lisa was recruited into the SWH project by her school administrator and Jenny, the other teacher participating in the study.

Jenny. Jenny is also an experienced teacher having taught for 28 years at the elementary level. She taught Kindergarten for 15 years (5 in a different school district), Grade 1 for 11 years, and Grade 5 for 2 years prior to participation in the microanalysis. Jenny holds a Bachelor of Science degree from a large university in the Midwestern United States in elementary education and remedial reading, during which time she completed 7 science credits. She completed a Master of Science in Education in effective teaching and curriculum development from an established private university in the same state, holds an Educational Specialist degree in curriculum development and elementary administration from the same institution, and has an early childhood endorsement. Due to her limited science credits and science teaching, Jenny did not view herself as conceptually strong in science. Jenny was recruited into the project by the project's principal investigator and volunteered to be involved in the pilot project. Prior to the project, Jenny would be considered a traditional teacher as she viewed her role in science, like Lisa, as provider of knowledge.

9.3.2 Data Collection

This microanalysis of the two teachers' journey through implementation of the SWH during Year 2 of the project utilized both qualitative and quantitative data. Data were collected over four units of study, which was 80% of the academic year. Building on the concerns-based adoption model, the researchers chose to utilize qualitative methods characterized by Loucks-Horsley and Steigelbauer (1991) as conversational assessment, specifically consisting of on-site observations, semistructured interviews, and informal conversations. Quantitative data presented as descriptive statistics (due to the small sample size) represented teacher progression as determined by use of the Reformed Teacher Observation Protocol (RTOP) during video and on-site observational analysis (Piburn et al., 2000).

9.3.3 Data Analysis

Teacher interviews were conducted four times throughout the study: the middle of unit two, the end of unit two, the end of unit three, and the middle of unit four. Questions were structured during each of the four interviews to determine barriers and obstacles to implementation of the SWH and, in particular, dialogical interaction. Follow-up questions were asked in an effort to probe initial responses. Interviews were audiotaped and transcribed. Researchers attempted to find patterns in the data between teachers and within teachers over the course of the four interviews. Notes regarding substantive aspects of informal conversations with teachers were documented by the researchers. Finally, an extended response questionnaire was completed by the two teachers after the conclusion of the fourth unit to extract information on follow-up questions that evolved out of the analysis of previous interview, observation, and informal conversation data.

Teachers were observed on-site and through video analyses. Each teacher was required to notify the researchers when she was teaching the units, and arrangements were made for observation. On-site observations were carried out monthly starting at the end of September 2005 and continuing through the end of April 2006. During on-site observations, the researchers recorded field notes focusing on pedagogical strategies adopted for student-centered classrooms, big ideas of the unit, nature of the dialogical interactions, and the management strategies employed.

Each teacher recorded three to four videotapes per unit of study. Overall Jenny was recorded 12 times while Lisa was recorded 16 times over the duration of the study. Videotapes were scored using the RTOP (Piburn et al., 2000). This scoring rubric was designed as a quantitative way to measure teachers' progression toward instructional practices identified by the *National Science Education Standards* (US National Research Council, 1996). The protocol utilizes a rating scale with a range from zero (the characteristic never occurred in the lesson) to four (the characteristic was very descriptive of the lesson). This study considered scores for dialogical interaction (RTOP items 2, 15, 16, 17, 18, 19, 20, 25) and control/focus of learning (RTOP items 1, 5, 21, 22, 24); Table 9.2 provides a detailed description of each item.

Items were thoroughly discussed among two researcher–raters prior to rating. Interrater reliability using the entire RTOP protocol was conducted and yielded a Pearson correlation coefficient of 0.962. In addition to the scores generated from the observation protocol, descriptive notes were taken during video and on-site observations focusing on pedagogical strategies adopted for student-centered classrooms, unit big ideas, nature of dialogical interactions, and management strategies employed. The notes taken during on-site and video observations were cross-referenced with interview data and notes from informal conversations with the teachers to identify themes across the datasets.

9.3.4 Results

By utilizing the aforementioned data sources the following conclusions were drawn regarding these two teachers' journey toward greater dialogical interaction in their classrooms via the SWH approach:

Table 9.2 Reformed Teaching Observation Protocol (RTOP)

Collapsed categories	Reformed Teaching Observation Protocol (RTOP) descriptors
Dialogical interaction	
2	The lesson was designed to engage students as members of a learning community
15	Intellectual rigor, constructive criticism, and the challenging of ideas were valued
16	Students communicated their ideas to others
17	Teacher questioning triggered divergent modes of thinking
18	High proportion of student talk and a significant amount was student to student
19	Students' questions and comments determined focus and direction of classroom discourse.
20	There was a climate of respect for what others had to say
25	The metaphor "teacher as listener" was very characteristic of this classroom
Control/Focus of learning	
1	Instructional strategies respected students' prior knowledge/preconceptions
5	Focus and direction of lesson determined by ideas from students
14	Students were reflective about their learning
21	Active participation was encouraged and valued
22	Students were encouraged to generate conjectures, alternative solution strategies, and ways of interpreting evidence
24	Teacher acted as resource person, supporting and enhancing student investigations

Conclusion 1: Teachers were able to promote greater amounts of dialogical interaction in their classrooms as the school year progressed. As illustrated in Fig. 9.1, RTOP scores relating to dialogical interaction increased from the beginning to the end of the year. This increase can be attributed to the teachers' improved questioning and removing themselves from the center of the discussion.

Conclusion 2: Teacher perceptions of the learning process and teacher content knowledge influenced questioning and, therefore, the quality of dialogical interaction throughout the study. Teachers struggled for the first half of the study with issues of control; however, as illustrated in Fig. 9.2, they began to remove themselves from the center of the discourse, leading to increased student voice and more challenging questions. Lisa articulated her shift in pedagogy when she indicated, "I think I'm becoming more comfortable with them (students) critiquing each other and turning over more of the responsibility to them" (Lisa interview, 2/27/06). Similarly, Jenny summarized her progression, "I guess the teacher has to let go and that is finally where I am getting to the part where I can put the decision making on the kids" (Jenny interview, 3/28/06).

The shift in questioning was evidenced in the final questionnaire. When asked if her ideas about the relationship between management and teaching had changed. She responded, "I'm not sure that my ideas have changed as much as they have been rearranged, and my emphasis is on asking the 'why' question continually" (Jenny, questionnaire, 5/1/06). When compared to Fig. 9.2, the pattern illustrated in Fig. 9.1 suggests a relationship between actions related to control/focus of learning and dialogical interaction.

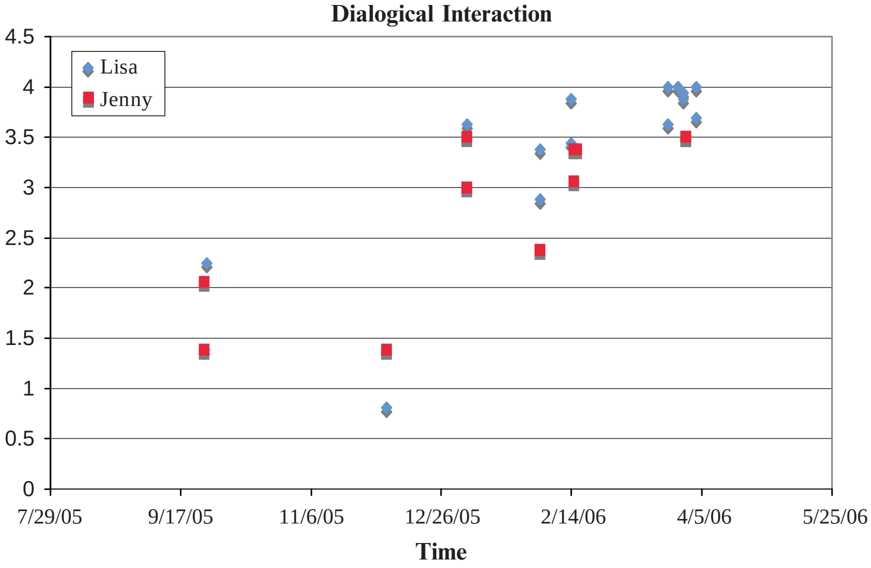


Fig. 9.1 Average RTOP scores relating to dialogical interaction over the study

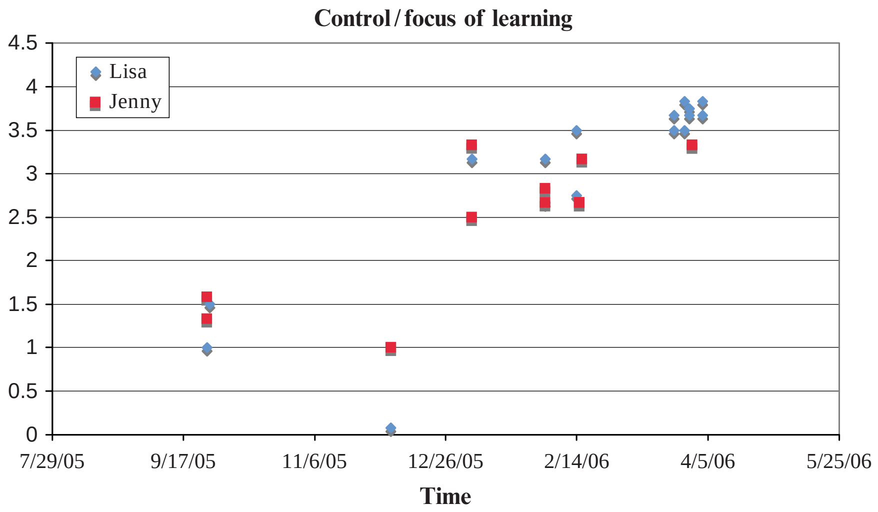


Fig. 9.2 Average RTOP scores relating to control/focus of learning throughout the study

The influence of content knowledge on ability to allow for greater student interaction and student input on direction of study was evident in Jenny's case. When asked if the unit of study influenced her implementation of the SWH approach, Jenny stated:

> Oh definitely. Definitely, I feel more comfortable teaching the biomes than I do light. Just because it's my third year teaching it, I'm still trying to get a hold on it, I'm not a science person so that keeps holding me back I think. (Jenny interview, 2/27/06)

Jenny's content knowledge concerns influenced her questioning ability and her ability to allow the classroom discussion to take unpredictable paths to the big idea.

Conclusion 3: Organizational components, such as increased time use and guidelines for student-to-student discussions, were necessary for students to meet the expectations of teachers with regard to quality of output and to maximize quality of dialogical interaction. Comparison of the two teachers' classroom choices over the course of the study yielded evidence that organizational components were necessary for fostering dialogical interaction. Early in the study, both teachers valued efficiency of time use; this was evident in the amount of time allocated for each phase of the SWH in the first two units. One day was provided to determine the question for exploration to design experiments to answer the question. As a result, students carried out experiments that oftentimes did not address the inquiry question. Realizing that students were not given adequate time to effectively conduct the investigations, both teachers began allowing greater time for this process. They began allowing 1 day for question generation and an additional 2 days for presentation and reconfiguration of their experimental design. As such, misalignment among questions and experimental designs were caught by other students during the presentations. Lisa summarized the importance of this evolution indicating, “I have learned that if they publicly defend their question and experiment, it makes it easier for the class and me to assist them in setting up more valid experiments” (Lisa, questionnaire, 5/1/06).

Other components that are often either overlooked or uncomfortable to address played a role in student-to-student dialogue. For example, public-speaking skills during formal and informal presentations can be inhibitory. Only when Lisa became fully committed to interrupting students to point out better speaking practices did presentation and respectful communications occur. Importantly, this can be difficult for teachers as they often feel as though they are continually interrupting. We found that dialogue increased as students became more adept at speaking respectfully to others. A more comprehensive description of the results is described by Cavagnetto (2006).

9.4 Macroanalysis of the SWH Teacher Project

In educational research, the impact of professional development programs—and, consequently, teachers’ practices and epistemological beliefs toward teaching and learning—are investigated using qualitative methodology (e.g., observations, interviews). The other important outcome, students’ academic achievement, is found to be a strong criterion to measure the accomplishments of in-service programs and the effectiveness of teacher implementation (Chinn & Malhotra, 2002; Songer, Lee, & McDonald, 2003). In particular, students’ performances on standardized tests and teachers’ classroom practices become crucial for evaluating educational settings and the impact of professional development programs (Sanders, Wright, & Horn, 1997). In this study, the following research questions were investigated:

1. What are the impacts (if any) of socioeconomic status (SES), individualized education program (IEP) status, and teacher implementation level on students’ standardized science test scores?
2. What are the barriers and obstacles to teacher progression toward student-centered, teacher-managed instruction?

9.4.1 Research Setting and Participants

There were 31, 32, and 31 teachers, along with their students, in Year 1, Year 2, and Year 3, respectively, across Prekindergarten–Grade 6 levels from five school districts in the Midwestern United States. Four of the school districts were in rural settings with two districts designated as rural poverty areas by the federal government. The fifth school district was a large urban district. Table 9.3 shows the distribution of the teachers and students across grade levels and years. Even though the majority of the participants continued in the project for 3 years, there were a few teachers who dropped out of the study at the end of the first or second year; hence, new teachers were asked to participate.

The participating teachers were involved in a 3-year implementation and analysis cycle of a yearly summer institute that included a science content update, critical reading experiences, and science inquiry teaching strategies. In addition, teachers and project staff worked together during the year on planning SWH units, implementing this curriculum in the classroom, and contributing to ongoing data collection and analysis. Each summer, the teachers engaged in 10 days of professional development to experience the SWH approach in action. This workshop focused on examining the SWH approach in the following ways:

- Science content knowledge update (teacher as learner).
- Learning theory knowledge update.
- Pedagogical content knowledge update.
- Embedded language practices.

In addition to the summer workshop, the teachers received 3 days of professional development during the school year, which focused on reviewing unit planning and reflecting on strategies for implementation, concerns about struggles being faced, obstacles to be overcome, and language connections. The underlying theme to all of the professional development work with teachers was *learning*—what is learning, and how do teachers support the learning of every student in the classroom? Close attention was paid to the new teachers; they were either matched with experienced SWH teachers in their school district or an SWH research member was present, or both, during the preparation and the implementation of the units.

Table 9.3 Distribution of teachers and students across grade levels and 3 years

	Grade levels								
	Pre	K	1	2	3	4	5	6	Total
Teachers									
Year 1	1	1	N/A	5	4	7	7	6	31
Year 2	1	2	N/A	6	3	6	7	7	32
Year 3	1	3	3	4	3	5	6	6	31
Students									
Year 1	13	19	N/A	98	82	153	231	184	780
Year 2	13	38	N/A	112	54	123	237	205	782
Year 3	13	58	59	71	70	100	193	197	761

9.4.2 Data Collection

There were multiple data sources to analyze the complexity of the research questions. The qualitative data included observations through both on-site observations and videotaped recordings, interviews with the teachers and students, and artifacts (e.g., teachers' unit plans, students' writing samples). The quantitative data included the Iowa Test of Basic Skills (ITBS) scores for students in Grade 2 and up, when available.

Observation and Videotaping. For each teacher, two types of observation occurred: on-site and videotaped. During the on-site observations, an observer who had SWH teaching experience was physically present in the classroom, following the teacher and taking field notes on teacher–student interactions. After the lesson, the observer had a short debriefing session with the teacher, where constructive feedback was provided after the teacher's identification/self-evaluation of strengths, weaknesses, and difficulties with implementation. Such debriefings targeted several areas of interest, such as promoting the teacher's awareness of certain observed behaviors, highlighting any pedagogical areas needing improvement, and suggesting some strategies to improve implementation of the required student-oriented approaches in the future. After the debriefing sessions, the observer electronically kept major points of the sessions. Each teacher was videotaped at least once during implementation. The videotaped recordings were used to make detailed analyses of implementation level by independent observers. The on-site and videotaped observations were conducted by three graduate students and two professors.

9.4.3 Data Analysis

The analyses consisted of the constant comparative method of data analysis. In terms of the qualitative data, the on-site and videotaped observations were analyzed multiple times in order to identify and confirm the teachers' levels of implementation. The SWH criteria matrix, jointly developed and improved by Omar and Gunel (2004), Gunel (2006), and Gunel et al. (2006), was used to determine the level of implementation. The percentage of agreement (or the interrater reliability) between any pairs of observers for teachers' level of implementation was 90–95%. If there was any disagreement about a score, all observers watched the videotape and made a decision based on a discussion revolving around the problematic part of the observed teaching. Such discussions resulted in 100% agreement by providing rationales for the scores.

The criteria matrix consisted of four major areas in pedagogical practice. These criteria placed a teacher in one of three categories for defining the quality of the implementation of the SWH approach: low, medium, or high. The teachers who continued in the project for at least 2 years were followed in terms of any changes in their implementation level.

Dialogical interaction is the first of the four criteria. Types of questions asked by teacher and students, teacher's response to students' answer and questions, and the direction of communication (e.g., from teacher to student) are of importance for creating dialogical interaction.

The second criterion is focus of learning. Focus of learning was defined in the SWH approach as creating a nonthreatening environment, choosing an inquiry investigation, and promoting public sharing of knowledge, which is an important step away from traditional science classroom practice. Teachers are expected to allow students to ask their investigation questions, build their models, and support their claims using the evidence they find during the investigation.

Unit preparation and making connection is the third criterion. Unit preparation refers to identifying the big ideas of the units, which reflects teachers' understanding of the content knowledge. In deciding the big ideas, teachers are engaged in an inquiry about their students' prior knowledge on which students build new concepts. Making those connections requires centering the concepts of the units on the big ideas and students' prior knowledge and supporting students in learning scientific language.

The last criterion is science argumentation. For the SWH approach, it is crucial that teachers encourage students to make a scientific argument among themselves by providing evidence for their knowledge claims. One role of an SWH classroom teacher is to create dialogical interaction (i.e., the first criterion) to promote scientific debate. Students argue based on the big ideas negotiated in the classroom.

In terms of the quantitative data, ANCOVA models were estimated to be able to probe the effects of other variables on students' performances and the possible interaction of those variables (Agresti & Finlay, 1997). ITBS science scores were the dependent variable; implementation level, year, SES, and IEP status were the independent variables; and ITBS mathematics and social studies scores were the covariates. To ensure the accuracy of the data collected, both frequency distributions and descriptive statistics were obtained using the SPSS Frequencies procedure (Mertler & Vannatta, 2002). The SPSS Casewise Diagnostic procedure was employed to examine whether outliers affected the results of the study (Levine & Roos, 2002).

In order to make a meaningful interpretation of the results, students' grade equivalent (GE) growth scores were used instead of raw scores. The GE is a number that describes a student's location on an achievement continuum, describing the performance in terms of grade level and months. For example, if a Grade 6 student obtains a GE of 8.4 on the vocabulary test, that score is one a typical student finishing the fourth month of Grade 8 would likely get on the vocabulary test. To calculate the growth, we also obtained the date of the test. For example, suppose that the test is taken after 6 months of schooling, indicating that the student's grade equivalency is 6.6. Thus, the grade equivalent growth for this student would be 1.8 (8.4 – 6.6).

In this study, we reported effect sizes using Cohen's *d* index, which is widely used in social science because it enables researchers to measure "the difference between two means expressed in standard deviation units" (Sheskin, 2004 , p. 141) and to recognize the magnitude of intervention on students' learning. There are three advantages to reporting effect sizes. First, it makes meta-analyses possible for a given report; second, it allows a researcher to determine more appropriate study expectations in future studies; and, third, it facilitates assessment and comparison of a study's results across existing related studies (Wilkinson & APA Task Force on Statistical Inference, 1999).

9.4.4 Results

In this study, several research questions were addressed: The first question attempted to articulate the relative effect of implementation levels and year on different characteristics of students (i.e., SES and IEP) for ITBS science scores. Therefore, this question captured various aspects of the study.

Before giving the statistical results, we provide the findings from the qualitative component of the study in order to create a base for the quantitative analysis. As mentioned earlier in Sect. 9.4.3, each teacher was ranked on their SWH implementation as low (L), medium (M), or high (H) every year. The results of the observational analysis indicated that there were 19, 10, and 10 low teachers in Year 1, Year 2, and Year 3, respectively (see Table 9.4). Similarly, 12, 22, and 12 teachers were ranked as medium in Year 1, Year 2, and Year 3, respectively. There were 9 teachers ranked as high in Year 3. No one was ranked as high in Year 1 and Year 2. Furthermore, we are able to follow 20 teachers from Year 1 to Year 3 in terms of the change in their levels of implementation of the SWH. As such, 1 teacher stayed at low (L-L-L); 3 teachers were at medium for all 3 years (M-M-M); and the change of others was in the following patterns: 1 teacher L-L-M, 2 teachers L-L-H, 2 teachers L-M-L, 4 teachers L-M-M, 3 teachers L-M-H, and 4 teachers M-M-H.

For science, two 3 × 3 × 2 ANCOVA models were estimated. The first model was conducted using year, implementation level, and SES as independent variables. In this model, there were three statistically significant effects: implementation level

Table 9.4 Distribution of levels of teacher implementation across year and grade

	Teacher implementation level						
	Year 1		Year 2		Year 3		
Grade level	Low	Med	Low	Med	Low	Med	High
Pre	–	1	–	1	–	–	–
K	1	–	1	1	1	1	–
1	N/A	N/A	N/A	N/A	3	–	–
2	4	1	2	4	3	–	3
3	2	2	1	2	–	3	–
4	4	3	2	4	2	3	1
5	5	2	1	6	–	2	3
6	3	3	3	4	1	3	2
Total	19	12	10	22	10	12	9

Table 9.5 Mean scores for implementation levels across 3 years

Year	Imp. level	Adj. mean	Std. error	*n*
1	Low	1.430	0.117	332
	Med	1.679	0.142	269
2	Low	1.089	0.130	218
	Med	1.851	0.107	387
3	Low	1.388	0.236	62
	Med	1.478	0.125	302
	High	1.700	0.161	162

main effect ($F(2, 1717) = 4.688$, $p < .009$, $\eta^2 = .005$); SES main effect ($F(1, 1717) = 13.408$, $p < .001$, $\eta^2 = .008$); and year–implementation level interaction effect ($F(2, 1717) = 10.657$, $p < .040$, $\eta^2 = .004$). That is, the implementation level main effect comes from the significant differences of medium and low implementation ($t(1568) = 3.003$, $p < .05$) and of high and low implementation ($t(772) = 2.112$, $p < .05$), and corresponding *d* effect sizes were 0.158 and 0.169, respectively. Second, students with high SES significantly outperformed students with low SES ($t(1730) = 3.605$, $p < .05$), and effect size was 0.189. Moreover, pairwise comparisons of levels of teacher implementation across 3 years showed that the higher the implementation level the larger the students' growth in ITBS science for each year (Table 9.5). For example, the mean difference between medium and low implementation levels in Year 2 is 0.762, which is significant at $\alpha.05$ ($t(603) = 4.526$, $p < .05$).

Even though there was no three-way interaction effect among year, implementation level, and SES, pairwise comparisons showed that students with low SES in medium and high implementation levels outperformed students with low SES in low implementation level within each year, with only Year 2 being significant. Due to the limited space, we are unable to report the results with IEP; yet, the same trend occurred when using the IEP status. That is, students with IEP outperformed their like-peers when they were in a high level of implementation within each year.

9.5 Discussion and Summary

In closing, it is essential to pause and reconsider the value of the methods chosen for this study. In recent years, mixed-methods research has received an adverse reaction—particularly in the United States and in the wake of NCLB—as preferencing quantitative data with the overwhelming opinion that to receive funding a qualitative researcher must adopt a quantitative component (Smith, 2006). We return to the words of Amelia, that Grade 3 student who was quoted at the beginning of this chapter: The value of any method is if it is providing answers for the specific research question. Our question could not be answered without our quantitative and qualitative data sources working simultaneously. It is the very theoretical underpinnings of our work and the nature of our two-part question that demands such complex and varied methods: What are the essential elements of the SWH approach, and how do these elements support the learning of students in science, language, and literacy practices?

Our work continues. A new question has emerged focusing on the nature of scientific argumentation in high implementation classrooms. The new question once again requires a mixed-methods approach utilizing the tool of discourse analysis from the qualitative tradition with the RTOP scores from the quantitative tradition to begin a new investigation and shed new light on our persisting question: What makes the SWH approach successful with a variety of students from different backgrounds, ability levels, and experience? The questions persist—thus, so does our research.

References

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd edn.). Upper Saddle River, NJ: Prentice Hall.

Akkus, R., Gunel, M., & Hand, B. (2007). Comparing an inquiry-based approach known as the Science Writing Heuristic to traditional science teaching practices: Are there differences? *International Journal of Science Education, 29*(14), 1745–1765.

Bybee, R. W. (1995). Achieving science literacy. *The Science Teacher, 62*(7), 28–33.

Cavagnetto, A. (2006). *Setting the question for inquiry: The effects of whole class vs. small group on student achievement in elementary science*. Unpublished doctoral dissertation, University of Iowa, Iowa City.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education, 86*(2), 175–218.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*(3), 287–312.

Duschl, R. A., & Ellenbogen, K. (1999). Middle school science students' dialogic argumentation. *Proceedings of the 2nd international conference of the European Science Education Research Association "Research in science education: Past, present, and future"*, Kiel, Germany. Retrieved from http://www.ipn.uni-kiel.de/projekte/esera/book/regf.htm

Duschl, R. A., & Ellenbogen, K. (2002, September). *Argumentation processes in learning science*. Paper presented at the international conference on Ontological, Epistemological, Linguistic and Pedagogical Considerations of Language and Science Literacy: Empowering Research and Informing Instruction, Victoria, British Columbia, Canada.

Gowin, D. (1981). *Educating*. Ithaca, NY: Cornell University Press.

Greenbowe, T. J., & Hand, B. (2005). Introduction to the Science Writing Heuristic. In N. J. Pienta, M. M. Cooper, & T. J. Greenbowe (Eds.), *Chemist's guide to effective teaching* (pp. 140–154). Upper Saddle River, NJ: Prentice Hall.

Gunel, M. (2006). *Investigating the impact of teachers' practices of inquiry and non-traditional writing on students' academic achievement of science during longitudinal professional development program*. Unpublished doctoral dissertation, Iowa State University, Ames, Iowa.

Gunel, M., Akkus, R., Hand, B., & Norton-Meier, L. A. (2006, April). *Effects of teacher level of implementation of the science writing heuristic on students' performance on post-test and standardized tests*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, San Francisco, CA.

Hand, B., Hohenshell, L., & Prain, V. (2004). Exploring students' responses to conceptual questions when engaged with planned writing experiences: A study with year 10 science students. *Journal of Research in Science Teaching, 41*(2), 186–210.

Hand, B., & Keys, C. (1999). Inquiry investigation. *The Science Teacher, 66*(4), 27–29.

Hand, B., Norton-Meier, L. A., Gunel, M., & Akkus, R. (2006, July). *Examining the impact of teacher implementation on student performance on standardized testing when using the Science Writing Heuristic in K-6 science programs*. Paper presented at the annual meeting of the Australasian Science Education Research Association, Canberra, Australia.

Hand, B., & Prain, V. (2006). Moving from border crossing to convergence of perspectives in language and science literacy research and practice. *International Journal of Science Education*, *28*(2/3), 101–107.

Hand, B., Wallace, C. S., & Yang, E.-M. (2004). Using a Science Writing Heuristic to enhance learning outcomes from laboratory activities in seventh-grade science: Quantitative and qualitative aspects. *International Journal of Science Education*, *26*(2), 131–149.

Hohenshell, L. M., & Hand, B. (2006). Writing-to-learn strategies in secondary school cell biology: A mixed method study. *International Journal of Science Education*, *28*(2/3), 261–289.

Howe, K. R. (2003). *Closing methodological divides: Toward democratic educational research* (Vol. 11, Philosophy and Education Series). Dordrecht, The Netherlands: Springer.

Keys, C. W., Hand, B., Prain, V., & Collins, S. (1999). Using the Science Writing Heuristic as a tool for learning from laboratory investigations in secondary science. *Journal of Research in Science Teaching*, *36*(10), 1065–1084.

Lemke, J. L. (1990). *Talking science: Language, learning, and values*. Norwood, NJ: Ablex.

Levine, J. H., & Roos, T. B. (2002). *Introduction to data analysis: The rules of evidence*. Retrieved May 25, 2008, from http://www.dartmouth.edu/~mss/Volumes%20I%20and%20II%20.pdf

Loucks-Horsley, S., & Steigelbauer, S. (1991). Using knowledge of change to guide staff development. In A. Lieberman & L. Miller (Eds.), *Staff development for education in the '90s: New demands, new realities, new perspectives* (pp. 15–36). New York: Teachers College Press.

Mehan, H. (2001). "What time is it, Denise?": Asking known information questions in classroom discourse. *Theory into Practice*, *18*(4), 285–294.

Mertler, C. A., & Vannatta, R. A. (2002). *Advanced and multivariate statistical methods: Practical application and interpretation*. Los Angeles: Pyrczak.

No Child Left Behind Act of 2001. Pub. L. No. 107–110, 115 Stat. 1425. (2002).

Omar, S., & Gunel, M. (2004, January). *The impact of teacher implementation on student performance when using the Science Writing Heuristic*. Paper presented at the annual conference of the Association for the Education of Teachers in Science, Nashville, Tennessee, USA.

Piburn, M., Sawada, D., Falconer, K., Turley, J., Benford, R., & Bloom, I. (2000). *Reformed teaching observation protocol (RTOP)* (ACEPT Technical Report No. IN00-1). Tempe, AZ: Arizona State University, Arizona Collaborative for Excellence in the Preparation of Teachers.

Rudd, J. A., II, Greenbowe, T. J., & Hand, B. (2001). Recrafting the general chemistry lab report. *Journal of College Science Teaching*, *31*(4), 230–234.

Sanders, W. L., Wright, S. P., & Horn, S. P. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *11*(1), 57–67.

Saul, E. W. (Ed.). (2004). *Crossing borders in literacy and science instruction: Perspectives on theory and practice*. Newark, DE: International Reading Association & National Science Teachers Association.

Sheskin, D. J. (2004). *Handbook of parametric and nonparametric statistical procedures* (3rd edn.). Boca Raton, FL: CRC Press.

Smith, M. L. (2006). Multiple methodology in education research. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 457–476). Mahwah, NJ: Lawrence Erlbaum.

Songer, N. B., Lee, H.-S., & McDonald, S. (2003). Research towards an expanded understanding of inquiry science beyond one idealized standard. *Science Education*, *87*(4), 490–516.

United States National Research Council. (1996). *The national science education standards*. Washington, DC: The National Academies Press. Available from http://www.nap.edu/catalog.php?record_id = 4962

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wallace, C. S., & Narayan, R. (2002, September). *Acquiring the social language of science: Building science language identities through inquiry-based investigations*. Paper presented at the international conference on Ontological, Epistemological, Linguistic and Pedagogical Considerations of Language and Science Literacy: Empowering Research and Informing Instruction Victoria, British Columbia, Canada.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

Chapter 10
Approaching Classroom Realities: The Use of Mixed Methods and Structural Equation Modeling in Science Education Research

Martina Nieswandt and Elizabeth H. McEneaney

Teaching and learning in schools is influenced by various factors, such as social factors and inter- and intra-individual factors. Social factors comprise, for example, gender, race/ethnicity, social class, school norms, and beliefs/perceptions about schools and their members, while inter- and intra-individual factors include such characteristics as group dynamics, attitudes, interests, motivation, and perceptions about oneself and others (e.g., Arum & Beattie, 2000). In addition, specific curriculum expectations as well as hidden curriculum objectives impact teaching and learning as do other contextual factors, such as subject-specific national standards and national assessment movements, teachers' preferences for specific teaching and assessment strategies that are often deeply rooted in personal beliefs about teaching and learning, and the level of support of school administration. In general, classrooms are complex phenomena, complicating not only efforts to change practice but also the act of research itself. Turner and Meyer (2000) summarized bluntly, "Classroom research is messy" (p. 69). Researchers can either ignore the messiness and complexity of classrooms so as to concentrate on simple interrelations of two variables or the relation to an outcome variable (mostly student achievement) or they can investigate multiple variables from multiple perspectives using a multimethod approach. The latter vision must be embraced as part of any Gold Standard vision, with variables interpreted in relation to an understanding of the whole context.

A systematic review of current research on teaching and learning shows that most educational research has moved far from the old theoretical models that treated the classroom as a "black box" with data collection focusing on quantitative measures of "inputs" and "outputs" (Metz, 2000, p. 65). However, studies in educational research designed as randomized controlled trials with entire groups (e.g., classrooms, schools)—rather than individual students randomly assigned to

M. Nieswandt
Illinois Institute of Technology

E.H. McEneaney
California State University, Long Beach

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

treatment and control group by lottery—still operate as classic, black-box analyses. Reasons for or mechanisms producing posttest differences between two groups are seldom the focus of such large-scale studies. Instead, the focus is usually on examining programs rather than single treatments. Examples of this approach include studies identifying academic effects of vouchers on African American students who used them to switch from public to private schools (e.g., Howell, Wolf, Campbell, & Peterson, 2002).

The contributions of qualitative researchers using designs beyond the black box have especially advanced the agenda of understanding classroom processes in naturalistic settings, yet quantitative research strategies continue to be important in this age of educational accountability. In this light, we believe it is time to take mixed methods as the default research approach, to be used in the vast majority of classroom-based research studies in science education and, in fact, in all serious attempts to understand teaching and learning. Researchers who rely on an exclusively quantitative or an exclusively qualitative design should, as a routine matter, justify their decision *not* to use mixed methods—an approach that Johnson and Onwuegbuzie (2004) hail as “inclusive, pluralistic, and complementary” (p. 16).

Similarly, within the realm of quantitative research, we believe it is time to take our efforts to measure teaching and learning much more seriously. Researchers deeply engaged in understanding learning processes, especially at the classroom level, know that key concepts—such as meaningful science understanding and subject-specific self-concept (as characteristics of students), student-centered pedagogy and effective use of group work (as characteristics of teachers)—are not adequately measured with a single, closed-ended attempt. Scholars in the field know that these concepts are complex—debating their nuances endlessly and vigorously, perhaps with the hope that a single, simple measure will trump all others. We believe it is time to give up this fantasy and to acknowledge that virtually all the key concepts in teaching and learning are truly multifaceted and that solid research must use multiple measures of each concept in statistically valid ways. Structural equation modeling (SEM) of multiple measures is one such approach. In the following sections, we will make the case for mixed methods and sophisticated statistical methods with multiple measures using various examples from current research.

10.1 Characteristics of Mixed-methods Research

Mixed-methods research “combines quantitative and qualitative research techniques, methods, approaches, concepts or language in a single study,” and it has long been used as a matter of course by practice-oriented researchers in a wide variety of social and behavioral sciences (Johnson & Onwuegbuzie, 2004; Teddlie & Tashakkori, 2003, p. 5). Morgan (1998) noted, however, that research designs can be mixed (qualitative versus quantitative) at any or all of three stages in the research process: formulating the research objective, collecting data, or analyzing data.

Moreover, even in terms of data collection, Tashakkori and Teddlie (1998) emphasized that mixed-methods research need not combine equal measures of qualitative and quantitative data. Instead, designs can be predominantly one or the other simultaneously (i.e., in their notation QUAL + quan or QUAN + qual) or sequentially (e.g., QUAL/quan or QUAN/qual). Thus, there are myriad possibilities to structure the mixed-methods approach.

The reluctance of methodologists in educational and other forms of social research to embrace mixed methods has been attributed to the "paradigm wars" between quantitatively oriented "positivists" and qualitatively oriented "interpretivists"—sometimes labeled "constructivists" (Johnson & Onwuegbuzie, 2004; Tashakkori & Teddlie, 1998, pp. 6–13). (NB We use the term *interpretivist* here so as not to confuse the concept with *constructivist* perspectives on learning. Though some may disagree, we see a constructivist view of learning as conceptually distinct from a constructivist view of research.) Grounded in distinct ontological and epistemological assumptions, the social research paradigms were deemed incompatible (Guba & Lincoln, 1994; Smith & Heshusius, 1986), with the shrillest pronouncements emerging from educational research. Lines drawn in the sand of the paradigm wars discouraged, on philosophical grounds, the adoption of mixed-methods approaches—particularly for academic researchers who have a stronger preference for philosophically valid bases for their work than applied researchers.

In their argument on behalf of mixed methods, Johnson and Onwuegbuzie (2004) urged methodologists to catch up with practice-oriented researchers. They cleared the way for this move by proposing the pragmatism of William James and John Dewey (i.e., the value of an idea lies in its consequences) as a philosophical foundation for mixed methods, a notion first offered by Tashakkori and Teddlie (1998). In our opinion, giving mixed methods the philosophical cover of *if it works, use it* seems like less of a contribution than Johnson and Onwuegbuzie's elaboration of basic points of agreement between qualitative and quantitative researchers. These shared understandings between quantitative and qualitative researchers are, therefore, foundational for mixed-methods research:

- What seems reasonable can vary across persons.
- Observation is theory-laden and conducted by individuals embedded in communities and belief systems and so is not a perfect window into *reality*.
- More than one theory can fit a single set of evidence.
- Evidence is only probabilistic, not *proof*.

Johnson and Onwuegbuzie (2004) described an eight-step, mixed-methods research process model:

> (1) determine the research question; (2) determine whether a mixed design is appropriate; (3) select the mixed-method or mixed-model research design; (4) collect the data; (5) analyze the data; (6) interpret the data; (7) legitimate the data; and (8) draw conclusions (if warranted) and write final report. (p. 21)

The strength of such a mixed-methods process stems from what they call the "fundamental principle" that researchers should "collect multiple data using different strategies, approaches and methods [so that] the resulting mixture … is likely to result in complementary strengths and non-overlapping weaknesses" (p. 18). If the paradigm purists have been correct, then mixing quantitative strategies with qualitative strategies yields the benefit of complementarity only if the researcher examines the design's fundamental assumptions and then determines which design is best for the research question(s). Greene, Caracelli, and Graham (1989) discussed five purposes for mixed-methods designs, which seem useful for making such design decisions:

- Triangulation—to seek convergence of results across methods.
- Complementarity—for one method to enhance or illustrate the results of another.
- Development—timed sequentially, results of one method inform the development or analysis of the other method.
- Initiation—in contrast to triangulation, to seek paradox and contradiction across results of different methods in order to generate new research questions.
- Expansion—to extend the range of inquiry by using different methods for different components of the study.

A key point by Greene and colleagues, and one that we want to emphasize, is that the research design must be deliberately crafted to achieve one or more of these purposes. The most powerful benefits of mixed-methods research do not emerge when methods are mixed post hoc.

In the next section, we focus on the first four design purposes and offer examples of various mixed-methods designs in practice. The fifth purpose (expansion) is often used in program evaluation studies with quantitative methods to assess the program outcomes and qualitative methods to assess the implementation. In general, such reports are not published in the journals that we reviewed or each component is reported in separate articles, which is beyond the purpose of mixed methods.

Based on the pragmatist paradigm, a vast variety of mixed-model designs is possible. For the sake of brevity, we focus our discussion on studies using mixed methods for data collection and data analysis. Thus, we will not discuss mono-method data-collection designs that are followed by mixed-methods data analysis. An overview of the studies that we offer as examples of various approaches to mixed methods research is presented in Table 10.1.

10.2 "Value Added": Mixed Methods in Practice

In order to find examples for the use of mixed methods in science education research, we reviewed research articles in *Science Education* (SE) and the *Journal of Research in Science Teaching* (JRST) published in 2006 and in 2007 (including the November issues of both journals). The review was based on three criteria: indication of mixed-methods research in the abstract, qualitative and quantitative data collection, and qualitative and quantitative data analysis. By no means does

Table 10.1 Overview of mixed-methods studies discussed

Study discussed	Mixed method characteristics	Primary purpose
1. Clary and Wandersee (2007)—Petrified wood in geology instruction	QUAN + QUAL Pre- and postinstruction questionnaire, examination scores; minilab discussion and discussion board content and field notes	Triangulation
2. Lee, Luykx, Buxton, and Shaver (2007)—Professional development to incorporate home language and culture	QUAN + QUAL Focus groups, classroom observations; questionnaires	Triangulation
3. Nieswandt (2008)—Preservice teachers' beliefs about science and science teaching	QUAN/qual Questionnaires with subsequent semistructured interviews using questionnaire answers as point of departure	Complementarity
4. Gilmartin, Denson, Li, Bryant, and Aschbacher (2007)—Representation of women teachers and students' science identity	QUAN/qual Questionnaires with follow-up interviews of subset of participants	Complementarity, emergent initiation
5. Shaver, Cuevas, Lee, and Avalos (2007)—Teacher perception of policy influences	QUAL + quan Focus groups and questionnaires administered in parallel over 2 years	Complementarity
6. Tretter, Jones, Andre, Negishi, and Minogue (2006)—Concept of scale of scientific phenomena	QUAN/QUAL Written questionnaires informed by subsequent explanations of card sort strategies	Development
7. Talanquer (2008)—Predictions of sensory properties of chemical compounds	QUAN + qual/QUAN Questionnaires, followed by interviews, used to develop final questionnaire	Development
8. Winberg and Berg (2007)—Computer-simulated activity and learning outcomes	QUAN/QUAL Attitude questionnaire, used to select interviewees; recorded student questions during laboratory work	Development
9. Nieswandt and Shanahan (2007)—Affect-motivation-cognition among high school science students, followed by Nieswandt and Turner (2008)	quan/QUAL Written survey questionnaires and interviews (Nieswandt & Shanahan), with revisions to questionnaire items based on interviews (QUAN/QUAL) (Nieswandt & Turner)	Development
10. Onwuegbuzie and DaRos-Voseles (2001)—Cooperative learning in graduate research methods	QUAN/QUAL In-class examinations; student reflective journals	Emergent initiation

this review provide a systematic analysis of the types of methodology being used in science education research; however, it presents examples of the most recently published studies using mixed-methods designs. In addition to this journal review, we did a search in Education Resources Information Center (ERIC) to find additional studies in science education; we limited the search to the last 2 years and used the same criteria. We now turn to a discussion of these examples, as outlined in Table 10.1.

10.2.1 Triangulation

Triangulation, a method used in qualitative research in order to verify results that are based on various data sources (e.g., observations, interviews, focus groups), is viewed in mixed methods as "parallel mixed analysis" (Tashakkori & Teddlie, 1998, p. 128). It is the most widely used mixed data collection and analysis strategy, a combination of QUAL and QUAN data. The design usually uses separate quantitative and qualitative instruments, which are then analyzed concurrently and the findings compared. More advanced designs might also include principles of complementarity or even development.

Our first example of a mixed-methods study that fulfills the purpose of triangulation is Clary and Wandersee's (2007) exploratory study. They investigated in three consecutive introductory college geology classes "whether petrified wood … could serve as a portal to deeper understanding of the geobiological topics of fossilization, geologic time, and evolution" (p. 1012). They chose a mixed-methods research design in order "to crossvalidate, confirm, or corroborate their findings within a single investigation" (Creswell, Clark, Gutmann, & Hanson, 2003, as cited in Clary & Wandersee, p. 1016). During the fall of 2003, a pretest and a posttest "Petrified Wood Survey" that was quantitative in nature was administered in a geology class that followed the traditional curriculum. The fall 2004 class became the experimental group; petrified wood was used as "the portal for geobiological learning" (p. 1017). A modified version of the petrified wood questionnaire was administered prior to and at the conclusion of instruction and analyzed quantitatively. In addition, the researchers used a variety of other instruments: an end-of-semester survey that mapped students' project feedback and was analyzed qualitatively, a minilab and a minilab follow-up electronic discussion (qualitative analysis), final examination results (quantitative), discussion board feedback that was analyzed qualitatively, and researchers' field notes. Triangulation of these multiple data results supported their findings that "petrified wood shows promise as a topic and material that can serve as a springboard to better student understanding of fossilization, geologic time, and evolution" (p. 1031). In Tashakkori and Teddlie's (1998) notion, Clary and Wandersee's design would be QUAN + QUAL, or qualitative and quantitative data have equal status but are administered in parallel with the aim to seek convergence of results.

This design added value to the researchers' understanding of students' learning in several ways. For example, the individual item analysis of the pre- and postinstructional questionnaire across the semesters revealed that the students receiving the special instruction gained the greatest knowledge. The item analysis also indicated in which concept area the students made progress. The electronic discussion board and the minilab discussion comments provided a more differentiated picture of the structure of students' scientific understanding, their alternative conceptions, and how both developed over time. In addition to triangulation, Clary and Wandersee's mixed-methods design also fulfilled the purpose of complementarity; thus, one result enhanced the results of others.

Another example of triangulation across different methods is Lee, Luykx, Buxton, and Shaver's (2007) study that examined the impact of a professional development intervention aimed at helping elementary teachers incorporate elements of students' home language and culture into science instruction. The researchers conducted focus group interviews with teachers, administered a questionnaire to the teachers, and conducted classroom observations. All three instruments operationalized and measured the same construct. Results of all three instruments did in fact converge, and Lee and colleagues demonstrated that teachers did not change their perceptions or their instructional practices throughout the 2-year intervention. In addition to the purpose of triangulation (QUAN + QUAL), the study followed a complementarity principle. While the questionnaire and interviews focused on teachers' perceptions and beliefs, the classroom observations concentrated on the extent to which the teachers integrated students' home language and culture into their instruction. Although teachers expressed the importance of such integration, classroom observations revealed that they generally did not incorporate students' home language and culture in their instruction. The additional qualitative classroom observations provided Lee and colleagues a rich view of how teacher beliefs are not seamlessly incorporated into practice, thus adding much value and nuance to their quantitative results.

10.2.2 Complementarity

The review of articles revealed that a majority of studies using a mixed-methods design favored the principle of complementarity, although this was not stated explicitly. It seems to be an emerging trend to follow a design in which data from a questionnaire are being enhanced and/or illustrated with interview data of a sample of participants (e.g., Barnett et al., 2006; Sadler & Donnelly, 2006; Trumper, 2006; Wu & Huang, 2007). Using Tashakkori and Teddlie's (1998) notation, this design would be QUAN/qual or predominantly quantitative followed sequentially by a less dominant, qualitative component. Such a design allows for developing a deeper understanding of key components that are being addressed in the questionnaire or explains the trends being depicted through the questionnaire data. In particular, while questionnaire data on teaching and learning might establish inputs and outputs to the black box, the qualitative data might address the process or mechanism linking the two. However, complementarity can also be achieved through enhancement of the dominant method by implementation of the less dominant method, for example, by using quantitative survey results to select purposively focus group participants with a wider range of experiences than might have occurred through random selection.

In classroom research using mixed methods conducted by Nieswandt (2008), semistructured interviews were conducted after quantitative written surveys were completed. The quantitative surveys were the dominant component of the project, with the subsequent interviews constituting a less dominant component (QUAN/qual). A key value added was that the interviews allowed opportunities to probe participants in particular areas, such as preservice teacher candidates' beliefs about how they want to teach science. These interview questions were able to build on the questionnaire items and use participants' prior questionnaire answers as a starting point. For example, the science teachers participated in a series of interviews throughout both their 9-month teacher education program and first year of teaching regarding how they would like to teach science, how they were able to teach science during their school practica and their first 2 years of teaching, and whether they were able to transfer their ideas about teaching into their teaching practice. The written, closed-ended questionnaire included more general questions about their beliefs about teaching science, which then in combination with the individual perspectives drawn from the interviews produced a specific, enriched view of their general beliefs. The ability to probe in an interview setting furthered the interpretive aim to understand deeply how research participants attached meanings to terms, especially those related to pedagogy and the nature of science. Since this was a longitudinal study—participants were followed into their first 2 years of teaching; the iterative use of survey and interview highlighted changes in the meanings that participants created. This would be an example where the mixed-methods design achieved Greene and colleagues' (1989) purpose of complementarity.

Another example of complementarity is Gilmartin, Denson, Li, Bryant, and Aschbacher's (2007) study that examined how the representation of women among secondary school science teachers affects aspects of their students' science identity. They tested their hypotheses with a questionnaire that included measures for eight dependent variables, such as students' perceptions of science in and beyond the classroom, their interest in studying science or engineering during college, their self-concept as a future scientist, and their performance in science class. In addition, Gilmartin and colleagues conducted interviews with a subsample of participants in which they expanded on some of the questionnaire topics (e.g., students' science attitudes and views, experiences in past and present science classes, plans for the future) and went beyond the questionnaire items in other areas (e.g., students' role models, social networks, family contexts) or QUAN/qual. Results of the questionnaire analysis conducted with hierarchical linear modeling contradicted their hypotheses: the results "indicated that the percent of female science faculty did not have an effect on multiple components and mediators of students' evolving science identities, nor did it affect gender differences in these measures” (p. 1000). Analysis of the interview data provided some explanation for these results: The teacher–student relationship and the teachers’ real-life science experiences and science credentials were more important than the sex of the teacher—students responded positively to science teachers who were caring, challenging, engaged, passionate, and fair. A second finding of the questionnaire data revealed that for some dimensions of students’ science identities (e.g., college major interests, science grades, science class perceptions) a small but significant difference was attributable to school contexts. Female teachers were more likely to be found at schools with a greater proportion of lower-performing and lower-income students. Again, qualitative interview data enhanced these results by indicating that at one of these schools “discipline was heavy, equipment could be scarce, and students perceived a lack of schoolwide support for science” (p. 1002). The key value added in this study was the strength of the qualitative data to provide initial explanations for the quantitative findings.

Studies that would be categorized as QUAL + quan or QUAL/quan, that is, where quantitative data are used to support qualitative findings statistically, either conducted simultaneously in the first instance or sequentially in the second, were less common in the literature. One such example is Shaver, Cuevas, Lee, and Avalos’ (2007) study investigating elementary teachers’ perceptions of policy influences on science instructions with a majority of English language learners. The researchers conducted a series of focus group interviews and administered questionnaires to the participants over a period of 2 years while the teachers were involved in a professional development intervention or QUAL + quan. Focus group interviews concentrated on teachers’ perceptions of policy influence on science instruction with culturally and linguistically diverse students; the questionnaire assessed teachers’ perception of their own knowledge and importance of policies. Shaver and colleagues analyzed both datasets independently and found that the quantitative results provided statistical support for the qualitative results. Both datasets indicated that throughout the project teachers consistently viewed standards as beneficial but asserted their right to use the standards at their own discretion; they also expressed more and more negative opinions regarding the need for test preparation for statewide assessment and accountability. Using quantitative data to support qualitative data adds value to the impact of the results, which in this case indicates the importance of considering teachers’ views in the debate on educational reform and policy evolution.

10.2.3 Development

An example of how one method informs the development or analysis of a second method is Tretter, Jones, Andre, Negishi, and Minogue’s (2006) study investigating middle school, secondary school, and doctoral students’ conceptions of scale of scientific phenomena. A questionnaire assessing students’ perception of size ranges for various objects was followed by a card sort task “completed individually in a location separate from other participants to maintain independence of results" (p. 289). For this task, students were asked to sort the objects on the cards according to similarity in size and to create piles with similar-sized objects. As soon as students were satisfied with their sorted piles, they were asked to explain their reasoning for their decisions. After an analysis of the questionnaire data using descriptive statistical methods, the researchers created quantitative categories clustering objects representing the concepts of scale, such as big, field size, room size, and small. They then used the interview data of students' strategies for creating similar-sized piles of objects in order to interpret and contextualize each category's rationales. The value added to this design was that the results of the card sort task and the interviews informed the analysis of the questionnaire data (development purpose) and at the same time complemented each other or QUAN/QUAL.

Talanquer's (2008) study of postsecondary students' predictions about sensory properties of chemical compounds is another example of a mixed-methods design that follows the principle of development. Like some of the other studies discussed above, this study also achieves complementarity. Talanquer used interviews to enhance his questionnaire data in a first study. Results of the initial part of the study resulted in the development of a final questionnaire, which was used with various populations in supplemental studies reported in the same article. Results not only confirmed the validity and reliability of the test items but also confirmed the results of the first study. The "majority of the students' answers to the sensory property questionnaire are consistent with the use of an additive framework in the prediction of color, smell, and taste of the product of a chemical reaction" (p. 110). Analysis of the interview transcripts revealed that these students believed that chemical compounds are mixtures of substances that preserve some of their original properties in the final product. Thus, the interviews enhanced the quantitative results and at the same time provided the researcher with more insight into students' everyday conceptions that guided these conceptions, a clear value added or QUAN + qual/QUAN.

Mixed-methods research fulfilling a development purpose can have various designs, as demonstrated above. Another example is Winberg and Berg's (2007) study on enhancing students' learning outcomes with respect to content knowledge using a computer-simulated activity. They administered an attitude questionnaire to determine student attitudes prior to the simulation, measured the effects of the simulation on students' content knowledge through interviews, and recorded and classified students' questions to teachers during subsequent laboratory work in order to determine whether the simulation had any effect on students' cognitive focus during the laboratory work. The value added of this study can be seen in the sequence of the instruments. Based on previous studies, the researchers assumed that students' attitudes toward learning and knowledge influence the outcome of the simulation. In order to test this assumption, they administered the attitude questionnaire at the beginning of the course and, based on the students' responses to the questionnaire, sampled two groups of participants for the interviews: students with low- and high-attitude positions. This procedure allowed the researchers "to maximize the contrast between groups" (p. 1119). This study is a good example in which the mixing of methods helped to refine subsequent methodology rather than just substantively adding to the results or QUAN/QUAL.

When a mixed-methods approach aims for development, the mixed character of the resulting, published study can sometimes be obscured as the editorial process makes distinctions between the back-stage analysis and the front-stage results of the research project. An example is the affect-motivation-cognition study by Nieswandt and Shanahan (2008) that investigated Grade 11 general science students' goal orientation and factors influencing these goals. Implemented as a mixed-methods study or quan/QUAL of a small class (*N* = 13) using questionnaires and interviews, the quantitative results proved so insubstantial that only the qualitative results were reported in the final published article. Specifically, questionnaire items intended to measure the key concept of instrumentality (e.g., *I do the work assigned in this science class because good grades lead to other things that I want—graduation, university acceptance, money, good job*; *I do the work assigned in this science class because my grades have a personal payoff for me—rewards from my family, graduation, scholarships*) suffered from severe ceiling effects as all participants scored at the maximum on this dimension. However, questionnaire results were used analytically to develop motivation profiles that were centrally useful in the analysis of subsequent qualitative interviews. These interviews allowed a much more nuanced understanding of instrumentality, with participants reporting that the science course had instrumentality for them both as a diploma requirement and as background knowledge for future social and work interactions. It was, however, instrumentality due to the diploma requirement that determined the boys' motivation and effort in the course. This motivation did not, however, result in more effort; students did only what they needed to in order to complete the course. This contrasts with research suggesting that perceived instrumentality precipitates increased motivation and engagement. Thus, in the end and not by design, the developmental aspect of this study lay purely in the way the quantitative component contributed to the analysis of the qualitative part of the study. Yet, as is true for many single-method qualitative studies published in journal article form, description of the gritty, behind-the-scenes analysis process of interview data (in this case, the interplay between ostensibly flawed questionnaire data with the interviews) was not deemed worthy of publication.

Development need not be limited to stages within a single study. In follow-up studies (Nieswandt & Turner, 2008), the questionnaire will be revised based on the qualitative results to integrate more specific items addressing this aspect of instrumental goal orientation. The developmental benefits of mixed-methods approaches will be fully realized, therefore, over the course of several related studies within a research program. The aim of development here can be seen in terms of methodology, but it can also be achieved in the area of theory.

10.2.4 Initiation

The purpose of initiation is to seek paradox and contradiction across results of different methods in order to generate new research questions. Our analysis of journal articles did not reveal a study that followed this purpose, which can be a result of either the short period of time that we covered (2006 and 2007), a matter of what is being presented in a journal article, or whether such a design purpose is meaningful for educational research or valued in the editorial review process. Greene and colleagues (1989) stressed that "purposeful initiation may well be rare in practice" (p. 268) while a more emergent initiation design may be more likely. An example of such an emergent initiation design is the Gilmartin and colleagues' (2007) study, which was described as an example of a complementarity design. Results of the questionnaire analysis contradicted the researchers' hypotheses; however, the analysis of the interview data provided some explanation for these results. Thus, the value added lay in the contradictory, complex findings that raised serious questions about the previous theory, justifying future rethinking of the theory with the possibility of a fundamentally new theory emerging.

Another example of an emergent initiation design is Onwuegbuzie and DaRos-Voseles' (2001) study of the effectiveness of cooperative learning (CL) in a graduate research methodology course in comparison to a course in which all assignments were done individually (IL). The researchers used a parallel mixed analysis of the datasets or QUAN + QUAL. The quantitative analysis revealed that students in the IL group obtained higher scores on the two in-class examinations, while the qualitative analysis of reflective journals in the CL group revealed positive and negative experiences with most students (70%) tending to have overall positive attitudes toward cooperative learning. While the statistical analysis demonstrated that CL techniques led to decreased performance, the qualitative analysis suggested that the majority of the class liked CL techniques. The researchers concluded: "The fact that students appear to like cooperative learning techniques despite not experiencing increases in their level of performance … suggest[s] that, for some students, the non-cognitive outcomes may be as important as subject matter achievement" (p. 72). The researchers also remarked that weaker students in particular liked cooperative learning although they did not increase their performance level. Thus, noncognitive factors, which were "not compatible with the instructional objectives of this method" (p. 72), seemed more important than achievement levels. The researchers would have missed this contradiction if they would have relied on only one method; and the results remind us of the complex nature of the classroom: among teacher and students there are social, noncognitive objectives in addition to instructional, cognitive objectives.

10.2.5 Difficulties in Implementation of Mixed Methods

To be sure, a mixed-methods approach may present difficulties to the classroom researcher. The major issue is that this approach—and in particular its qualitative component—requires a greater investment of time on the part of the researcher in various stages of the study. First, combining quantitative and qualitative components multiplies the number of decisions about the appropriate research design that must be made based on the research questions. Tashakkori and Teddlie

(1998) and Johnson and Onwuegbuzie (2004) outlined the ways in which this is true—although both sources view this as a strength of mixed methods rather than an obstacle. Second, data collection is likely to take considerably more time, especially when compared to relying solely on a standard quantitative survey. Third, some forms of qualitative data collection are seen as more invasive and more difficult to assure confidentiality; therefore, securing ethical approval from school officials to, for example, videotape classroom observations or students' interactions in group work tasks can be a difficult, time-intensive project. Finally, analyzing and synthesizing two distinctive types of data requires more time. Qualitative data analysis, particularly from a grounded theory perspective, in which themes emerge from the data rather than being imposed upon the data, is notoriously time consuming. Additional time requirements obviously increase the cost of such projects. Standard qualitative data analysis requires full or substantially partial transcription of, for example, interviews and focus group sessions, which is a pricey endeavor.

Aside from cost and time demands, the mixed-methods literature has yet to offer substantial tips on how to analyze data from a thoroughly mixed-methods perspective, leaving researchers to switch from specialized qualitative techniques, such as Nieswandt and Bellomo's (in press) six-step procedure for analyzing written extended response questions, to statistical quantitative techniques as the data require, a less than ideal bifurcation. However, some progress has been made, including strategies developed by Caracelli and Greene (1993) in the area of evaluation research and by Onwuegbuzie and Leech (2004) more generally about the meaning of significant findings in mixed-methods research. The examples of studies in science education as described above highlight concurrent, parallel, or sequential mixed-methods analysis. While each of them has its value depending on the design of the study and the research questions, all are equally time consuming despite advances in software for statistical and qualitative data analysis. A truly synthesized approach to analyzing mixed-methods data is still needed.

10.2.6 Barriers to Communication of Mixed-methods Results

For all of the potential value of a mixed-methods approach, communicating the full richness of data generated with this design can be difficult. Despite the recent quelling of the so-called paradigm wars, we contend that there are still two mostly distinct scholarly audiences—qualitatively trained researchers and quantitatively oriented researchers—and that the gap between these groups must be addressed as a first step. In some ways, this gulf approaches C. P. Snow's (1959/1998) classic description of two cultures—the humanists and the scientists. Although less of a problem in science education, many educational and sociological journals have clear identities as qualitative or quantitative, making it difficult to reach both audiences with a single publication. We believe that the greater challenge, in general, is to communicate the value of qualitative results to statistically oriented researchers. Backed by the assurance that their research activities count as science, either within a positivist or postpositivist paradigm (Guba & Lincoln, 1994), quantitatively oriented researchers (and journal reviewers) insist on standards of validity, reliability, and generalizability that remain at odds with qualitative analysis. At the same time, quantitative scholars tend to discount the power of *thick description* and *verstehen* (deep understanding) that qualitative research often generates. On the other hand, we have both experienced situations when reviewers, presumably better trained in qualitative methods, demanded the inclusion of lengthy explanations of statistical techniques that assumed virtually no background knowledge, a sure way to extinguish the elegance of any piece of scholarly work.

These conflicts dramatically influence the very last stage of mixed-methods research: the writing. Our sense is that research projects that are fundamentally mixed in their conception and implementation are forced into qualitative and quantitative boxes during the writing phase, with savvy researchers spinning off qualitative or very much predominantly qualitative analysis into one publication and the lion's share of quantitative analysis packaged separately (see Nieswandt & Shanahan, 2008). Ideally, of course, the written communication of results should be as mixed as the research stages leading to that point. To reach this ideal, it is vitally important that the methodological development of a distinctive mixed-methods paradigm continues, rather than relying on an uneasy patchwork of tenets that are often not acceptable to substantial numbers of researchers. Sandelowski (2003), in her chapter on writing and reading mixed-methods studies, concluded: "Writing mixed methods studies requires an understanding of differences, of how aesthetic considerations enter into the creation of convincing write-ups of both qualitative and quantitative research, and of whether and how diverse aesthetic sensibilities can be brought together" (p. 344).

10.3 "Value Added": The Case for Multiple Measures and Structural Equation Modeling

We move on now to the case for structural equation modeling, as a subordinate element of an overall strategy of using mixed methods. In the following discussion, our intent is to encourage mixed-methods researchers to consider employing more complex statistical models like SEM, models that acknowledge the complexity of real-life teaching and learning. Readers who are persuaded by our case for SEM should consult sources such as Schumacker and Lomax (2004) for a complete guide to the theory and implementation of this statistical method.

The point of departure for structural equation models is the notion that most concepts of interest in social research should properly be thought of as latent variables (Bollen, 1989; Schofer & McEneaney, 2003). Latent variables, also known as constructs or factors, are variables that are "not directly observable or measured

… [they] are inferred from a set of variables that we do measure" (Schumacker & Lomax, 2004, p. 3). Why are latent variables particularly relevant to classroom-based research on teaching and learning? We believe that using latent variables that are inferred from sets of multiple, direct measures acknowledges the complexity of the classroom setting and the identities of teachers and students within that setting in a fundamental way. For this reason, we see the use of SEM as an ideal extension of the mixed-methods approach.

In this section, we describe briefly examples from the science education literature that will help illustrate ways in which SEM can be used to better reflect complexity in teaching and learning. These examples include studies by Baumert, Evans, and Geiser (1998), Mattern and Schau (2002), and Nieswandt (2007). A student's use of television, everyday experiences, and control beliefs regarding technology were all treated as latent variables in Baumert and colleagues' analysis of technical problem solving, which itself was treated as a latent variable. The latent variable of technical everyday experiences was based on two separate scales: (a) construction with technical objects and (b) creativity with technical objects, while use of television was measured not only by average number of hours of television watched daily but also by number of programs. Finally, the latent variable of technical control beliefs was based on a 6-item scale directly measuring self-concept of technical ability and a 12-item inventory measuring attributions about failures in technical–mechanical encounters. In each case, a multiple-measure approach honors the complexity of the concept. It broadens and enriches our understanding, for example, of what everyday experiences might contribute to better problem-solving by developing a model that reflects both experiences constructing technology *and* being creative with technology. As our theories about teaching and learning become more sophisticated, testing these theories with singly measured concepts is not likely to advance the field.

10.3.1 Mirroring Complexity Statistically

Yet, it is not enough to acknowledge complexity with multiple measures (or indicators) of a latent variable. These indicators need to be handled in a valid way statistically. SEM enables testing models that more closely mirror the complexity of classroom realities by:

- Allowing indicators to contribute differentially to latent variables.
- Allowing specification of measurement error.
- Permitting correlated error between indicators, especially useful for longitudinal studies.
- Allowing more sophisticated modeling of group differences, interaction effects, and noncontinuous indicators.

We briefly discuss each in turn.

10.3.1.1 Indicators Contribute Differentially to Construct

Like its less sophisticated analytic cousin, exploratory factor analysis, SEM allows each indicator to contribute more or less strongly to the latent variable. In nearly all cases, this is far preferable to simply summing indicators together, which forces all indicators to be weighted equally. The weighting is calculated relative to a reference indicator, whose weight is typically set to 1. This reflects the reality of classroom-based research, because our theories are not usually so detailed as to justify particular weights for indicators. Theory comes into play because indicators must be specified for each latent variable; but the SEM software calculates weights that best fit the data, thus confirming whether the selection of indicators was appropriate or not. In some ways, this strategy of allowing the weighting of indicators to emerge from the data is similar to the qualitative researcher's grounded theory approach.

10.3.1.2 Acknowledging the Ubiquity of Measurement Error

Structural equation models also allow for explicit specification of measurement error. Having multiple indicators of a construct does not eliminate error; but in a structural equation model, every observed indicator can be specified as having an error component. In contrast, for example, an ordinary least squares regression model only acknowledges that there is some fuzziness or error in the dependent variable and that the independent variables are measured perfectly. Researchers collecting data in classroom situations from distracted students, for whom your research study is not nearly as important as whether they can sit with friends at lunch, or from harried teachers who can only spend 2 minutes on your 10-minute survey know instinctively that every response gathered is covered with a bit of random fuzziness. Instead of ignoring this unfortunate fact of an educational researcher's life, a structural equation approach allows us to build this reality into the model.

10.3.1.3 Admitting that Measurement Error Can Be Repeated

In general linear models such as regression analysis or ANOVA, a mathematical assumption that makes the model work mathematically is that the error terms are uncorrelated (i.e., where the error term is the component of the dependent variable left unexplained by the independent variables). Combined with the assumption that there are no explanatory variables omitted from the analysis, these two conditions are very rarely met in social science research. This is true because humans tend to be embedded in a web of interrelated characteristics. The academic support students receive from their families is certainly related to the students' socioeconomic status; and attempts to measure these two characteristics are likely to be biased for the same reasons, for example, lack of fluency in the language in which the survey is written. Both constructs are known to influence academic achievement. If one simply creates a scale or index to measure family support, socioeconomic status, and achievement, a regression analysis might show statistically significant effects on achievement. However, those estimates of significant effect are likely to be biased (i.e., wrong) because regression requires that there is no correlated error between family support and socioeconomic status. Instead of ignoring this violation of the assumption, SEM deals with the complexity by allowing the analyst to build this correlation into the model.

This feature of SEM does not resolve the problem of omitted variable bias. Perhaps the most common class of omitted variables is variables characterizing different levels of analysis. For example, statistical analysis based on data collected on students often fails to control for classroom-level or neighborhood-level characteristics. An explanation of hierarchical or multilevel models and latent growth curve models is beyond the scope of this chapter. However, we want to point out that contextualizing individual-level processes within broader social circumstances is a common reason for undertaking qualitative—especially ethnographic—research, thereby providing another reason to adopt a mixed-methods approach.

Another extremely important situation in which researchers are faced with correlated error is in longitudinal research, where measures of the same latent variable administered at different points in time are quite likely to have correlated error (Arbuckle & Wothke, 1999; Schumacker & Lomax, 2004). For instance, suppose one attempts to measure family support at the beginning of a school year with three indicators: degree of parent involvement with school personnel, nights in a typical week that the family sets aside specific time for homework, and frequency of discussing science topics at home. Imagine as well that respondents to these items tended to *fudge* answers a bit, maybe overstating how often time is set aside for homework. That measurement error is very likely to occur in the same way—the errors are correlated—if we use the same or similar indicators at the end of the school year to measure family support. SEM allows researchers to build this tendency into the model rather than ignoring it. This strategy of not simply taking measures at their face value is a bit like the qualitative interviewer who knows to discount a research participant's words of praise for a new instructional module when they are uttered with a slight smirk. In the science education literature, good examples of this use for SEM are Mattern and Schau's (2002) longitudinal analysis of the relationship between science attitudes and science achievement or Nieswandt's (2007) study exploring the relationship between affective and cognitive variables in Grade 9 chemistry students.

10.3.1.4 Modeling Group Differences, Interaction Effects, and Various Types of Indicators

There are additional reasons for SEM's utility. Recent advances in the field have made yet more sophisticated models possible, such that tests of group differences (e.g., Arbuckle & Wothke, 1999) and interaction effects (e.g., Schumacker & Lomax, 2004) are possible. Mattern and Schau (2002) make use of SEM by comparing the model results for boys against the results for girls. They found that the relationship between attitudes and achievement over time varied between girls and boys. While earlier versions of software packages like LISREL, EQS, and AMOS only permitted continuous scale indicators, newer versions allow use of indicators at various levels of measurement (e.g., dichotomous "dummy" variables) in "mixture models" (Schumacker & Lomax, p. 342). Also, advances in the user-friendliness of software have made the analysis process more intuitive, though it has not obviated the need to understand the underlying mathematics. In our opinion, the intuitive graphical interface first introduced in AMOS represented a great step forward.

Thus, although the technical details of the way in which SEM mathematically mirrors the complexity of classroom research can be daunting, we believe that it is worth the effort to master these details. Embracing complexity in the statistical modeling allows quantitative work to avoid the charge of being reductionist and perhaps paves the way to easier linkages between quantitative and qualitative approaches in mixed-methods research.

10.3.2 Difficulties in Implementation of Structural Equation Models

For researchers who received statistical training in the general linear model tradition (e.g., correlation, ANOVA, multiple regression, logistic regression, loglinear models), it is understandable if SEM is not easily grasped on an intuitive level. While there are intuitive bridges that lead between ANOVA and multiple regression or between multiple regression and logistic regression, nonstatisticians may think that they have to build their own conceptual bridge to SEM!

There are several key differences between SEM and general linear models that may impede researchers who want to add SEMs to their analytical repertoire. The latent, unmeasured character of the key variables in SEM might pose some conceptual difficulties for novices. The essential mathematical engine that drives SEM is to compare the observed variance–covariance matrix (comprised of the variance of all observed indicators and the covariance between all pairs of indicators) with the variance–covariance matrix that one would expect if the specified model were correct. In essence, the null hypothesis in SEM is that the model fits the data well and that the observed and expected variance–covariance matrices are quite similar. This contrasts substantially with the general linear model approach where the null hypothesis is that the model does not fit well. As a result, for example, a good fit for an SEM has a χ^2 test with high p values, much above 0.05, while the objective in fitting general linear models is to achieve a fit statistic with a low p value.

Another intuitive difficulty for regression and ANOVA users is that there is no single fit statistic for SEM that enjoys the authority among scholars that R^2 and F

tests enjoy in a general linear model framework. SEMs are complicated enough mathematically that the various proposed fit statistics assess different desirable aspects of the model. Guidelines suggest appropriate fits with Comparative Fit Index (CFI; model compared with independence model) greater than 0.95 and Root Mean Square Error of Approximation (RMSEA) less than 0.05. With the Parsimony-adjusted Comparative Fit Index (PCFI), values closer to 1 represent a comparatively better fit (see Schumacker & Lomax, 2004, pp. 79–106, for a detailed discussion). Researchers must, therefore, report a set of fit statistics as they document their results.

Unlike general linear models, SEMs have two fundamentally distinct levels. The first part is the measurement model, which estimates how the measured indicators load on their respective conceptual factors (or latent variables). The second part is the structural model, which estimates how the set of latent variables and any single-measure, nonlatent variables relate to one another. When a model does not appear to have an acceptable fit, the problem may lie at either or both levels of the model. This complicates the process of specifying a model that fits the data well, in a way that has no analog in general linear models.

There are relatively few threats to the stability of general linear models. Collinearity has to be rather extreme to be judged unduly influential in an ordinary least squares regression. Specifying a model that converges (i.e., produces a unique set of parameter estimates) can be very challenging in SEM. The model must be *identified*. In brief, the observed variance–covariance matrix must have enough information to estimate the model parameters required (otherwise the model is *underidentified*) while not introducing so many constraints that no unique solution can be found (otherwise the model is *overidentified*). There is a range of choices for how to estimate the model once it is identified, such as maximum likelihood and generalized least squares. In ordinary least squares (OLS) regression, these two estimation methods produce the same result; but in SEM they are likely to produce substantially different results; therefore, the analyst should justify the estimation method used a priori. Finally, a variety of conditions can result in estimation of a nonpositive definite variance–covariance matrix, making the solution inadmissible (Schumacker & Lomax, 2004). In short, simply coaxing an admissible set of estimates from a structural equation model can be an achievement in itself, unlike ANOVA or regression.

Perhaps most challenging of all is that SEM estimates are very sensitive to the number of cases. Ding, Velicer, and Harlow (1995) raised considerable concern about bias in model estimates where there are fewer than 5 or 10 cases per parameter estimated. Even a simple SEM model may require estimates of 50 or 60 parameters since not only paths from indicator to latent variables and paths between latent variables are estimated. SEM also estimates, depending on the specification, error terms for indicators, correlations between indicators and between latent variables, and error terms for endogenous latent variables. Hence, an SEM that seeks to control for relevant latent variables with multiple indicators is very demanding indeed in terms of the number of participants that must be recruited. Moreover, the instrument used to collect the data is likely to be fairly lengthy since multiple indicators are sought, also contributing to difficulties in recruitment.

10.3.3 Barriers to Communication of SEM Results

Some of the difficulties in implementation of SEM analysis exacerbate the challenges of writing up results. As noted above, for a quantitative researcher, the very notion of a latent, unobserved variable is fairly abstruse, while qualitative researchers may be more accepting of the inability to measure a characteristic directly. One complication that this presents for research using SEM that aims to speak to educational policy is that there is no common metric in the model (e.g., effect size). Unlike a general linear model approach, significant effects at the structural level in SEM cannot be summed up by stating that an increase of 1 unit in (latent variable) X is associated with an increase of Y units in (latent variable) Z. This can be a problem in the realm of policy, where it is crucial to be able to argue that a mandated change will generate enough of the desirable effect to be a worthy investment.

Communicating results of SEM is also challenged by the fact that editorial standards vary concerning which details of the estimated model should be reported. While most estimates in a regression or ANOVA model can be thought of as substantively interesting even for control variables, not all estimates produced in SEM results are central to a substantive argument. Nevertheless, enough information needs to be given about both the measurement and structural model to assure the reader that it is sufficiently stable. As mentioned above, there is no single fit statistic that describes the degree to which the model approximates the observed data, adding to possible confusion. Efforts have been made to generate normative standards for reporting SEM results, such as Schumacker and Lomax (2004), which should ameliorate this problem in the future.

10.4 Synthesis: Working Toward the Gold Standard

In this chapter, we have argued that Gold Standard research on teaching and learning in science education should include a strong preference for a mixed-methods approach that combines both quantitative and qualitative components. In this, we support the subtle change in language first used by Tashakkori and Teddlie (1998), when they referred to studies that are strictly qualitative or strictly quantitative as “monomethod” (p. 17). For phenomena as complex as classroom teaching and learning, we believe that the norm should be a mixed approach and that a limited methodological approach should be labeled as such—monomethod—with an expectation that researchers justify their choice. To date and as a result of the paradigm wars, it has been incumbent on mixed-methods researchers to justify the combination of both quantitative and qualitative elements in the same study. We will know that we are enacting classroom research that approaches a Gold Standard when monomethodologists sound a little defensive when they communicate results, and when mixed methodologists (perhaps to be known as methodologists) report qualitative and quantitative results seamlessly and with no need for an extensive rationale.

For researchers aiming to influence in a significant way how teachers go about their work and how children learn, mixed methods hold genuine promise for leveraging change at all relevant levels of modern school systems. The current organizational cultures of school administrations and provincial or state governments are such that quantitative results are essential components for making arguments to change policies in schools. Accountability, as currently constructed, demands quantitatively based accounts. In contrast, but equally important in terms of implementing lasting reform, teachers are often more convinced by qualitative case studies, concrete examples, and thick description. Mixed-methods approaches allow the concerns of both groups to be addressed.

One might wonder why we have chosen to emphasize structural equation models as a desirable complement to a mixed-methods approach. Although technically demanding, we see SEM as having a kind of metaphorical resonance with qualitative approaches, making it a particularly good choice for mixed-methods studies. Multiple indicators of key concepts, differentially weighted, reflect a qualitative researcher's understanding that there can be varied manifestations, varied lived experiences of the core traits of good teaching and effective learning. Qualitative researchers know instinctively that there is no single way to demonstrate academic self-efficacy or science literacy. The use of multiple indicators in SEM acknowledges, potentially, a kind of humility about the act of measuring human attitudes and behavior that has been a central tenet of contemporary qualitative research. The capacity to model explicitly the many correlations among indicators and latent constructs in SEM mirrors the holism inherent in most qualitative research—an understanding that an individual is more than the sum of her measured attributes—but rather that these characteristics are apt to be interrelated in interesting and important ways. Finally, the flexible but delicate process of specifying models in SEM is less prescriptive in nature than other statistical approaches. Though theory-guided, it is often slightly speculative and, at its best, is rather artful in balancing demands of data, theory, and mathematics. This iterative character of SEM requires the analyst to rely on patterns to emerge from the data in ways that a qualitative researcher would find familiar.

Thus, we think the solution for making mixed methods part of the Gold Standard is not to simplify the quantitative part of the research but rather to choose a statistical approach (along with an appropriate data-collection strategy) that acknowledges and approximates the everyday complexity of classroom reality. Qualitatively oriented researchers are correct to reject statistical accounts that dramatically reduce this complexity simply to conform to mathematical assumptions in the model. Structural equation modeling does not impose such a reductionist view of teaching and learning. As such, it should be part of any mixed methodologist's repertoire, in the interests of reaching both quantitatively and qualitatively oriented audiences.

References

Arbuckle, J. L., & Wothke, W. (1999). *AMOS 4.0 user's guide*. Chicago: SmallWaters Corporation.

Arum, R., & Beattie, I. (2000). Introduction: The structure of schooling. In R. Arum & I. Beattie (Eds.), *The structure of schooling: Readings in the sociology of education* (pp. 1–11). Mountain View, CA: Mayfield Publishing.

Barnett, M., Lord, C., Strauss, E., Rosca, C., Langford, H., Chavez, D., et al. (2006). Using the urban environment to engage youths in urban ecology field studies. *Journal of Environmental Education*, *37*(2), 3–11.

Baumert, J., Evans, R. H., & Geiser, H. (1998). Technical problem solving among 10-year-old students as related to science achievement, out-of-school experience, domain-specific control beliefs, and attribution patterns. *Journal of Research in Science Teaching*, *35*(9), 987–1013.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Caracelli, V. J., & Greene, J. C. (1993). Data analysis strategies for mixed-method evaluation designs. *Educational Evaluation & Policy Analysis*, *15*(2), 195–207.

Clary, R. M., & Wandersee, J. H. (2007). A mixed methods analysis of the effects of an integrative geobiological study of petrified wood in introductory college geology classrooms. *Journal of Research in Science Teaching*, *44*(8), 1011–1035.

Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling: A Multidisciplinary Journal*, *2*, 119–143.

Gilmartin, S., Denson, N., Li, E., Bryant, A., & Aschbacher, P. (2007). Gender ratios in high school science departments: The effect of percent female faculty on multiple dimensions of students' science identities. *Journal of Research in Science Teaching*, *44*(7), 980–1009.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation & Policy Analysis*, *11*(3), 255–274.

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (pp. 105–117). Thousand Oaks, CA: Sage.

Howell, W. G., Wolf, P. J., Campbell, D. E., & Peterson, P. E. (2002). School vouchers and academic performance: Results from three randomized field trials. *Journal of Policy Analysis and Management*, *21*(2), 191–217.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, *33*(7), 14–26.

Lee, O., Luykx, A., Buxton, C., & Shaver, A. (2007). The challenge of altering elementary school teachers' beliefs and practices regarding linguistic and cultural diversity in science instruction. *Journal of Research in Science Teaching*, *44*(9), 1269–1291.

Mattern, N., & Schau, C. (2002). Gender differences in science attitude-achievement relationships over time among white middle-school students. *Journal of Research in Science Teaching*, *39*(4), 324–340.

Metz, M. H. (2000). Sociology and qualitative methodologies in educational research. *Harvard Educational Review*, *70*(1), 60–74.

Morgan, D. L. (1998). Practical strategies for combining qualitative and quantitative methods: Applications to health research. *Qualitative Health Research*, *8*(3), 362–376.

Nieswandt, M. (2007). Student affect and conceptual understanding in learning chemistry. *Journal of Research in Science Teaching*, *44*(7), 908–937.

Nieswandt, M. (2008, March-April). *Between theory and practice: Beginning high school science teachers' beliefs about science and science teaching over time*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Baltimore, MD.

Nieswandt, M., & Bellomo, K. (in press). Written extended-response questions as classroom assessment tools for meaningful conceptual understanding of evolutionary theory. *Journal of Research in Science Teaching*.

Nieswandt, M., & Shanahan, M.-C. (2008). “I just want the credit!” – Perceived instrumentality as the main characteristic of boys’ motivation in a grade 11 science course. *Research in Science Education*, *38*(1), 3–29.

Nieswandt, M., & Turner, J. E. (2008). *Implementation of inquiry-based instruction and the Motivation-Affect-Cognition Cycle (MACC): Effects on meaningful science understanding and further science study for minority high school students*. Unpublished proposal submitted to the National Science Foundation, Washington, DC.

Onwuegbuzie, A. J., & DaRos-Voseles, D. A. (2001). The role of cooperative learning in research methodology courses: A mixed-methods analysis. *Research in the Schools*, *8*(1), 61–75.

Onwuegbuzie, A. J., & Leech, N. L. (2004). Enhancing the interpretation of “significant” findings: The role of mixed methods research. *The Qualitative Report*, 9(4), 770–792. Retrieved from http://www.nova.edu/ssss/QR/QR9-4/onwuegbuzie.pdf

Sadler, T. D., & Donnelly, L. A. (2006). Socioscientific argumentation: The effects of content knowledge and morality. *International Journal of Science Education*, *28*(12), 1463–1488.

Sandelowski, M. (2003). Tables or tableaux? The challenged of writing and reading mixed methods studies. In A. Tashakkori & C. B. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 321–350). Thousand Oaks, CA: Sage.

Schofer, E., & McEneaney, E. H. (2003). Methodological tools and strategies for the study of globalization. In G. Drori, J. W. Meyer, F. O. Ramirez, & E. Schofer (Eds.), *Science in the modern world polity: Institutionalization and globalization* (pp. 32–54). Stanford, CA: Stanford University Press.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner’s guide to structural equation modeling* (2nd edn.). Mahwah, NJ: Lawrence Erlbaum.

Shaver, A., Cuevas, P., Lee, O., & Avalos, M. (2007). Teachers’ perceptions of policy influences on science instruction with culturally and linguistically diverse elementary students. *Journal of Research in Science Teaching*, *44*(5), 725–746.

Smith, J. K., & Heshusius, L. (1986). Closing down the conversation: The end of the quantitative-qualitative debate among educational inquirers. *Educational Researcher*, *15*(1), 4–12.

Snow, C. P. (1959/1998). *The two cultures*. (Introduction by S. Collini). Cambridge, UK: Cambridge University Press. (Original work published 1959)

Talanquer, V. (2008). Students’ predictions about the sensory properties of chemical compounds: Additive versus emergent frameworks. *Science Education*, *92*(1), 96–114.

Tashakkori, A., & Teddlie, C. B. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

Teddlie, C. B., & Tashakkori, A. (2003). Major issues and controversies in the use of mixed methods in the social and behavioral sciences. In C. B. Teddlie & A. Tashakkori (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 3–50). Thousand Oaks, CA: Sage.

Tretter, T. R., Jones, M. G., Andre, T., Negishi, A., & Minogue, J. (2006). Conceptual boundaries and distances: Students’ and experts’ concepts of the scale of scientific phenomena. *Journal of Research in Science Teaching*, *43*(3), 282–319.

Trumper, R. (2006). Teaching future teachers basic astronomy concepts – seasonal changes – at a time of reform in science education. *Journal of Research in Science Teaching*, *43*(9), 879–906.

Turner, J. C., & Meyer, D. K. (2000). Studying and understanding the instructional contexts of classrooms: Using our past to forge our future. *Educational Psychologist*, *35*(2), 69–85.

Winberg, T. M., & Berg, C. A. R. (2007). Students’ cognitive focus during a chemistry laboratory exercise: Effects of a computer-simulated prelab. *Journal of Research in Science Teaching*, *44*(8), 1108–1133.

Wu, H.-K., & Huang, Y.-L. (2007). Ninth-grade student engagement in teacher-centered and student-centered technology-enhanced learning environments. *Science Education*, *91*(5), 727–749.

Chapter 11
Mixed-methodology Research in Science Education: Opportunities and Challenges in Exploring and Enhancing Thinking Dispositions

Tamar Levin and Tili Wagner

Questions about the value of research in education, its paradigmatic orientation, and potential use and importance for advancing knowledge have resurfaced in the past decade, mainly as a result of the proliferation of standards-based reforms and high-stakes accountability policies. The political agenda of accountability, manifested in such issues as the call for common standards, quality indicators, and evidence-based instructional programs, has created a demand for *proven* research strategies among educators, including literacy and science educators. Two acts in the United States—the No Child Left Behind Act of 2001 (NCLB, 2002) and the Education Sciences Reform Act of 2002 (ESRA, 2002)—have prompted epistemological questions regarding the disciplinary profile of educational research, maintaining that education research can and should shape knowledge, policy, and practice. Their demand for research-proven strategies along with hopes of making education an evidence-based field (US National Research Council, 2004) have helped to challenge and encourage fresh and mindful dialogue on the nature of worthwhile research in literacy and science education and the relation of research to both theory and practice.

For many decades, the quantitative paradigm has been the paradigm of choice for science education research—chiefly when investigating the relation between different types of instruction and student learning. More specifically, the dominant paradigm or Gold Standard for science studies and program evaluation has been experimental methodology. This standard is considered to have merit in particular because experimental research makes explanations possible as to cause and effect. This paradigm has often involved comparing instructional innovation with more traditional forms of instruction in an experiment building or drawing on theory-driven hypotheses, random assignment of participants to treatments, controlled manipulations uniformly applied to all participants under rigorously controlled conditions, and the use of quantitative measurement and statistical analysis (McCall & Green, 2004). The parameters of evaluating performance outcomes have also remained well within

T. Levin
Tel Aviv University

T. Wagner
Beit Berl College

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

the empirical–analytic paradigm. This has involved assessing student responses and conceptions as right or wrong, giving little interpretation or consideration to the context. Scores have usually been standardized against such norms as statistical distributions or judgment by a panel of experts (Aikenhead, 1997).

Although numerous other methodologies are often used in educational research and although educators emphasize the need to address questions relating to cause and effect (internal validity), ecological validity, and practical importance, researchers still show a persistent preference and loyalty toward the traditional methodological approach (McCall & Groark, 2000). Studies implementing qualitative methodologies with rich sets of measurement instruments, such as classroom observations and interviews, appear more frequently over the last 20 years, augmenting science education research and mainly complementing quantitative data (Erickson, 1986). The advantages and disadvantages of both quantitative and qualitative research in science education have been discussed (Libarkin & Kurdziel, 2002) although Patton (2002) believes this is only the infancy of methodological enlightenment and tolerance for mixed methodologies.

The critique of previously accepted methods of studying educational phenomena in the academic fields, including science education, and debates between proponents of differing positions have been so extensive that some authors call the present period the era of “paradigm wars” (Hammersley, 1992). Given a new context with questions of disciplinary identity at stake and the fact that realist, constructivist, multiplist, evaluativist, and critical epistemologies all receive equal respect, the question is whether just one theory or methodology can or should describe, predict, explain, or change all the phenomena in a discipline dealing with how diverse human beings achieve science literacy. This chapter and its central study aim to demonstrate that—when broadening our methodological value system to recognize the advantages and limitations of each methodology used—there are valuable benefits to considering a plurality of methodologies and paradigms in science education research, which may also maximize our ability to understanding science learning.

11.1 Theoretical Background

The present study suggests methodological ideas and highlights issues related to mixed-methodology research in the context of studying the relationships between writing and thinking in science learning. Conceptually, it was inspired and informed by constructivist views of learning approaches to science literacy (Hand, Lawrence, & Yore, 1999; Yore, 2000), cognitive and social theories of the writing process (Sperling, 1995), and theories of thinking dispositions (Tishman, Perkins, & Jay, 1995). It focuses on the enhancement of student scientific literacy—the abilities, values, and emotional dispositions for understanding fundamental concepts, major scientific ideas, and scientific processes—and the communication skills to clearly articulate their written and oral understanding of the material, describe what they have learned to others, and persuade others to take informed action (Yore; Yore, Pimm, & Tuan, 2007).

Recognizing that ability alone does not qualify as intelligent behavior and that all too often people are able to think more effectively than they do but are not disposed to do so (Tishman et al., 1995), this study focuses on exploring how students' writing-in-science experiences affect their thinking dispositions. The study regards writing as a constituent of science literacy (Norris & Phillips, 2003), a dynamic mode of knowledge transformation, and a meaning-making process involving the constant reevaluation and transformation of a writer's knowledge and understanding about the content space and the discourse space (Keys, 1999). Drawing on the ideas that good writers must cross the boundaries of conventions, discourse, and communities rather than adopt a narrow template for their writing (Chaopricha, 1997) and that production of multiple, expressive genres is likely to make science more appealing to an assortment of students and groups in society (Prain & Hand, 1996), this mixed-methodology study explores whether, to what extent, and under what conditions writing in science using different writing genres has the potential to enhance students' thinking dispositions.

11.1.1 The Qualitative and Quantitative Continuum

Researchers have long debated the relative value of qualitative and quantitative inquiry (Patton, 1990; Phillips, 2005). What distinguishes qualitative and quantitative research is not simply their methodology but also various other differences relating to ontology, epistemology, and philosophy of science (Eberle, 2005). *Qualitative inquiry* is usually associated with an interpretive paradigm, such as phenomenology, social hermeneutics, and constructivism. It manifests a transition from objective to constructed multiple realities and is based on viewing reality as complex, non-deterministic, perspectival, and contextual (Bresler & Stake, 1992). *Quantitative inquiry*, in contrast, seeks causal determination, prediction, and generalization of research findings. Because the underlying principles of each research method reflect very different assumptions, the research procedures are also very different. Qualitative methods are often considered soft, flexible, subjective, political, case-specific, speculative, and grounded. Quantitative methods are often described as hard, fixed, objective, value-free, survey oriented, hypothesis testing, and abstract (Halfpenny, 1979).

However, as with all dichotomous classifications, when such distinctions are made within a complex field, they obscure the inherent diversity within each paradigm and method (Phillips, 2005). For example, it is becoming increasingly recognized that all data collection, both quantitative and qualitative, occurs within a particular cultural context and is affected to some extent by the perceptions and beliefs of the investigators and data collectors. The fact that in reality the methodological applications of the two paradigms, which are not as distinct as theory would propose, challenge the existence of pure paradigmatic research methodologies and methods and indicate a reality that is more integrative than polarized. Indeed, Borland (2001) argued that the relationship between qualitative and quantitative research should not be considered as mutually exclusive or dichotomous but as a continuum of complementary paradigms that when used in concert produce complete or useful knowledge.

Claiming that there are many ways to represent our understanding of the world, Eisner (1991) placed the two research approaches on a continuum from fictional to highly controlled, reflecting the conceptualization that a continuum is more appropriate for research paradigms than a simplistic dichotomy. Firestone (1987) identified two groups in the qualitative and quantitative debate: the *purists* and the *pragmatists* (see also Tashakkori & Teddlie, 1998). The purists believe that the two methodologies are incompatible because they are inextricably linked to paradigms that make different assumptions about the world and what constitutes valid research. They argue that there is a logical relationship between these paradigms' principles and the research methods chosen so that epistemology informs method. The pragmatists do not agree, seeing methods as a collection of techniques with no inherent link to paradigmatic assumptions and method types can be associated with both paradigms. Patton (1990, 2002), who took a pragmatist position, argued that there is no need to see the two paradigms as rivals, and strongly advocated for a "paradigm of choices" (1990, p. 39), which seeks methodological appropriateness as the primary criterion for selecting, applying, and judging methodological quality. More specifically, he posited the importance of matching the research method and paradigm to the problem, question, purpose, and issues addressed by the research. He suggested it is important to use aspects of quantitative and qualitative methods that are responsive to the nuances of particular empirical questions and the idiosyncrasies of specific stakeholder needs in evaluation studies.

11.1.2 Methodological Dualism

Other researchers believe that qualitative and quantitative methodologies can be combined effectively in a single research project. Flick (2002) argued that methodological dualism is possible and that different research perspectives can be combined and supplemented. However, this conflicts directly with Guba and Lincoln's (1988) view of methodological choices and the analogies they provided in support of their arguments, "Like water and oil, they do not mix; like similar magnetic poles, they repel one another; to hold them in contact requires force, and when the force is released, the methodologies fly apart" (p. 111). Budd (2001) suggested this opposition to mixed-method research relates more to the inability of individual disciplines to talk to one another in their different language and terminology rather than the essential characteristics of each approach.

Rocco and colleagues (2003) suggested that research is generally concerned with *best-use* techniques and procedures for answering specific problems but with no a priori commitment by the researcher to using a mixed method. According to this view, mixed methods may be used when the researcher believes it would enhance the accuracy of the data collection and analysis procedures and the usefulness of the resulting inferences. In contrast, the dialectical position (Maxwell & Loomis, 2003) explicitly requires synergistic benefits from integrating of postpositivist and constructivist paradigms. Greene and Caracelli (1997) posited that the underlying belief of the dialectical position is that it is more ethical to mix methods and thereby describe a plurality of interests, voices, and perspectives rather than to limit the study to just one point of view.

Sale, Lohfeld, and Brazil (2002) also suggested that combining research methods is not only legitimate but also useful for several reasons, including (a) the complexity of phenomena investigated requires data representing a large number of perspectives; and (b) the two paradigms share a commitment to understanding and improving the human condition, possess the common goal of disseminating knowledge for practical use, and hold a shared allegiance to rigor, conscientiousness, and critique in the research process (Reichardt & Rallis, 1994). In other words, the two approaches share the tenets of theory—leadenness of facts, fallibility of knowledge, indetermination of theory by fact—and a value-laden inquiry process. King, Keohane, and Verba (1994) added that qualitative and quantitative studies share a unified logic and the same rules of inference apply to both. However, Sale and colleagues noted that, despite the arguments presented for integrating methods, it is important to realize that each method is based on a particular paradigm, a patterned set of assumptions concerning reality (ontology), knowledge of that reality (epistemology), and the particular ways of knowing that reality (methodology) (Guba, 1990). Therefore, one should only combine multiple methods in a single study if they are complementary, study different phenomena, and play the role of an additive partner. Ercikan and Roth (2006) suggested an integrative framework with low levels of inference at one end and high levels of inference at the other. Thus, the formerly distinct forms of quantitative and qualitative research are now located at different locations of the same scale; and research at different positions on this continuum addresses different questions.

While the debate still rages between proponents of the two research traditions, certain theoretical contributions have strongly influenced the direction of recent methodological developments. Researchers (Howe, 1988; Yin, 2003) and evaluators (Cook, 1995; Patton, 1990; Visser, n.d.) have increasingly claimed that a single investigation can use quantitative *and* qualitative methods, acknowledging that—although both approaches reflect different ontological, epistemological, and methodological assumptions and although both have limits and strengths—it is possible and useful to combine aspects of both. However, they also argue that such integration requires careful consideration and justification regarding the specific research context, problem space, and questions; and it should be based on formulated theoretical assumptions, coherence, and critical reflection (Yanchar & Williams, 2006). This suggests the emergence of a hybrid approach that is characterized by mindful flexibility and allows the researcher to combine both qualitative and quantitative research techniques. Libarkin and Kurdzile (2002) claimed that, although some researchers choose one research paradigm over another, the mixed-method design of statistical analysis with contextual data has been used with great success by a number of researchers. They suggested that qualitative analysis provides the context lacking in quantitative research and that quantitative analysis widens the implications of a purely qualitative study. Most importantly, they claim that such dual studies can guide educational practice for both the local setting under study and the wider context.

11.1.3 Mixed-methodology Framework

Many now view qualitative and quantitative methods as complementary, choosing the most appropriate method for their inquiry (Mackenzie & Knipe, 2006). Moreover, it is argued almost invariably that both approaches need to be applied for the research to be fully effective. Despite the various disagreements in view, some of which are mentioned above, there seems to be a consensual understanding that the research method applied in different stages of the inquiry should be determined by development of the problem space (see Yore & Boscolo, Chap. 2), the research question (Creswell, 2003), and the complexity of the study (Shulha & Wilson, 2003). Collaborative, mixed-method research involves the purposeful application of a multiperson, multiperspective approach to research and evaluation issues. A nonpurist, or mixed, position allows researchers to mix and match design components that offer the best chance of answering their specific research questions (Johnson & Onwuegbuzie, 2004; Onwuegbuzie & Leech, 2004). This agrees with Classen and Lopez (2006) who stressed that, rather than simply collecting qualitative or quantitative data, mixed-methodology research—a third paradigm—calls for data to be integrated, compared, contrasted, appraised, and synthesized and that, when used in combination, this form of research yields a more complete analysis.

This movement between two opposite worldviews of separateness and interconnectedness resembles the transition from Newtonian physics to quantum and relativistic physics. The analogy from quantum physics is helpful for combating our tendency to dichotomize and demonstrates that the either/or positions of classical Cartesian reasoning are not necessarily valid—because in order to understand the nature of light, one must incorporate the findings that light is both wave-like and particle-like; otherwise, the account is incomplete. Thus, instead of insisting on understanding light in terms of either waves or particles, Bohr's complementarity principle recognized that light is neither a wave nor a particle but *both* wave *and* particle (wave-particle). It seems, therefore, that contradictory ontological, epistemological, and methodological assumptions and explanations can indeed coexist although they call for a broader worldview, which encompasses separateness while transcending its limitations (Gilman, 1993).

Therefore, the combination of quantitative and qualitative research, which many methodologists have regarded as incommensurable opposites, not only seems feasible and beneficial for solving educational puzzles but can also unravel problems that pure designs cannot solve (Niglas, 2004). By combining quantitative and qualitative research, we can describe and understand educational phenomena using their own constructs to give a more contextually situated understanding or *thicker* description of the phenomenon as we try to simultaneously describe and explain phenomena in terms of external standards in terms of quantifiable and generalized dimensions. The combined quantitative–qualitative approach perceives the combination of paradigms as complementary ways of studying educational phenomena rather than mutually exclusive states; it asks us to live with the paradoxes in the ontological and epistemological spheres while trying to find a solution within them. We can, therefore, view mixed-methods research as "a separate methodological orientation with its own worldview, vocabulary, and techniques [or] the third methodological movement" (Tashakkori & Teddlie, 2003, p. 679).

What we have here is an acceptance of the dialectical position (Greene & Caracelli, 1997), which calls explicitly for a synergistic benefit from integrating the positivist and the constructivist paradigms. The underlying assumption is that research will accomplish more if it mixes research paradigms since this offers a fuller understanding of human phenomena and because it is more mindful and ethical to mix methods "in order to represent a plurality of interests, voices, and perspectives" (p. 14).

This view is reflected in Johnson and Onwuegbuzie's (2004) definition of mixed-methods research, which is also the point of departure for the present study. According to these scholars, mixed-methods research is "the class of research where the researcher mixes or combines quantitative and qualitative research techniques, methods, approaches, concepts, or language into a single study" (p. 17). In other words, mixed-methods research is not simply about mixing methods and techniques but about mixing concepts and approaches as well. For example: a combination of experimentation and survey, which are both forms of *quantitative* methods, could better answer the dual needs of addressing internal and external validity than either method alone (Berends & Garet, 2002). Thus, according to Yin (2006), the use of experimentation and survey in a single study is an example of mixed-methods research even though no *qualitative* method is included. Researchers have similarly recognized that we can use different qualitative methods (Guba & Lincoln, 1994) in a single, mixed-methods study without using quantitative methods.

So conceptually freed from the quantitative–qualitative dichotomy, an understanding of the relevance and reality of using a wide variety of *mixes* has emerged, which recognizes the true diversity between and within the research methods with their own individual strengths and weaknesses that we use in education. It is, therefore, more appropriate to term such research as *mixed methodology* rather than simply mixed methods; we view methods as concrete procedures toward a particular end, whereas methodology is a higher-level construct providing a rationale for choosing between different methods. In other words, mixed-methodology research encompasses the general methodological framework, the local methodology, and the methods or techniques.

Though we still need to resolve both theoretical and practical issues when discussing the mixed-methodology approach to research (Greene, 2008), it is important to note that, particularly in light of the quantitative–qualitative debate, methodologists have been suggesting divergent lists of general standards and criteria for research be it quantitative, qualitative, or mixed methodology. For example, Eisenhart and Howe (1992) argued that “all educational research is subject to the same general criteria of validity even though quite distinct and specialized criteria are required to conduct and evaluate specific kinds of research studies” (p. 644). They suggested five general standards that should be applied to all forms of educational research, including mixed-methods studies. These general standards require that research studies be (a) cogently developed, namely there should be a fit between research questions, methodological tools, and inferences from data; (b) competently produced, meaning that data collection and analysis techniques should be competently and effectively applied; (c) coherent with respect to previous work; (d) important and ethical; and (e) comprehensive, meaning that there should be a balance between the technical and theoretical quality, the scientific and practical value and importance of the study, the risks involved, and an alertness to knowledge from outside the tradition in which the author is working. Though these general standards of validity form a unitary holistic construct, according to Niglas (2004), design-specific standards for new emerging research strategies must still be developed.

When, therefore, should we carry out mixed-methodology studies? Relying mainly on examples from research practice, different authors have listed various reasons for combining methodologies, which address the quantitative and qualitative facets in a study. Greene, Caracelli, and Graham (1989), for example, identified five reasons for using mixed-methods design strategies: triangulation, complementarity, development, initiation, and expansion (see Nieswandt & McEneaney, Chap. 10). Triangulation, seeks to improve the accuracy of the results by collecting and analyzing different types of data. It refers to the convergence or corroboration of the data and interpretation of a phenomenon. Complementarity, also termed elaboration, goes beyond triangulation by focusing not only on overlapping or converging data but also on the different features of phenomenon, thereby providing a greater insight and perspective. The goal of complementarity is thus to use “the strengths of one method to enhance the performance of the other method” (Morgan, 1998, p. 365). Development, which also increases validity, combines or uses the findings from one method of studying a phenomenon to develop another. This requires a sequential research design since the results of the secondary study inform and shape the primary, dominant study. Initiation involves deliberately analyzing inconsistent qualitative and quantitative findings to add depth and breadth to results and interpretations. This search for “fresh insights” (Greene et al., p. 260) is more likely to emerge spontaneously than be a planned part of the research design. It, therefore, involves identifying paradoxes and contradictions following the intentional analysis of new perspectives for a phenomenon. Finally, expansion concerns the overall widening of the scope, breadth, or range of a study. Mixed methodologies can be used to expand the scope and breadth of a problem by studying multiple phenomena through a desire to provide a more comprehensive solution or understanding of a problem.

The five reasons for conducting mixed-methods research are not mutually exclusive however. In fact, in their review of 57 studies, Greene and colleagues (1989) reported that authors often listed multiple reasons for using mixed methods, giving a total of 70 reasons for mixed-method research. Petter and Gallivan (2004) expanded this by suggesting an additional dimension to the five-pronged scheme outlined above, which relates to the implementation phase of mixed-methods research. They suggested three designs for mixed-methods data collection: sequential, parallel, and independent. Sequential design has two or more distinct phases and uses different research methods at each stage. In sequential design, it is common for one study or methodology to predominate. The second design is parallel design, in which two or more simultaneous studies are conducted by separate researchers using different methods but allowing for some interaction between the researchers, data, and results of each study. The third design is the independent research design, which is typically conducted at the same time but with no interaction between researchers during data collection and analysis.

Petter and Gallivan's (2004) two-dimensional framework helps researchers to choose a mixed-methodology research design based on the goals of their study. However, Greene (2008) suggested that mixing methods opens up possibilities while the type of methods used depends on why mixed-methodology research was chosen in the first place. Therefore, we still need to develop different sets of methodological principles and guidelines for this research. Having said this, the focus on a *single* study seems critical to mixed-methodology research mainly due to the need to provide converging evidence, which is presumed more compelling than the evidence any single method alone might produce (Yin, 2006). According to Yin, if a research effort consists of multiple, related studies rather than a single study, the use of multiple methods offers little distinctive contribution. Furthermore, the more a single study integrates mixed methods in the study's five basic procedures (research questions, units of analysis, sample, instrumentation, and data collection and analytic strategies), the more valuable and practical the contribution of mixed-methods research will be (Yin).

The following methodological assumptions were, therefore, formulated based on the above review and the idea that, when moving to a larger worldview, we can retain useful elements of the old:

1. Methodology links ontological and epistemological assumptions and the practical and technical issues involved in methods.
2. The methodologies chosen for any study—whether quantitative or qualitative or both—are largely determined by the questions asked and the development of the problem space under investigation and not by a predetermined commitment to a specific paradigm or methodology.
3. Both qualitative and quantitative paradigms have strengths and weaknesses for exploring complex phenomena; therefore, it would be wise to integrate their complementary elements in what Shulman (2004) called a "union of insufficiency [or a] marriage of complements" (p. 355).
4. Research should be informed by theoretical sense of the topic and its methodologies; therefore, methods and interpretation should fit together within a larger theoretical framework or purpose.

11.1.4 Assessing the Problem Space and Clarifying Research Questions

Further review of the science literacy, writing-to-learn science, critical thinking, and science learning literatures revealed that these constructs have varying degrees of development and the combined constructs lack sufficient canonical knowledge to venture hypotheses about tentative relationships and causal mechanisms. The science literacy literature proposes a cognitive symbiosis or interaction between writing, thinking, and understanding (Hand, Prain, & Yore, 2001; Norris & Phillips, 2003; Yore et al., 2007), which has not been fully tested and established with research evidence. Keys (1999), for example, illustrated how writing as the transformation of knowledge or construction of understanding involves the movement between the writer's prior conceptual knowledge about the nature of science and the target topics and knowledge about science discourse conventions, practices, and traditions. This problem-solving process involves moving to the other space in attempts to solve problems that appear anchored in the original space. However, her speculations are not fully explored and supported with research.

We think that mixed-methodology research for the two constructs of writing-in-science and thinking is likely to be valuable for two reasons. The first is that it will help us describe and elaborate on the complementary relationships between the two constructs: writing-in-science and thinking. Specifically, it will provide richness, additional detail, and a better understanding of whether, how, under what circumstances, and for what dimensions scientific habits of minds can be developed through writing experiences in science classrooms. Although the literature shows that writing can enhance critical thinking in students, few studies have explored the actual processes for developing these skills. For example, most studies compare students' thinking skills before and after writing, mostly concluding that both formal and informal writing tasks/genres enhance higher-order thinking because writing invokes cognitive strategies for processing, encoding, and expressing information and requires students to reflect, consolidate, elaborate, hypothesize, interpret, synthesize, and persuade. However, not many studies offer evidence demonstrating the manifestations of these processes.

Furthermore, since each method of measurement—in this case, a self-perception questionnaire and discourse analysis—sets its own context, we are not likely to expect complete correspondence between the two sets of measurements. For example, it is not very feasible to examine sensitivity to context in a self-perception questionnaire of disposition, whereas this can easily be observed in the student's actual writing. Similarly, it is almost impossible to determine students' enjoyment of a defined thinking disposition from their actual writing, whereas one can estimate their degree of enjoyment with a Likert-scale self-perception questionnaire.

The second reason why mixed-methods research is helpful for examining writing-in-science and thinking is that it can expand the scope and breadth of the problem-space by studying multiple phenomena as well as provide a more profound and comprehensive understanding of the problem at hand. Whereas most studies examine students' thinking abilities, few focus on their thinking dispositions that reflect a person's inclination to act a certain way and their sensitivity to do so when it is appropriate (Ennis, 1996). Since thinking dispositions develop through enculturation and are best enhanced in an environment that reinforces good thinking in a variety of ways (Tishman et al., 1995), it was considered important to explore whether, how, and to what extent students' thinking dispositions are explicitly reflected in their actual writings during their writing experiences. This contrasts with limiting our investigations to simply measuring students' self-stated perceptions of their thinking dispositions prior to and even following their writing-in-science experiences. Mixed-methods research also encourages us to examine whether and how thinking dispositions develop as a function of their cumulative experience in writing-in-science.

Similarly, assuming that different writing tasks encourage students to invoke different cognitive strategies (Langer & Applebee, 1987) and that written genres are tools that students use to learn about the rhetorical contexts, or the rhetorical representations of the discipline, it seems meaningful to explore not only whether students' thinking dispositions develop as a consequence of their writing experiences in science but also whether similar dispositions are manifested in different genres of writing. Likewise, it is consequential to explore which indicators of thinking dispositions are observable in student writings on different genres.

Specifically, the study addressed two questions that are very good candidates for a mixed-methodology design to use quantitative measures of well-established constructs paired with qualitative information about a less well-developed idea:

1. How can we characterize mixed-method research, and how can it enhance the quality of science education investigation?
2. What can a mixed method of study teach us about the development of students' thinking dispositions in the context of their writing-to-learn experiences in the science classroom?

11.2 Methodology

The study, which lasted three school semesters, was based on the ideas that theory alone has little power to create change in schools and that there is a need for a more complex interplay between theory and practice. We, therefore, chose the scientific approach to action research (Glanz, 1998) as the research model, which allowed us to provide schools with useful, practical knowledge while eliciting new forms of understanding through reflection on action. Although the scientific action research model requires comparative and intervention groups, its participatory paradigm differs from an experimental or quasi-experimental design in terms of its flexibility and freedom to accommodate research conditions to the evolving needs or concerns of the intervention group. The design, therefore, included intervention and comparative groups, which resembled in terms of the students' personal and social characteristics, science curriculum, and science teachers.

11.2.1 Participants and Context

The study was conducted in two junior secondary schools in a city in central Israel whose students were from middle-class backgrounds. The sample comprised 97 Grade 8 students (42 boys and 55 girls) divided into two intervention classes ($n = 48$) and two comparative classes ($n = 49$). Two teachers participated in the study; each taught one intervention group and one comparative group. Thus, although no random assignment of students to groups was feasible, the assignment of two teachers to both the intervention and the comparison groups and the use of the same curriculum controlled two potential influences that were likely to affect outcomes. The intervention and comparative groups studied five science topics in six learning units in the following order: heat and temperature, an introductory unit on reproduction, fibers, electrical circuits, an advanced unit on reproduction, and energy.

Before and at the end of the action research, a questionnaire was used to measure students' self-perceived thinking dispositions in both groups. During the study, students in the intervention group received writing assignments at the end of each learning unit and a reflection task on the subject of their writing. Throughout the study, after each unit's instruction, these students received the same writing tasks genres (debate, story, diary)—with one exception: the fourth unit (electrical circuits), when the teachers asked that the students be assigned only one genre (debate). This approach of allowing the intervention group students to choose their own task reflects the constructivist approach to science learning. Free choice increases the likelihood of students finding a task that speaks to them personally. It also reflects the belief that choice is positively related to the student's intrinsic motivation to carry out the task (Cordova & Lepper, 1996).

Students in the comparative groups were not asked to write anything during unit instruction. They only wrote at the end of the study for comparative purposes. The writing genre used after the last writing task of the intervention group, which was also the only writing task given to the comparative group, differed from the genres used in the previous units. These genres were a plan and two related letters (putting forward two different viewpoints through a series of letters) referring to a specific, proposed program. The type of writing genre used in the last unit was changed so as to enable (a) a valid comparison of the writing quality of the two groups following the intervention, thereby eliminating the possibility of familiarity with a particular genre influencing the intervention group's writing, and (b) an exploration of the cumulative effects of the students' writing experiences by assessing the transfer effect within the intervention group.

11.2.2 The Intervention: Use and Description of Informal Writing Tasks

Informal writing tasks are items of writing not regularly used in the science class as part of the science discourse (Keys, 1999). These tasks can take the form of a genre that is uncharacteristic of typical science discourse. Three informal writing genres were designed and assigned to five of the units: a story, a debate, and a diary. Naturally, the task subject was related to the unit's content. Altogether 13 tasks were developed for the first five units. In addition, two different tasks, a plan, and two related letters were used for the final unit for both the intervention and comparative groups.

The debate genre belongs to the general genre of argumentation (Kuhn, 1991) and involves a socially constrained dialogue in which the arguer attempts to establish a position against actual or potential divergence from the audience. The diary and story genres are two forms of narrative requiring writers to assign meaning to events and establish connections between them (Nicolopoulou, 1997). Both the diary and story create opportunities to describe and explain phenomena from an individual–idiosyncratic viewpoint. Thus, one can expect these genres to elicit new conceptual distinctions and generalizations about different phenomena and situations (Peled, 1997). Diary writing involves systematic documentation of activities imagined by the writer and encourages the writer to create relationships or a temporal sequence between events, goals, hopes, and emotions. The plan involved global energy distribution; it required students to suggest several criteria for energy distribution and then consider the relationships between them in order to generalize and achieve a holistic view of the benefits and drawbacks. Similarly, the two related letters addressed to a committee dealing with global energy distribution provided a potential platform for designated dialogue with a defined, prospective reader. This task required students to present justifications and persuasive argumentation regarding the difference of opinion expressed in the two letters. It also required students to justify claims, persuade by using assumptions and principles, and establish judgment criteria.

The selection of these writing genres was based on a theoretical analysis of their potential to enhance thinking dispositions. Although the genres differed in terms of the strategies they encourage students to use and in their rationale and purpose, they are all open-ended, exploratory, elicit personal voice, and—most importantly—share the potential to enhance students' thinking dispositions. Naturally, they all require an elaborated, coherent text that explicitly connects ideas. Given these characteristics, we assumed and expected that each genre would contribute to the development of student thinking and serve as a motivational means to deepen interest and enhance enjoyment. The following writing task examples are from the heat and temperature unit as part of the intervention and the energy unit as the comparison across the two groups. Section 11.3 also describes additional writing tasks.

Example of Writing Tasks—Heat and Temperature Unit

Story: *Tell the story of a group of water particles heated from a temperature of 20°C below zero to 150°C above zero.*

Diary: *An additional sun has appeared over our planet. The sun radiates continuous heat on the Earth. Write a diary of your own or the diary of someone else describing the effect on our world.*

Debate: *Two children, Dan and David are arguing:*

Dan: In my computer game, the hero shoots a tank using a small rifle and the tank evaporates.

David: That's impossible. The tank is made of iron. Iron is a metal and there is no heat that can cause the tank to evaporate.

Continue the argument.

Example of Writing Tasks—Energy Unit

The year is 2030. The Third World War, which was caused by international disputes over energy resources, is over. The United Nations has decided to distribute energy resources differently. According to the resolution, no matter where energy resources are found they belong to all nations. An international committee is being formed to discuss the distribution of these resources.

Plan: *As a member of the committee, write a proposal for the new energy distribution.*

Related letters: *Write two letters to the committee, one from a boy in Saudi Arabia and the other from someone in Israel.*

11.2.3 Discourse Analysis of the Writing Tasks

In order to describe and explain the processes and outcomes of writing in science, we looked for thematic categories in the students' writings. The study employed the phenomenographic (Marton, 1986) approach to data analysis, which grouped the expressions students used in their writings according to similarities, differences, and complementaries—what Glaser and Strauss (1967) called the "constant comparison method" (p. 116). Thus, the data were constantly reorganized and reinterpreted according to the categories emerging from the raw data. This process was repeated for every unit separately. The established categories were compared and refined to reveal important sources of similarity and differences until finally several qualitative dimensions were obtained, which reflected different dimensions and thinking disposition indicators.

The categories we established were partly built on the data and partly theoretical; for example, when relying on Langer's (1993) definition of the mindfulness construct, we used the following three components of mindfulness: (a) alertness to new distinctions, (b) sensitivity to different contexts, and (c) awareness of multiple perspectives. A priori categories arise from the characteristics of the phenomenon studied; from accepted professional definitions found in literature reviews; from local, commonsense constructs; and from researchers' values, theoretical orientations, and personal experiences. Strauss and Corbin (1990) called this aspect of discourse analysis theoretical sensitivity. Mostly, however, the categories were deduced from the empirical data, texts, and textual images. In other words, knowing that we could not possibly anticipate every theme the students would write before we analyzed the data, we also looked for other themes/categories that were more subtle, symbolic, and even idiosyncratic. It is worth mentioning that by using both the theoretical knowledge and the empirical data we were able to examine the data for what was not mentioned in the literature. In other words, we tried to identify issues that students had perhaps intentionally or unintentionally avoided. We believed that this type of design steers a middle path between applying prior theorizing and grounded theory to the theme-identification aspect of the analysis and examining the writings with both a well-prepared mind and a fresh eye.

There was 90% agreement on the interpretation of the data and the categories obtained between the three evaluators: two science teachers, one specializing in teaching science and technology, one who had studied the role of writing for learning in science for several years; and an educational expert in interdisciplinary curricula, mainly science and mathematics. Consensus was established after discussing minor differences of opinion.

A thinking disposition questionnaire was developed especially for the study. The questionnaire comprised 28 items in the form of a 5-point Likert scale ranging from 1 (totally disagree) to 5 (totally agree). Construct validity of the questionnaire was established with principal component factor analyses performed on two different samples. The identical results for the two groups provided five factors and explained a high proportion (78%) of the total variance. The five factors were: tendency to view a phenomenon from different angles, inclination to draw new distinctions, need and willingness to think, appreciation of metacognition, and enjoyment of thinking. The reliability indices for the factors ranged from 0.88 to 0.94. The reliability indices for the total questionnaire were 0.85 prior to the study and 0.88 at its completion. This questionnaire was administered to all groups on a prestudy-poststudy schedule.

11.2.4 Analysis

Using two sets of measures (indicators of the thinking dispositions identified in the student writings and students' self-perceived measures of their own thinking dispositions), we explored the effects of the science writing experiences on the development of student thinking dispositions in several complementary ways. First, we qualitatively examined the content and discourse indicators of the thinking dispositions revealed in the students' writings (intervention group). Second, we quantified these indicators and carried out an intragroup comparison for the students' first and last writings of the thinking disposition measures. Third, we compared the discourse analysis measures of the students' writing at the end of the study for both the intervention and comparison groups. Fourth, we assessed the change in students' self-perceived thinking dispositions between the beginning and the end of the study. Finally, we compared the self-perceived thinking disposition measures of the intervention and the comparison groups—both intact—before and at the end of the study.

To determine how the content and the writing genres influenced students' thinking dispositions, we performed a descriptive discourse analysis and examined the thinking disposition indicators found in the different writing genres for a random selection of science learning units. We expected to see that the categories derived from analyzing every thinking disposition would not be limited to a particular writing genre or content unit. We also performed a descriptive analysis of the three thinking dispositions found in the students' writings in each genre, expecting to find clear differences between the indicators for each thinking disposition examined and each genre.

Methodologically, it is important to bear in mind that since each method of measurement (self-assessed questionnaire versus discourse analysis) sets its own context (limits and possibilities for measuring the variables) we anticipated that it would be impossible to obtain a complete correspondence between the two sets of measurements. For example, it is far more challenging to examine *sensitivity to context* in a self-assessment questionnaire of disposition and far easier to observe it in the student's actual writing. Similarly, it is almost impossible to establish a student's enjoyment of a given thinking disposition from their actual writing and a straightforward matter to establish enjoyment using a self-assessment questionnaire. We believed that the use of quantitative and qualitative measures to examine student thinking dispositions both through their writing and from their responses to self-report questionnaires would be more reflective of the current, established understanding of these constructs and more helpful in expanding our understanding of the relationships between students' writing experiences and the development of their thinking dispositions.

11.3 Results

11.3.1 Qualitative Analysis: Indicators of Thinking Dispositions in the Actual Writings

In this section, we discuss the results from the qualitative analysis of the student writings. We provide first the descriptions of the science-content space and discourse space indicators revealed in the thinking dispositions and manifested in the students' actual writings. Table 11.1 shows the categories obtained from the qualitative analysis of all the genres and content units for the three thinking dispositions. Generally, we see that content elaboration (content space) and linguistic formulation (discourse space) are the two main features that were revealed as a consequence of the discourse analysis. Table 11.1 presents the four dimensions identified with respect to the writer's disposition to examine data from multiple perspectives concerning the content of different subjects and topics:

Table 11.1 Description of qualitative categories for analyzing a writing task in terms of thinking dispositions and dimensions

	Science content		Discourse context	
Category	Dimension	Description	Dimension	Description
Viewing information from different perspectives	Domain	Diversity of subject areas or topics (biology, physics, social science, etc.)	Structure	1. Existence of a structure 2. Expression of structure with more than two components
	Focus	Specification of reference within topic (micro–macro, personal–universal, far–close)	Coherence	1. Writing that fits the topic 2. Situational coherence 3. Global coherence
	Attitude	Expression of opinion (negative–positive)	Rhetoric	Explicit dialog with a reader
	Complexity	Detailed explanation, elaboration, and relations within topic	Improvement	Occurrences of elaboration
Sensitivity to context	Task	Sensitivity to the task	Register	Fit between language and genre
	Writer	Sensitivity to the writer's state	Presentation mode	Fit between genre and presentation mode
	Integration	Combination of sensitivity to both task and writer	Increment	Changes due to context sensitivity
Drawing new distinctions	Comparison	Expressions of similar and different features of phenomena, procedures, and principles	Comparison	Occurrences of differences and similarities in lexical and rhetorical components of the text
	Generalization	Occurrences of generalizations of concepts or principles	Generalization	Occurrences of combining and generalizing language ideas to a concept or principle
	Awareness	Expressions of reflection concerning personal and new science insights of the writer	Awareness	Expressions of reflection concerning personal new language insights of the writer

- *Domain*: Concerns the subject areas referred to (e.g., economic, technological, social, chemistry).
- *Focus*: Reflects the writer's positioning vis-à-vis the text (close–far, personal–universal, micro–macro).
- *Attitude*: Reflects the author's opinion (positive–negative, for–against).
- *Complexity*: Refers to the structure and number of links between elements.

When referring to the discourse context, the four dimensions represent:

- *Structure*: Concerns the existence and textual expressions of text structure (arguments, counter-arguments, explanations, conclusions).
- *Coherence*: Reflects the writing's fit to topic and both situational and global coherence.
- *Rhetoric*: Reflects the form and style of discussion.
- *Discourse improvement*: Concerns the occurrences of discourse elaboration.

Table 11.1 also shows the three dimensions of contextual sensitivity identified with regard to the science-content space:

- *Task*: Concerns with sensitivity to the task (sensitivity to scientific, social, or application aspects).
- *Writer*: Expresses sensitivity to the writer's state (related background, disposition, previous experience).
- *Integration*: Reflects sensitivity of the writer to the combination of sensitivity to both task and writer.

In terms of the discourse space, three dimensions were identified:

- *Register*: Concerns the fit between language and genre.
- *Presentation mode*: Reflects the style of the discussion and the fit between presentation mode, genre, and writing context.
- *Increment*: Reflects changes occurring in the writing due to context sensitivity (transitions in writing style from personal to more scientific and cautious language).

Finally, looking for indicators of drawing new distinctions demonstrated three similar dimensions referring to both the science-content space and the discourse space:

- *Comparison*: Concerns expressions of similar and different features of scientific phenomena, procedures, or principles and occurrences of differences and similarities in lexical and rhetorical components of the text.
- *Generalization*: Expresses occurrences of generalizations of scientific processes, concepts, or principles as well as instances of combining and generalizing language ideas to a concept or principle.
- *Awareness*: Refers to expressions of reflections concerning personal and new science insights as well as expressions of reflections on personal new language insights.

In order to examine the general nature of the indicators reflecting student dispositions as expressed in the writings across the genres and science contents, it was important first to determine whether students had selected the different genres evenly across the learning units. The results showed that about 30% of students had selected tasks from all three genres, 63% had chosen two genres, and only about 4% always selected the same genre. The results also showed that in the first four units, the students' selection of genres was equally distributed over all the genres (about 35% selected a story, 30% diary, and 35% argument-debate). However, after completing the electricity unit, which required that students write an argument, more students (64%) selected the argument genre in the unit following electricity. The distribution of tasks selected by intervention group students according to units showed significant differences ($X^2 = 17.48$, $p < .008$, $df = 6$), demonstrating that the students' preference for writing genres was not consistent across all units.

In summary for the qualitative analysis, we found that we could use a similar set of indicators to describe student writings that reflected the three thinking dispositions explored in the different units containing the different science contents. We also found that the sets of indicators emerging from the writings could be generalized to the five genres used in this study.

11.3.2 Quantifying Measures of the Qualitative Categories

To establish measures expressing the thinking dispositions in the students' writings and to ensure we did not apply judgmental measures when evaluating the thinking skills, we examined the frequency of occurrence of the indicators in Table 11.1 and, for each criterion, counted the number of indicators found in the writings. Each indicator received 1 point. For example, regarding the criterion *Specification of topics used* in the context of the disposition *Viewing information from different perspectives*, we gave 1 point for each topic mentioned.

To make comparisons within categories of a particular thinking disposition, we applied a standardization procedure. We set the maximum frequency at 100; all category frequencies received a proportional value. The reliability coefficients of the total measures for the pre- and postfrequency measures of the thinking dispositions found were 0.75 and 0.81, respectively. The reliability of the specific measures was 0.67 to 0.80.

11.3.2.1 Differences between First and Final Writing within the Intervention Group

To estimate the impact of science writing on the manifestations of thinking dispositions in students' writings, a within-group comparison was conducted for the intervention group between the first and final writing tasks. This involved comparing the measures of occurrence of thinking dispositions in the writings (Table 11.2). Multivariate repeated analysis of the data showed an overall significant difference between the beginning and end of the study ($F = 13.31$, $df = 1.46$, $p < .001$), significant differences between the different thinking dispositions beyond the measurement

Table 11.2 Measures of thinking dispositions as reflected in the first and last writing tasks of the intervention group

	First learning unit		Last learning unit		Difference	
	Content mean (*SD*)	Discourse mean (*SD*)	Content mean (*SD*)	Discourse mean (*SD*)	Content difference *t df* = 47	Discourse difference t *df* = 47
Viewing different perspectives	45.39 (27.7)	43.06 (24.5)	75.65 (17.7)	61.72 (20.4)	30.07 (30.5) 6.79^{***}	18.66 (31.7) 4.07^{***}
Drawing new distinctions	15.63 (29.4)	18.75 (39.44)	29.86 (28.5)	14.58 (27.2)	14.23 (39.6) 2.49^{*}	−4.17 (51.39) 0.56
Sensitivity to context	34.03 (27.1)	41.67 (20.8)	39.06 (22.4)	57.29 (22.4)	5.035 (37.8) 0.92	15.62 (28.54) 3.79^{***}

* denotes $p \leq .016$

*** denotes $p \leq .0001$

time (F = 61.94, df = 5.46, p < .00001), and a significant interaction between the measurement time and the disposition categories (F = 6.92, df = 5.46, p < .0001).

Significant interaction implies a different change in frequency of occurrence for the different thinking dispositions, with the greatest increase found for the content-based dimensions associated with *Viewing information from different perspectives* and the smallest change occurring for *Sensitivity to context*. Of the discourse-based dispositions, the greatest change was again apparent for *Viewing information from different perspectives* while the smallest change was found for *Drawing new distinctions*.

11.3.2.2 Differences between the Intervention and the Comparative Groups on the Final-transfer Writing Task

A one-way multivariate analysis of the frequency of occurrence measures for the dispositions in the students' final writings for both the intervention and comparison groups (see Table 11.3) showed significant differences between the groups (F = 8.806, df = 1.94, p < .0001). A univariate analysis demonstrated that the differences between the groups were significant for all content-based measures and one discourse-based measure. The differences for the content-based measures in standard deviation units ranged from 0.7 to 1.5. For space limitations, we do not show additional data concerning the differences between the two groups on the more specific dimensions for the content and discourse categories (see Table 11.1). It is, however, important to say that significant differences between the groups existed for all the content and discourse dimensions, with the exception of the discourse structure dimension.

Table 11.3 Measures of thinking dispositions as reflected in the last writing task: Differences between intervention and comparative groups

	Intervention group		Comparison group		*F* (1, 96)	
Measures	Content mean (*SD)*	Discourse mean (*SD)*	Content mean (*SD)*	Discourse mean (*SD)*	Content	Discourse
Viewing different	75.46	61.72	48.38	43.49	50.6***	20.79***
perspectives	(17.7)	(20.4)	(19.5)	(18.8)		
Drawing new	29.86	14.58	9.03	5.21	16.96***	3.54
distinctions	(28.5)	(27.2)	(20.3)	(21.2)		
Sensitivity to	39.06	57.29	21.35	50.52	13.51***	2.63
context	(22.4)	(22.4)	(24.7)	(18.2)		

*** denotes $p \leq .0001$

11.3.3 Quantitative Analysis of Self-assessed Thinking Dispositions

In this section, we analyze the quantitative measures of students' self-assessed thinking dispositions. To study the effects of writing experiences in science on the students' self-perceived thinking dispositions, a repeated multivariate analysis was performed. Table 11.4 presents the means and standard deviations for each thinking disposition for the two groups as measured pre- and post study.

The analysis demonstrates a significant difference and interaction between the group and the measurement time, indicating that the growth in thinking dispositions of the intervention group was significantly higher than the growth in thinking dispositions of the comparative group. Similar results were found for all thinking dispositions.

Table 11.4 presents the interaction effect, showing no differences in the thinking disposition *Viewing information from different perspectives* between the groups prior to the study and a significant difference between the pre- and postmeasurements for the intervention group (1.66 *SD*). No significant difference was found for the comparative group (0.21 *SD*). Furthermore, the intervention group scores at the end of the study were significantly higher than the comparative group scores (1.13 *SD*).

11.4 Summary and Discussion

This study adopted a dialectical approach to mixed-methodology research and synergetically integrated principles from the quantitative and qualitative research styles. Since the study focuses on interrelated issues involved in research methods

Table 11.4 Differences between intervention and comparison groups on the self-perceived measures of student thinking dispositions after the first and final learning units

	First learning unit		Final learning unit		*F* ratios with *df* = 1,94 and (*p* values)		
Measures	Inter. M (*SD*)	Comp. M (*SD*)	Inter. M (*SD*)	Comp. M (*SD*)	Between groups	Time	Interaction T X G
Viewing different perspectives	2.65 (0.95)	2.76 (1.12)	3.97 (0.64)	2.99 (1.13)	5.29 (0.024)	147.6 (0.0001)	70.8 (0.0001)
Drawing new distinctions	2.85 (0.98)	2.57 (1.16)	4.20 (0.53)	2.94 (1.09)	17.03 (0.00001)	160.2 (0.00001)	52.23 (0.00001)
A need and willingness to think	3.35 (1.07)	3.12 (1.10)	4.01 (0.73)	3.36 (1.08)	5.04 (0.027)	54.36 (0.0001)	2.56 (0.001)
Appreciation of meta-cognition	2.89 (0.99)	2.46 (1.04)	3.67 (0.75)	2.79 (1.06)	11.86 (0.001)	90.92 (0.0001)	14.23 (0.0001)
Enjoyment of thinking	3.08 (0.91)	2.67 (0.95)	3.63 (0.70)	2.96 (0.96)	10.18 (0.002)	47.25 (0.0001)	4.86 (0.030)

and science education, the discussion examines the study from two different but related perspectives: the methodological perspective and the general science perspective, which we examine both separately and together.

The study supports the use of informal writing tasks—not for their role in acquiring science knowledge but for the thinking dispositions they can encourage. According to the prevailing fundamental sense (Yore et al., 2007) of scientific literacy, these dispositions are among the characteristics required of scientifically literate individuals and encapsulate abilities and emotional dispositions that reflect the speculative, personal, temporary, and rational attributes of science knowledge and the scientific process. They also reflect skepticism in generating temporal explanations and plural rather than singular interpretations of world phenomena (Hand et al., 2001). The study also highlights the importance of the dialogical connections between content and discourse when writing is viewed as a process of knowledge transformation and when student tasks are open-ended, exploratory, and personal. In terms of enhancing scientific literacy and understanding the processes of writing-in-science and its effect on students' habits of mind, the study confirms our expectations that—when linked to the study of science topics—informal writing not only helps to develop students' self-perceived thinking dispositions but also their tendency to use these thinking dispositions in their writing. Furthermore, the improvement in students' self-perceived and use of thinking dispositions was not just a consequence of practicing familiar writing genres but followed writing experiences, which stimulated participants to utilize mindful thinking skills in novel writing situations.

The study demonstrates that in the context of science education research quantitative and qualitative research can be methodologically integrated thereby bringing opportunities, responsibilities, and challenges to the researcher regarding the development of effective and informative research designs while advancing scientific literacy. The study offers an example, a case study if you will, of mixed-methodology research or a hybrid of principles from different methodologies. Rather than seek causal determination, prediction, and generalization of findings or alternatively attempt to provide illumination and understanding of a particular situated phenomenon, this study seeks a different class of knowledge by using the strengths of more than one paradigm (positivist–constructivist) or methodology (quantitative–qualitative) in order to enhance the performance and implications of the other.

For the purposes of our discussion of the methodological principles applied in the study, we sought generally accepted criteria for characterizing a mixed-methodology approach and judging its values, advantages, and limitations. However, because a consensual list of criteria does not seem to exist and because the principles underlying the qualitative and quantitative methodologies do not always match, we used several characteristics of research methodology—some of which are based on parallel methodological concepts found in quantitative and qualitative research while others are based on our experiences in this study. These should be regarded as strategic ideas that provide a direction for developing specific mixed-method designs and data collection procedures.

11.4.1 Predictable, Standardized Design Combined with Emergent, Intervening Processes

The study design was almost entirely predefined; it was not structured as a dynamic and flexible process emerging from a given situation prompted by the unique features and needs of a specific classroom. Rather, the teachers and students were asked to follow directions regarding specific writing assignments as well as when to use them during science classes. They were not at liberty to decide whether and when to write nor whether and how to use science writing in the classroom.

Nevertheless, there were three occasions when the planned design proved problematic and required adjustments. The first occasion was when students sought feedback on their writing after the first writing assignment. We consequently altered the design so that feedback became part of the routine study design. The second occasion involved the fourth unit when only the argumentative writing task was used. Following the teachers’ request to find out whether their students had understood the analogy between electricity and water, we allowed use of the debate genre. The third occasion was when the students became partners in writing the last two writing tasks.

11.4.2 Controlled Manipulation of an Intervention Combined with Freedom of Choice

The study used a natural setting in that it did not extract participants from their classes but kept classes intact, which means that the study manipulated and helped to change science learning in one group—the intervention group—with no random assignment of students for treatment. The study observed, described, and interpreted the setting during the intervention without any reference to the setting's prestudy conditions. Moreover, the intervention conditions were predefined in terms of the sequence of learning units, number and types of writing genres, use of reflective writing tasks specially designed for the study, and the strategy of exploring both students' self-perceived thinking dispositions and which dispositions were reflected in student writings. Furthermore, the use of a comparison group that was similar to the intervention group in several important variables actually established controlled conditions that approximated quasi-experimental design. The methodological concepts closest to these experimental control elements are internal validity in quantitative research and credibility in qualitative research.

However, within this controlled manipulation of the learning process, the study used a set of three defined writing genres for all learning units, barring the last. Within the controlled and preset framework of writing opportunities, students were free to choose one genre from a set of three. They could choose the same genre for every unit or a different genre. Although this reflects the constructivist view that students should be able to choose tasks that speak to them personally, their choices also furthered our understanding of student writing genre preferences. As one would expect of quality research, this illustrates how the study can provide data on the processes implemented during the study as well as information on the outcomes of the writing process. Naturally, this has its own limitations, which in this case prevented us from assessing the differential contribution of each writing genre separately.

11.4.3 Predefined Intentional Measures and Locally Situated, Constructed, and Evolving Measures

The analytical approach in the discourse analysis of student writing was constrained neither by standardized methods nor by theoretical categories. Similarly, it only partially utilized predefined dimensions of thinking dispositions.

Both factors facilitated the identification of thinking dispositions, which not only related to the science content but also highlighted features associated with language and discourse. The study used two sets of measures: one derived from critical thinking theory and the other from mindfulness theory. Finally, the conceptualization that arose from the study was based on a statistical analysis of the students' writing. The two sets of measures shared common theoretical themes. However, since one pertains to self-perception and the other to the actual use of the thinking dispositions, they offer their own unique qualities—providing depth and detailed tools for better explaining the benefits of using informal writing in the science classroom. The discrepancy between self-perceived dispositions and thinking dispositions revealed in written texts highlights the value of the integrative nature of a mixed-methodology study. This discrepancy will probably also appear in future research involving both the theoretical and empirical study of the relationship between perceptions and actions, between implicit and explicit behaviors, and between two different representations of student thinking dispositions.

11.4.4 Specific, Situational, and Generalizable Methods

No research—either quantitative or qualitative—is free from its cultural context. Although this study was conducted in a particular setting with a particular group of students, the fact that it builds on several conceptually related variables implies that it can be used to develop a single explanatory model of the effects of writing-in-science on student thinking dispositions. Also, given that the design largely controls for the intervening variables by using a comparative group, the results seem generalizable as long as the generalization applies to similar contextual conditions, namely, a science classroom where students use a variety of informal writing genres including reflective writing.

Although we almost take generalizability for granted in quantitative research, it is also very carefully considered in qualitative research when using random sampling. In qualitative research, instead of simply assuming that we can generalize findings, we evaluate the transferability of findings to a particular group or population on a case-by-case basis (see Rossman & Yore, Chap. 26). Thus, generalizability does not depend on sampling strategy but on substantive data (Patton, 1980) and, therefore, applies to this study as well. Nevertheless, it is unclear whether we can generalize the findings of this study to different writing genres. It is doubtful whether we should generalize each genre's contribution to enhancing students' thinking dispositions. In this context, we note that at one end of the epistemological spectrum are researchers who reject the notion of dependability or consistency on the grounds that every phenomenon is rooted in a place and time that is viewed by a particular observer–researcher who is inextricable from the phenomenon and, therefore, the study can never be replicated.

11.4.5 Interpretation of Raw Data Coupled with Statistical Analyses

The study used two types of data interpretation: manipulation of raw data directly relating to the data source (discourse analysis), which is affected by the researcher’s individual beliefs, and statistical analysis of the measures formed as a result of the data interpretation based on qualitative analysis and quantitatively predesigned measures. Naturally, the predesigned quantitative measures are far less dependent on the researcher’s personal belief systems. In any case, using more than one researcher to analyze the raw data reduced subjective bias in data interpretation. The reliability of this study was, therefore, based on logical reasoning, consensus, and the use of statistics—implying that neither quantitative nor qualitative data were privileged per se. The most important methodological criteria of the study are objectivity and confirmability. Quantitative research is predicated on an assumption of objectivity, implying that the researcher does not bias the phenomenon studied or the results in any way. In qualitative research, however, the closest one gets to objectivity is confirmability (Lincoln & Guba, 1985), which refers to the quality of the data—in other words, whether it is confirmable by other observers or interpreters. Again, some researchers using qualitative methods may reject the notion of confirmability altogether for different epistemological reasons.

11.4.6 A Differentiated, Detailed View Combined with Interconnectedness and Authenticity

A mixed-methodology study can inform educational practice in both the local setting and the broader context. Typically, qualitative analysis provides the detailed context or processes lacking from quantitative research, whereas quantitative analysis broadens the implications of a purely qualitative study by measuring products or outcomes. Additionally, the use of multiple datasets can inform the research itself, yielding insights and methodological changes that improve the study and strengthen its findings. In this study, the unique pattern of the choice of genre between the fourth and fifth learning units stimulated ideas regarding the attractiveness of a genre and insights into each genre’s particular characteristics. Furthermore, the use of multiple genres is more likely to reflect authentic, practical situations in the classroom. Had we only used one genre at a time, we would have needed many more classrooms working with different informal writing genres and would still have not come any closer to generalizing our findings to the real-life classroom.

11.4.7 Generating and Testing Theories

It is generally agreed that qualitative analysis chiefly concerns direct experience in a given setting whereas quantitative analysis documents occurrences and tests hypotheses. Inherently, exploratory qualitative research is expected to generate new theories and ideas. Quantitative data, on the other hand, are most valuable for evaluating established hypotheses and theories. This was indeed our experience with this study, which tested theories relating to the contribution of writing-in-science to developing students' critical thinking dispositions. In addition, writing-in-science plays a role in emergent theories that examine the interaction between writing genres and writing contents. Since this study also analyzed student writings with reference to thinking skills and dispositions that are not presented in this chapter, we can actually suggest an even broader theory concerning the value of informal writing on diverse genres in enhancing students' habits of mind; however, this is beyond the scope of this chapter.

Finally, we realize that researchers may encounter problems when trying to combine two divergent research paradigms and that they will find they cannot perform either well. To avoid this, we must bear in mind Weinreich's (2006) argument; namely, an integrative research approach demands expertise in both research methods. Such integration is founded on a complex dialogue in which multiple voices converge and diverge through the tensions imposed by researchers' strong internal prejudices, epistemological and methodological beliefs, and paradigmatic commitments as well as their traditional habits of actions. Since meaningful mixed-methodology study requires integration and the juxtaposition of habits, knowledge, and research procedures, coupled with open-mindedness, courage to view practice and theory as equally significant, and an ability to depart from traditional paradigmatic frameworks, we uphold Wasser and Bresler's (1996) notion of interpretive zone. They emphasized the need for researchers with different methodological expertise, experiences, and research beliefs and dispositions to interact and create new meaning and understandings through the process of joint inquiry. Thus, it is likely that the more mixed-methodology research becomes accepted the more experienced researchers will become in its use—until they no longer need to judge studies using criteria that only reflect a one-sided view of the research paradigm continuum.

References

Aikenhead, G. S. (1997, May). *A framework for reflecting on assessment and evaluation* (pp. 195–199). Paper presented at the Korean Education Development Institute international conference Globalization of Science Education, Seoul, Korea.

Berends, M., & Garet, M. S. (2002). In (re)search of evidence-based school practices: Possibilities for integrating nationally representative surveys and randomized field trials to inform educational policy. *Peabody Journal of Education, 77*(4), 28–58.

Borland, K. W., Jr. (2001). Qualitative and quantitative research: A complementary balance. *New Directions for Institutional Research, 2001*(112), 5–13.

Bresler, L., & Stake, R. E. (1992). Qualitative research methodology in music education. In R. Colwell (Ed.), *Handbook of research on music teaching and learning* (pp. 75–90). New York: Schirmer Books.

Budd, J. M. (2001). *Knowledge and knowing in library and information science. A philosophical framework*. New York: Scarecrow Press.

Chaopricha, S. (1997). *Coauthoring as learning and enculturation: A study of writing in biochemistry*. Unpublished doctoral dissertation, University of Wisconsin, Madison.

Classen, S., & Lopez, E. D. S. (2006). Mixed methods approach explaining process of an older driver safety systematic literature review. *Topics in Geriatric Rehabilitation: The Older Driver, Part 2, 22*(2), 99–112.

Cook, T. D. (1995, November). *Evaluation lessons learned*. Paper presented at the International Evaluation Congress *Evaluation '95*, Vancouver, British Columbia, Canada.

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology, 88*(4), 715–730.

Creswell, J. W. (2003). *Research design: Qualitative and quantitative, and mixed methods approaches* (2nd edn.). Thousand Oaks, CA: Sage.

Eberle, T. S. (2005, May). Promoting qualitative research in Switzerland. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, 6*(2), Art. 31. Retrieved from http://www.qualitative-research.net/fqs-texte/2-05/05-2-31-e.htm

Education Sciences Reform Act of 2002. Pub. L. No. 107–279, 116 Stat. 1940. (2002).

Eisenhart, M., & Howe, K. R. (1992). Validity in educational research. In M. D. LeCompte, W. L. Millroy, & J. Preissle (Eds.), *The handbook of qualitative research in education*. San Diego, CA: Academic Press.

Eisner, E. W. (1991). *The enlightened eye: Qualitative inquiry and the enhancement of educational practice*. New York: Macmillan.

Ennis, R. H. (1996). *Critical thinking*. Upper Saddle River, NJ: Prentice Hall.

Ercikan, K., & Roth, W.-M. (2006). What good is polarizing research into qualitative and quantitative? *Educational Researcher, 35*(5), 14–23.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd edn., pp. 119–161). New York: MacMillan.

Firestone, W. A. (1987). Meaning in method: The rhetoric of quantitative and qualitative research. *Educational Researcher, 16*(7), 16–21.

Flick, L. B. (2002). *An introduction to qualitative research* (2nd edn.). London: Sage.

Gilman, R. (1993). The next great turning: A growing awareness of our interconnections could revolutionize our culture. *In Context, 33*(Winter), 11–12.

Glanz, J. (1998). *Action research: An educational leader's guide to school improvement*. Norwood, MA: Christopher-Gordon.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.

Greene, J. C. (2008). Is mixed methods social inquiry a distinctive methodology? *Journal of Mixed Methods Research, 2*(1), 7–22.

Greene, J. C., & Caracelli, V. J. (1997). Defining and describing the paradigm issue in mixed-method evaluation. *New Directions for Program Evaluation, 1997*(74), 5–17.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation & Policy Analysis, 11*(3), 255–274.

Guba, E. G. (1990). The alternative paradigm dialog. In E. G. Guba (Ed.), *The paradigm dialog* (pp. 17–30). Newbury Park, CA: Sage.

Guba, E. G., & Lincoln, Y. S. (1988). Do inquiry paradigms imply inquiry methodologies? In D. M. Fetterman (Ed.), *Qualitative approaches to evaluation in education: The silent scientific revolution* (pp. 89–115). London: Praeger.

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (pp. 105–117). Thousand Oaks, CA: Sage.

Halfpenny, P. (1979). The analysis of qualitative data. *Sociological Review*, *New Series, 27*, 799–825.

Hammersley, M. (1992). *What's wrong with ethnography? Methodological explorations*. London: Routledge.

Hand, B., Lawrence, C., & Yore, L. D. (1999). A writing in science framework designed to enhance scientific literacy. *International Journal of Science Education*, *21*(10), 1021–1035.

Hand, B., Prain, V., & Yore, L. D. (2001). Sequential writing tasks' influence on science learning. In G. Rijlaarsdam (Series Ed.) & P. Tynjälä, L. Mason, & K. Lonka (Eds.), *Writing as a learning tool: Integrating theory and practice* (Vol. 7 of Studies in Writing, pp. 105–129). Dordrecht, The Netherlands: Kluwer.

Howe, K. R. (1988). Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational Researcher, 17*(8), 10–16.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26.

Keys, C. W. (1999). Revitalizing instruction in scientific genres: Connecting knowledge production with writing to learn in science. *Science Education*, *83*(2), 115–130.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

Kuhn, D. (1991). *The skills of argument*. Cambridge, MA: Cambridge University Press.

Langer, E. J. (1993). A mindful education. *Educational Psychologist*, *28*(1), 43–50.

Langer, J., & Applebee, A. (1987). *How writing shapes thinking: A study of teaching and learning* (NCTE Research Report No. 22). Urbana, IL: National Council of Teachers of English.

Libarkin, J. C., & Kurdziel, J. P. (2002). Research methodologies in science education: The qualitative-quantitative debate [Column]. *Journal of Geoscience Education*, *50*(1), 78–86.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

Mackenzie, N., & Knipe, S. (2006). Research dilemmas: Paradigms, methods and methodology. *Issues in Educational Research*, *16*(2), 193–205. Retrieved from http://www.iier.org.au/iier16/mackenzie.html

Marton, F. (1986). Phenomenography — a research approach to investigating different aspects of reality. *Journal of Thought*, *21*, 28–94.

Maxwell, J. A., & Loomis, D. M. (2003). Mixed methods design: An alternative approach. In A. Tashakkori & C. B. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 241–271). Thousand Oaks, CA: Sage.

McCall, R. B., & Green, B. L. (2004). Beyond the methodological gold standards of behavioral research: Considerations for practice and policy. *Social Policy Report*, *18*(2), 3–19.

McCall, R. B., & Groark, C. J. (2000). The future of applied child development research and public policy. *Child Development*, *71*(1), 197–204.

Morgan, D. L. (1998). Practical strategies for combining qualitative and quantitative methods: Applications to health research. *Qualitative Health Research*, *8*(3), 362–376.

Nicolopoulou, A. (1997). Children and narratives: Towards an interpretive and sociocultural approach. In M. Bamberg (Ed.), *Narrative development: Six approaches* (pp. 175–195). Mahwah, NJ: Lawrence Erlbaum.

Niglas, K. (2004). *The combined use of qualitative and quantitative methods in educational research* (Dissertations on Social Science). Tallinn, Estonia: Tallinn Pedagogical University.

No Child Left Behind Act of 2001. Pub. L. No. 107–110, 115 Stat. 1425. (2002).

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, *87*(2), 224–240.

Onwuegbuzie, A. J., & Leech, N. L. (2004). Enhancing the interpretation of "significant" findings: The role of mixed methods research. *The Qualitative Report*, *9*(4), 770–792. Retrieved from http://www.nova.edu/ssss/QR/QR9-4/onwuegbuzie.pdf

Patton, M. Q. (1980). *Qualitative evaluation methods*. Beverly Hills, CA: Sage.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd edn.). Newbury Park, CA: Sage.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd edn.). Thousand Oaks, CA: Sage.

Peled, N. (1997). *Genres in speaking and writing — Theory and practice in the arts of language in teacher training institutions*. Tel Aviv, Israel: Mofet Institute.

Petter, S. C., & Gallivan, M. J. (2004, January). *Toward a framework for classifying and guiding mixed method research in information systems*. Paper presented at the Hawaii International Conference on System Sciences (HICSS-37), Waikoloa, Hawaii.

Phillips, D. C. (2005). The contested nature of empirical educational research (and why philosophy of education offers little help). *Journal of Philosophy of Education*, *39*(4), 577–597.

Prain, V., & Hand, B. (1996). Writing for learning in secondary science: Rethinking practices. *Teaching and Teacher Education*, *12*(6), 609–626.

Reichardt, C. S., & Rallis, S. F. (1994). Qualitative and quantitative inquiries are not incompatible: A call for a new partnership. In C. Reichardt & S. F. Rallis (Eds.), *The qualitative-quantitative debate: New perspectives* (Vol. 61, pp. 85–91). San Francisco, CA: Jossey-Bass.

Rocco, T. S., Bliss, L. A., Gallagher, S., Perez-Prado, A., Alacaci, C., Dwyer, E. S., et al. (2003). The pragmatic and dialectical lenses: Two views of mixed methods' use in education. In A. Tashakkori & C. B. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 595–615). Thousand Oaks, CA: Sage.

Sale, J. E. M., Lohfeld, L. H., & Brazil, K. (2002). Revisiting the quantitative-qualitative debate: Implications for mixed-methods research. *Quality and Quantity*, *36*(1), 43–53.

Shulha, L. M., & Wilson, R. J. (2003, May). *Collaboration and mixed method inquiry: Theory and practice*. Paper presented at the XXXI annual conference of the Canadian Society for Studies in Education, Halifax, Nova Scotia, Canada.

Shulman, L. S. (2004). *The wisdom of practice: Essays on teaching, learning, and learning to teach*. San Francisco: Jossey-Bass.

Sperling, M. (1995). Uncovering the role of role in writing and learning to write: One day in an inner-city classroom. *Written Communication*, *12*(1), 93–133.

Strauss, A. L., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Thousand Oaks, CA: Sage.

Tashakkori, A., & Teddlie, C. B. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

Tashakkori, A., & Teddlie, C. B. (2003). The past and the future of mixed methods research: From "methodological triangulation" to "mixed model designs". In A. Tashakkori & C. B. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 671–702). Thousand Oaks, CA: Sage.

Tishman, S., Perkins, D. N., & Jay, E. (1995). *The thinking classroom: Learning and teaching in a culture of thinking*. Needham Heights, MA: Allyn & Bacon.

United States National Research Council. (2004). *Advancing scientific research in education*. Committee on research in education. L. Towne, L. L. Wise, & T. M. Winters (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Visser, R. M. S. (n.d.). Trends in program evaluation literature: The emergence of pragmatism. *TCALL Occasional Research Paper*, (No. 5). Retrieved from http://www-tcall.tamu.edu/orp/orp5.htm

Wasser, J. D., & Bresler, L. (1996). Working in the interpretive zone: Conceptualizing collaboration in qualitative research teams. *Educational Researcher*, *25*(5), 5–15.

Weinreich, N. K. (2006). *Integrating quantitative and qualitative methods in social marketing research*. Retrieved April 27, 2008, from http://www.social-marketing.com/research.html

Yanchar, S. C., & Williams, D. D. (2006). Reconsidering the compatibility thesis and eclecticism: Five proposed guidelines for method use. *Educational Researcher*, *35*(9), 3–12.

Yin, R. K. (2003). *Case study research: Design and methods* (3rd edn.). Thousand Oaks, CA: Sage.

Yin, R. K. (2006). Mixed methods research: Are the methods genuinely integrated or merely parallel? *Research in the Schools*, *13*(1), 41–47.

Yore, L. D. (2000). Enhancing science literacy for all students with embedded reading instruction and writing-to-learn activities. *Journal of Deaf Studies & Deaf Education*, *5*(1), 105–122.

Yore, L. D., Pimm, D., & Tuan, H.-L. (2007). The literacy component of mathematical and scientific literacy. *International Journal of Science & Mathematics Education*, *5*(4), 559–589.

Chapter 12
New Directions in Science Literacy Education

E. Wendy Saul and Brian Hand

Research on science education and research on literacy education have been informed by different theoretical traditions, different programmatic efforts, and different goals or conceptions of competence. The issue we address in this chapter, however, is whether the methods used to examine science and literacy education should be the same. More specifically, in seeking to bring together and develop new research that informs the convergence of science and literacy pedagogy, what questions and balance do we advocate for, given the constructs on which the two fields build and their varying degrees of development? Like Munby (2003), we will argue that “research at its most fundamental is an *argument* that leads us through purpose, related literature, data, and analysis to a specific point” (p. 155). Thus, the issue of method—the appropriateness of approach, the rigor of the study, and the generalizability of the conclusions—needs to be based on a strong connection among the question(s), claims, and evidence. Although the evaluation of method can be viewed as a paradigm war, we further agree with Munby that these differences in perspective are best viewed as opportunities for all researchers to examine various perspectives of the “human conditions within our research” (p. 154). To highlight this conception, we have conducted lengthy discussions and identified four issues that can be used to frame the dialogue around methods, both for us and for our readers:

- The identification of important and useful questions.
- The relationship among research question, method, and program.
- Unaddressed and under-researched questions.
- Practitioner research and research on practice.

In providing a context for this discussion, we turn to Prain (see Chap. 8), who offers a framing of science literacy research using a continuum that provides different

E.W. Saul
University of Missouri-St. Louis

B. Hand
University of Iowa

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

methodological perspectives and orientations in terms of outcomes. In expanding on different areas of research that are being conducted or need to be conducted, Prain highlights his concerns over the need for a variety of methods to be adopted in the explorations of literacy and science pedagogy. He suggested that new theoretical insights on which flexible interpretations of existing methods or new methods of research can be linked were needed to explore and build persuasive arguments about disciplinary literacy achievement involving diverse learners. Prain's notions align well with arguments like that posited by Grossman (2008):

> Our field needs to regain a critical balance between carefully designed, small, in-depth qualitative studies and larger scale research programs that intentionally take advantage of different methodologies. Larger scale studies may lose some of the nuanced descriptions of teacher education—in fact, it is almost inevitable that they will—but sometimes our attention to nuance leaves us unable to answer pressing policy questions, questions that ask for broad scope rather than telescopic detail. (p. 20)

While Grossman calls for a rebalancing in methodological approach and targeted subjects, each individual study would not necessarily employ mixed methods. Instead, the underlying assumption is that useful research, in the aggregate, will avail itself of both quantitative and qualitative approaches, both small-scale and large-scale studies. However, chapters in Part III argue for mixed methods and offer up examples of research in which a single research team has used both quantitative and qualitative data. Although Nieswandt and McEneaney (see Chap. 10) have a much more quantitative emphasis than Levin and Wagner (see Chap. 11), both chapters can be viewed as employing mixed methods and sitting relatively close to one another on the continuum. The differences in method choice and orientation can be justified because the target constructs under consideration were at different stages of conceptual development and instrumentation. The chapter by Norton-Meier and colleagues (see Chap. 9) further highlights potential differences in approach to mixed-methods research. Because they are seeking to examine interactions that occur at the project level rather than at the individual study level, they have employed a combination of approaches that enables them to address a problem space related to teacher implementation and the consequential impact on student learning. In this case, quantitative and qualitative approaches are used to study a variety of science content, grade levels, and classroom settings across the overall project; and the choice of method is determined and informed by both the larger question posed and the available data.

Munby (2003) suggested that discussions on appropriate research, in fact, should be framed around questions of purpose and rigor. He argued that we need to move past a purely technical view of reliability and validity and focus on the essence of research, that is, to persuade others of the trustworthiness of the results. Rigorous studies are designed so that the argument reflects the quality of the data and analysis and also responds appropriately and convincingly to the questions posed. For Munby, the argument needs to show a strong connection among the question(s), claims, and evidence.

Interestingly, Raudenbush (2005) made the point that “education research is an interdisciplinary effort long characterized by methodological diversity … [and then asks] why, then, do we hear an urgent call for mixed methods now?” (p. 25). The adaptation of mixed-method approaches allows researchers to engage a broader range of question(s) than any single method by itself can. Munby’s (2003) comments reinforce the notion of a continuum where the appropriate balance of qualitative and quantitative methods is not about a checklist of technical and procedural items but rather about how well the particular aspects of each method support the argument put forward by the researcher(s). We believe that such an orientation has special import because pedagogical research must consider pragmatics; that is, we are interested in the concept of how can we best connect traditional research to practitioners within classrooms and also how practitioner work and assessment can inform research practices that take place at the university or in research institutes.

Like Brickhouse (2006), we believe that literacy and science education researchers need to be held to “a higher ethical standard [in] that research should have at least some potential to improve the quality of education and the lives of children” (p. 4). In attempting to relay our perspectives on issues raised in Part III of this book, we have created a dialogue. To that end, we developed the four broad issues about literacy and science education research approaches and purposes mentioned earlier to frame our deliberations. Second, we made our thinking visible by sharing our initial positions on these issues and our negotiated common understanding from diverse perspectives of literacy and science learning and teaching by providing a summary of our deliberations on each issue.

12.1 The Identification of Important and Useful Questions

Brian: I have to agree with Munby (2003) and say that we often get trapped into a paradigm of research and thus only ever frame a question around the restrictive structure that it enables us to address, disregarding the primacy of importance and usefulness. This means that we are trapped into going through a checklist related to technical and mechanistic aspects of method rather than focusing first on the question. For me, the critical starting point of any research is the question, not the method. One size does not fit all problem spaces and research questions; therefore, as the 2nd Island Conference concluded: Gold Standard needs another ‘s’! The concept of quality research, that is, the trustworthiness and rigor of the research, is framed around the appropriate method adopted to address important and worthwhile questions. Furthermore, since pedagogy must be practical, it means the questions that pass the importance and worthiness filters must also have the potential of delivering useful results. These fundamental criteria not only may be beyond a single research approach, they may be bigger than a single researcher. We need to adopt the model of science and view research as a team endeavor where the necessary people and expertise are recruited to the process so that research questions can be addressed fully. We need to be much more open to both the value of team research and to the ideas that such a collective brings required expertise, diverse perspectives, and multiple disciplines to science literacy for all questions under investigation.

Munby (2003) stated that quality research is about building compelling and persuasive arguments about these important and useful questions. The concept that research is about persuasion applies not only to the results but also to the method. However, a current problem for Eisenhart and DeHaan (2005) is that many researchers do not have the background, experience, and emotional disposition (habit of mind) to be able to engage in such openness and diversity of approaches. They have suggested that educational researchers need expertise and insights in five broad areas: “(a) diverse epistemological perspectives; (b) diverse methodological strategies; (c) the varied contexts of educational practice; (d) the principles of scientific inquiry; and (e) an interdisciplinary research orientation” (p. 7).

Well-prepared research teams need to be open, flexible, and opportunistic to engage the range of questions and the development of constructs embedded in the questions and to address embedded issues as they arise during an inquiry or series of inquiries. This means that, like the researchers in Part III have shown, mixed methods are the most appropriate research approach by which to address the important questions they identified. Mixed-methods research is not about some diplomatic middle ground between dominant positions at the dipoles of the empirical approaches; rather, it is about the research question and, thus, should be viewed as a broad continuum of methods. The richness and complexity of the classroom cannot be examined easily by any one method, which requires us to be vigilant in examining carefully what are the method and the adjustments within the selected method to answer the question. As Mayer (2000) suggested, the question about method is not one about which method is best but how data are used to support argument.

Wendy: This book, and the conference on which it is based, has sought to explore what is meant by Gold Standard research, particularly as it affects our understanding of science and literacy pedagogy. In asking if the elements we found are in fact the real thing, academics, practitioners, and policy planners have tended to focus on methods: Are they rigorous enough? Are the results generalizable? Using the same map, could others expect the same results? My worry is that in this gold rush for verifiable studies, too little attention has been given to:

- A careful examination of the questions being posed.
- Issues related to context (i.e., what, in fact, can be generalized and to what populations?).
- Follow-up studies that seek to explore how and why particular effects are seen in particular groups.
- The interrogation of the Gold Standard metaphor itself.

Any scientist worth his or her salt will surely tell you a research study can be no better than the questions it seeks to investigate. Although we all would like to know how cancer can be cured or how climate change can be slowed, no funding agency or peer review group would support a study based on such broad questions. The other end of the spectrum has to do with questions that appear impossibly narrow and that may appear to have no applicability. The best scientific research—those elegant studies—surprise the community and allow for cognitive leaps in understanding; they enable us to get past blocks in thinking and upend faulty assumptions.

In summary, like Brickhouse (2006) and Phillips (2006), we also caution researchers to be open to selecting appropriate and rigorous research designs, data collection, and interpretative frameworks aligned with the research question. We believe to do otherwise will leave important and worthwhile questions and applications without proper consideration and investigations. Furthermore, literacy and science education research runs the chance of converging onto the easy research questions and a single research approach without consideration of the more difficult, important, and useful questions and approaches that will influence public policy and inform classroom practices. Many of these difficult research questions deal with the complexity of learning, pedagogy, assessment, classrooms, all students, and the fuller participation in the public debate about science, technology, society, and environment issues.

12.2 The Relationship among Research Question, Method, and Program

Wendy: Several years ago the Holmes group was asked to identify studies that had a profound impact on education (Johnson, 1990). Their answer was Rowe's (1974a, 1974b) study of wait time. Although wait time is certainly an important concept, simply waiting the requisite number of seconds will not, in fact, solve the problems facing teachers today. Just as identification of a particular gene will not cure cancer, Rowe's study, at best, offers a small, incremental step in improving the prospects of schoolchildren today. What the example of wait time does ask us to consider, however, is this: What studies related to science and literacy offer insights of the same scale as those offered by Rowe? What questions related to science and literacy have the potential to serve as a springboard to understanding the salient issues that confront scholars and practitioners working in the field?

The Purcell-Gates, Duke, and Martineau (2007) study might serve as one example of research having such potential. For many years, scientists in US National Science Foundation-sponsored curricula and practitioners have promoted authenticity as an important value in the teaching of science. Purcell-Gates and colleagues explored what happens when authentic text and activities were used and concluded that inauthentic materials (i.e., have no analog in the real world and are just designed for school-teaching) were less effective than authentic texts in teaching both science and reading.

Although these researchers, and we, would like to see more work on this topic, this study should be valued, first, because of the importance of the problem space it sought to address and, second, for its potential to enable us to rethink and revise the materials we offer students. As the authors noted in their prefatory remarks:

> This study addresses the long-held debate regarding how language is best learned, particularly language forms that are not acquired as one's primary discourse (Gee, 1992) such as reading and writing. … The question becomes what combination of experience and explicit instruction best facilitates learning of new language forms. (Purcell-Gates et al., 2007, p. 8)

The debate to which they refer is significant to both the research and practitioner communities, and the study benefits both researcher and practitioner knowledge.

This study was also impressive in that the authors sought to complicate the questions they asked by looking at issues related to context. Through methodology and assertion, they recognized the complexity of learning, teaching, classrooms, schools, and the sociocultural and socioeconomic influences. Too many research models oversimplify schooling in that they only consider instructional resources, professional development, classroom practices, and student achievement and forget contextual factors. In the Purcell-Gates and colleagues' (2007) study, children in families with better-educated parents (which served as a proxy for higher income) were compared to children from families with less education. This added element confounded the data and greatly reduced the number of students in each cell, thus opening up the research to additional criticism; but without such conversation about student background, the research sits comfortably and impotently in a bell jar, pristine and useless. Importation and generalization of pedagogical research results without consideration of contextual and cultural factors in the target setting is unwise.

Interestingly, this takes us back to the notion of wait time. For the past year, I have been working in Liberia where class size often tops 100 students. Will wait time—clearly a good idea in the developed world—work in the Liberian context, I wonder, especially in the rainy season, when water pounds noisily on the metal schoolhouse roofs and almost no child voices can be heard singly?

Quality pedagogical research is reflected in the program of study and not in a single study. With even the best studies, follow-up research is necessary. For those seeking stable, large-scale answers, the promise of rigorous methods that can be used to reshape education is tantalizing. But I propose that in seeking answers we look to the variety of well-regarded methods used in the sciences, the place from which social science has borrowed its Gold Standard notions. In the double-blind drug study, for instance, we have learned to look carefully at side effects; when they are too dangerous or disturbing, we pull the drug from the market. Are we, in fact, willing to pull a newly purchased set of textbooks from classroom shelves if it does not prove effective for a given population? Are funding agencies willing to fund research methods akin to those used by Charles Darwin—careful observational studies designed primarily to build theory? Does Gold Standard empirical research include work comparable to that undertaken by epidemiologists? Some indications and insights into these questions can be found in the international funding priority, practices, and patterns for literacy and science education research (see She et al., Chap. 23).

Brian: Quality research agendas should consider the type of questions that need to be asked and also value the different questions that make us think outside of the box. For me, the question of rigor is one we need to investigate with exploratory research. That is, exploratory research needs to persuade others that a line of inquiry does provide opportunities to move further into an area that requires researchers to think outside of the box. The example of wait time is a good one because it raises issues related to the context of a study. It seems to me the context of a study is important in helping us locate the relative value of the outcome. Berliner (2002) used the phrase “the power of contexts” (p. 19) to highlight the critical difference between science research and educational research. He suggested this is what makes educational research so difficult and that qualitative components of research are so important because they do provide explanations of context.

A number of areas in science education have demanded lots of research time, but as yet I am uncertain of their impact on learning. For example, we are racing toward technology use in classrooms in large part because the technologists are moving forward at incredible rates and these off-the-shelf technologies are looking for purpose and market. We are told that these technologies will assist learning, but as yet I am uncertain how these technologies impact learning, what the critical pedagogies required for their use are, and how they promote students’ critical thinking and logical reasoning. Some educational technologies and software have promise in promoting knowledge-building communities, inquiry abilities, and metacognition; but their demands on memory and for computing power and a lack of stability have hindered their broad classroom applications. My argument is not about being negative, but where are the explorative studies addressing these questions or large-scale studies to demonstrate the educational value of these technologies?

Another similar area of research in science education is the nature of science studies. Again, my critique is not about the value of the nature of science but about related issues, such as: Does learning about the nature of science improve student learning of science? Does it help students be more active in the public debate about science, technology, society, and environment issues and take advantage of the science and technology career opportunities in their adult lives? We have researchers who ask different groups about their understanding of the nature of science without telling us how this impacts on the group’s understanding of science or how they can shape pedagogies that begin to address both learning and the nature of science. In other words, much of this work is replication studies. However, the work by Ford (2008), a relatively recent graduate, is shaping discussions about the role of scientists and how this should or could be reflected in classrooms. His work is interesting and could lead to some very interesting perspectives and a productive program of study on this very area.

My point here is that we need to think through more carefully about what our question is and how this will impact learning and teaching generally. Osborne (2007) highlighted this idea when he suggested that we need to be more careful in what we are doing as researchers. He suggested that we need to sit and stare for a while in order to shape our work. For him, it is not a case of not enough data but rather of not enough theorizing. For me, this translates into the concept of a research program, rather than single studies, as this will help us begin to shape studies from conceptual, exploratory studies to broader studies that will have a chance to have impact. The other critical element of a research program is that we have to be open to the critique necessary to move the work forward. This, I believe, certainly enables rich discussions about the relative merit and value of areas and to encourage thinking outside of the box.

In summary, we believe there is a need to be much more critical about what counts as useful research and how we count the value of this research. To move beyond a centimeter deep and a kilometer wide sort of approach, that is, replication studies or a single-focus study, we need to be more open to a topographical view of research program. By this we mean, instead of a continuum notion that does not necessarily deal with depth of research, we need to begin to see research more in a three-dimensional capacity. As such, the interaction among questions, methods, and programs of study can provide a richer sense of what is possible, what can or will have potential value, and what needs to be discarded because it is not leading anywhere.

12.3 Unaddressed and Under-researched Questions

Brian: In reflective pieces flowing from the 1st Island Conference, at least two research areas were highlighted: the concepts of multiple literacies (Hand et al., 2003) and representations (Yore et al., 2004). One overall issue is that the literacy and science education communities run the risk of squandering the opportunity of making a difference embedded in *science literacy for all* that targets all learners, fuller participation in the public debate about science, technology, society, and environment issues, constructivist learning models, and authentic assessment. The specific acceptance of what science literacy involves is far from consensus, but there is general agreement on it culminating in better citizenship and in two interacting components of fundamental literacy in science and derived understanding of the big ideas in science. For me, the area of representation is one that I believe is critical within science literacy. We are beginning to understand that students need to engage all of the representational forms of a concept to construct and demonstrate understanding of the concept. Although students are constantly exposed to different forms of representations of a concept because of the textbooks used or through the Internet, we as researchers are a long way from really understanding all the dimensions to how students learn using these representations or how we can best help them learn to engage better with these representation forms as cognitive tools. As with any new problem space, there are always going to be different perspectives and research orientations to this work. For example, can we provide the critical representation for students that will help them understand the concept(s)? This is a line of research going on currently (e.g., Ainsworth, 2006). Do we provide opportunities for students to engage some of the language-to-learn strategies as a vehicle to produce and move between different representations of the concepts? This is an emerging line of research (e.g., Hand, Gunel, & Ulu, in press). Are there clusters of representations used by students for different topic areas, and how do these clusters assist or detract from learning? This is a further line of research (e.g., Airey & Linder, 2006).

Each of these lines of research will have different methodological orientations and will lead to different perspectives on the topic. However, at the risk of sounding redundant, rather than seeing these as being radically different positions, we need to examine the rigor of the arguments put forward and look at the convergence points. Much of the work on representation at the moment is being looked at independently of the bigger picture of science literacy for all. For example, is graphing research viewed as a separate activity or as a skill that is needed to build a pathway to mathematical understanding? How does graphing fit the overall picture of representation?

We need to engage constantly and consistently in the theoretical orientations underpinning the research as well as the outcome of the research studies. For me, the significant work of Klein (1999, 2006) is his theoretical papers that push the field to think about issues and what are possible orientations to research that need to be engaged. This allows the outcomes of research, across the various methodologies, to both inform and push theory forward. Of course, the caveat here is that the connection to theory is that not all theory has to be new. As researchers, we need to review theory from older studies and reexamine their value. A colleague recently has introduced me to the early Systemic Functional Linguistics work as a means to look at our work on science literacy—this work is from 30 years ago.

Wendy: I, too, would put multiliteracies high on the list of under-researched topics, but I am as interested in students' ability to encode as well as decode various discourse modes—and the relationship between their encoding and decoding skills. Frankly, I suspect that it is in the back and forth that the greatest learning occurs. I think that there are also fascinating developmental issues to be unpacked. For instance, 50 years ago we assumed that young children would have no concept of gravity and that the teaching of space sciences was fairly inappropriate in a developmental sense, but recent generations have grown up looking at TV images of somewhat weightless astronauts floating in capsules. What has this, in fact, done to our perceptions? Perhaps what was viewed as developmental is often cultural. Similarly, young people growing up with computer screens flashing by may be able to perceive the world—or parts of it—differently. Are we, as educators, taking advantage of those experiences and skills as we study teaching and learning?

In the contemporary culture of research, we are fostering too much *black box* research. We study curricula as educator inputs and student outputs—Was X a successful intervention?—but spend too little time figuring out why a particular intervention worked with particular students or populations of students. I think that we need fewer studies of instructional programs and more and closer studies of classroom practices. School systems are being asked only to buy or buy into researched programs when, in fact, it could be a single aspect of a given program that produces the desired results. My examples tend to come from the literacy community but easily can be generalized to science. For instance, the notably successful and very expensive “Reading Recovery” program includes many distinct elements. Some of these elements appear to be particularly useful in promoting better reading and are not particularly expensive. Are there analogous programs to parse and study within the science community (e.g., aspects of programs that are particularly useful in increasing students’ knowledge and skills)? I strongly suspect that the answer is yes.

More important, perhaps, we have almost stopped looking at what people who produce double-blind drug studies would call side effects of instruction, curriculum, and assessment studies. Although test scores may rise, what are students really learning about what counts as knowledge? Years of work with students in the former Soviet Union, where scores on standardized tests dramatically top those in American schools, make me worry about the hidden curriculum of the current focus on testing. Are we teaching students to generate their own ideas? To evaluate and revise their work? To make reasoned decisions about what to do next … and, if that does not work, then what? Are we, in fact, using tests to replicate the social inequalities apparent in our society writ large—that wealthier people get to make decisions and have choices—and denying the less wealthy such opportunities for critical thinking, assuming that one must learn to follow orders before enlisting a more creative spirit? We need research that brings a moral lens to the tasks at hand—what Brickhouse (2006) called a “good standard” (p. 2), instead of a gold standard.

I also would like to see work that explores the ways in which science supports literacy learning—an implied possibility embedded in the interactive components of science literacy. To date, research has focused almost entirely on using talking, reading, and writing to learn more science; but I think that there is a claim to be made about science teaching students to understand better and interrogate text more fruitfully. It is not simply that doing a hands-on investigation helps to support prior knowledge—although it surely does. Rather, I think that we need to look carefully at the conversations—oral and written, as well as the kinds of interactions (communicative moves, use of resources, gestures) that take place around challenging hands-on activities.

Again, I would like to see more work on what promotes an attitude of inquiry. Both habits of mind and science inquiry are outcomes specified by the science literacy for all reforms. However, are those of us who work in schools taking full advantage of students’ ability to make meaningful choices? Years ago, Apple and King (1977) studied kindergarten students’ perceptions of work and play. It seems that the same activity, depending on whether students chose to do it or the teacher told them to do it, was defined variously as play or work. Play, as we know, engenders a different attitude and attention to activity than work, an attitude much more akin to what we call inquiry. What moves that can take place within a classroom support deep engagement and creative thinking about science tasks?

There is really so much we do not know and have not even begun to explore about learning, teaching, classrooms, schools, and sociocultural influences. My deepest worry is that in an effort to assert the importance of work undertaken to date we have moved prematurely to take a seat at the table of canonical knowledge.

The analogy I come up with is putting up a sign saying *ZOO* when all we have is an elephant, a kangaroo, and two pigeons. It is not that such animals are not an important part of a zoo, but are we ready to call what we have assembled comparable to the information gathered to date about other areas of science?

In summary, we recognize that the work undertaken to date in science and literacy is but a small fraction of the work that needs to be done in order to offer a robust understanding and canon of excellent practice. Science needs to be viewed as a subject that employs a variety of discourses. The best research recognizes this fact and works to find authentic instances in which these discourses can be brought into the classroom, which promotes science literacy for all and fuller participation in the public debate about science, technology, society, and environment issues.

12.4 Practitioner Research and Research on Practice

Brian: Brickhouse (2006) suggested that a big problem with science education is the orientation that universities do research and then teachers will use it. She stated “for universities to produce research, and for those research results to then be applied to practice, is a strategy that frequently has been tried and consistently has failed” (p. 5). Recently, a colleague and I explored the mismatch between the studies being reported in research journals compared to practitioner journals in relation to writing in science (Hand & Choi, in press). Our results suggested that the practitioner journals were really focused on strategies and these tended to be removed from current research work. For us, this was troubling because it raises issues related to the relevancy of our work and the question of translation into the practical world of the teacher. Two issues exacerbate this situation—one related to research and the other to the translation of research.

During the many professional development activities in which I have been engaged with teachers, they have never really questioned what my research methods are, how did I arrive at these, and why can I make claims about my work. There appears to be a divide between research and practice, in which teachers lack interest in the actual research process. Issues related to generalizability, validity, interrater reliability, effect sizes, significance, triangulation, etc., are not of great interest to practicing teachers. I believe that they tend to treat research data, claims, and evidence as being something for the researchers to deal with. Windschitl (2005) pointed out that very few teachers are exposed to educational research. Teachers are interested in what works and what impact it will have on their classroom. In terms of relating research to classrooms, I believe that mixed-method research is more teacher-friendly because it enables teachers to engage in some form of statistical outcomes as well as beginning to understand what is going on inside the heads of their students. We researchers have to show that there is a pragmatic outcome to our work and that it does have meaning. Single case studies tend not to apply to teachers because they cannot see the connection to their classroom, while large-scale studies often appear to be cold and not connected to individual classrooms. This is a disconnect that researchers need to address, and this is the second issue of translation of research results.

For us to have impact on classrooms, I believe that it is important that we undertake the translation of our work into language that engages teachers. In essence, when dealing with new strategies or approaches, teachers want to know if it will work and how my students will respond. While many of the articles in the teacher journals are teachers writing to other teachers, we as researchers need to be active in helping them understand how we can improve learning and teaching based on our research. This requires us to translate the language of research into the language of the classroom. For example, I am often asked to translate what a significant result for students using writing-to-learn strategies means in terms of percentage improvement. Typically, there is a tendency to hedge because we are reluctant to be definitive; somehow we need to address this issue. I strongly believe that the combination of some statistical outcome with commentary from teachers or students helps make the new work much more understandable and helps teachers become more engaged. The problem for researchers is to achieve this engagement while maintaining the integrity of the research process.

Wendy: I think it is wrong to assume that teachers are not interested in research processes. Rather, I think that as a group they have been taught to trust researchers in the same way that we trust our doctors when they prescribe or do not prescribe a certain test or procedure. Sure, there are plenty of people on the Internet proclaiming that this drug is harmful or useful; but when faced with a decision, we consult and ultimately trust our own practitioner. The real question for me is: What would it take for a teacher or policy maker to be convinced that a certain practice is wrong or needs to be revised? What evidence would be enough to change practice? What concerns me at present is that virtually no attention seems to be given to method or sampling or even to researcher-expressed caveats. Instead, practitioners and policy makers alike are seeking data they can use to justify their existing ideas. Frankly, given the difficulties of sorting through and analyzing data, I can hardly blame them. They tend to assume, rightly or wrongly, that we in the research community are policing ourselves and that if conclusions are not warranted we would not be publishing them or advocating for particular programs or practices.

The *hard science* research community—what we in education take as the Gold Standard—has pointed with some regularity to the use of unwarranted claims, for example, work designed to manufacture doubt about data on smoking or global warming (see Union of Concerned Scientists, 2007). If you go back to my zoo analogy, I think that we in the research community need to ask ourselves if we have enough data to support policy decisions or teacher-in-the-classroom decisions about what to do on a given day or in a given week?

In truth, we are working on different levels. When we in the research community make assertions, we are happy when we realize that a particular practice is *likely* to work—but our statistical level of significance ($p \leq .05$) does not guarantee success in a specific classroom of 25 students. Teachers, on the other hand, feel a deep moral imperative to help and find appropriate strategies and materials for every single child in their charge. This is not a question of statistics, but it does require every bit as much time as any researcher spends considering issues related to method. What most good teachers assume is that researchers offer a general direction—something to try first—but if that does not work, they are obliged to dig deeper, try harder, and think more creatively. Designers of curriculum too often assert that the research base assures success if only teachers would stick to their prescribed materials.

If the issue, as I proposed earlier, is that finding meaningful and rightly proportioned questions is key to improving this educational enterprise, I think we often make a mistake by not promoting and building on teacher questions. For instance, two teachers with whom I worked last term used a question board through several units in the school year. At the year-end, they looked at the questions students had generated, which seemed to suggest that the questions born from physical science units tended to be testable, the questions in the unit on genetics tended to be personal, and the questions from units on weather and space tended to be about large systems and were answered most satisfactorily through books. I am not sure if their conclusions would hold in a large study, but this is a fascinating thesis. Each term I am showered in my teacher-research class with questions of this scale, questions that beg, ultimately, for some methodologically sophisticated way to explore them. What are we as a research community missing by seeking to build curriculum and policy in a trickle-down model, where research is generated and conducted by methodologically sophisticated researchers and used by teachers who have little or no input into question formation or method?

In summary, we believe that pragmatic research is about working with and learning from practitioners. Over time we can develop communities of people—researchers and teachers—applying a multitude of methods that will benefit us all. As Pellegrino and Goldman (2002) suggested, “the community of educational researchers must include practitioners if it is to understand and draw its problems from practice and study them in practical as well as theoretically relevant ways” (p. 16). We are constantly amazed at teachers’ capacity to engage in this process—for example, an exceptional teacher who when videotaping her class carries the camera around and talks with the students. This does not bother the students at all, and we get to see a much richer picture of the negotiations they undertake. As researchers, we are part of this teaching and learning community—not above it.

12.5 Closing Remarks

As we began our conversation, the two of us wanted to examine what methods and research questions meant to each of us from our respective disciplines and then from the concept of science literacy and pedagogy. For both of us, the conversation needs to revolve around the ideas of border crossing (Saul, 2004) and convergence (Hand & Prain, 2006). Border crossing lives in the intent to move beyond our own disciplines, to listen to, act upon, and respond to concerns and critiques from those outside our respective fields. We also need convergence because there is a need for researchers to begin to coalesce around some critical issues as we move beyond our disciplinary borders.

However, there are also issues within each of our disciplinary areas where there is divergence that needs to be challenged and engaged. A small example has to do with a fairly well-agreed-upon assumption in reading: it is good to preteach difficult vocabulary words so that when students encounter them fluency will not be interrupted. Science, on the other hand, generally seeks a visceral understanding of concepts before the vocabulary word is associated with that notion. This same issue is played out on a larger screen as we talk about learning to use language as opposed to using language as a learning tool. For example, do we need to teach students the language structures of the discipline prior to using this language or do we encourage students to use language as an embedded component of the lesson? This issue informs current work on argumentation in science classrooms.

Importantly, in moving past discipline-based knowledge, the question of research has prompted us to view research methods more in a topographical manner rather than as a linear notion of a continuum as suggested earlier. Rather than arguing about the methods per se—though when we began this piece we both expressed a clear preference for mixed-methods work—we have come to believe that the emphasis on critiquing and using single studies as models is misplaced. Instead, we are ready to argue for programs of study where a combination of methods allows researchers, and the community they seek to persuade, to benefit from a range of studies using a range of methods. Such a program allows us to benefit from the richness of the detailed description and argument and from attention to issues related to reliability, validity, and other notions associated with large-scale studies. It should be noted that rigor and care are notions to be associated with both large- and small-scale studies. The perspective we proffer is really a call for us as a community to move beyond the two-dimensional continuum to a three-dimensional model that values the use of a purely quantitative, purely qualitative, or mixed-methods approach. It is the detail of that topogragraphical map that allows us to draw the best policy and classroom decisions from the complex information at hand. Said differently, it is in the combination, in the amalgamation and complexity, of studies that a research picture and program is built most successfully. In this way, exploratory studies are viewed as critical within a research program since they provide potential parameters for more large-scale efforts.

In summary, we believe that as a community we need to encourage scholars to adopt and become part of research programs that require long-term commitments to lines of inquiry that can impact student learning. Such an orientation, rather than jumping from one current fad to the next, requires a commitment to a topographical view of method that provides researchers, educators, policy makers, and politicians with the rich, quality data that are necessary and useful in decision making and persuasive argument.

References

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, *16*(3), 183–198.

Airey, J., & Linder, C. (2006, June-July). *Languages, modality and disciplinary knowledge*. Paper presented at the Integrating Content and Language in Higher Education conference, Mastricht, The Netherlands.

Apple, M. W., & King, N. R. (1977). What do schools teach? *Curriculum Inquiry*, *6*(4), 341–358.

Berliner, D. C. (2002). Educational research: The hardest science of all [Comment]. *Educational Researcher*, *31*(8), 18–20.

Brickhouse, N. W. (2006). Celebrating 90 years of *Science Education*: Reflections on the gold standard and ways of promoting good research [Editorial]. *Science Education*, *90*(1), 1–7.

Eisenhart, M., & DeHaan, R. L. (2005). Doctoral preparation of scientifically based education researchers. *Educational Researcher*, *34*(4), 3–13.

Ford, M. J. (2008). Disciplinary authority and accountability in scientific practice and learning. *Science Education*, *92*(3), 404–423.

Grossman, P. (2008). Responding to our critics: From crisis to opportunity in research on teacher education. *Journal of Teacher Education*, *59*(1), 10–23.

Hand, B., Alvermann, D. E., Gee, J. P., Guzzetti, B. J., Norris, S. P., Phillips, L. M., et al. (2003). Message from the “Island group”: What is literacy in science literacy? [Guest editorial]. *Journal of Research in Science Teaching*, *40*(7), 607–615.

Hand, B., & Choi, A. (in press). Writing in classroom science. In W.-M. Roth & K. A. Tobin (Eds.), *The world of science education: Handbook of research in North America*. Rotterdam, The Netherlands: Sense Publishers.

Hand, B., Gunel, M., & Ulu, C. (in press). Sequencing embedded multimodal representations in a writing-to-learn approach to the teaching of electricity. *Journal of Research in Science Teaching*.

Hand, B., & Prain, V. (2006). Moving from border crossing to convergence of perspectives in language and science literacy research and practice. *International Journal of Science Education*, *28*(2/3), 101–107.

Johnson, W. R. (1990). Inviting conversations: The Holmes group and tomorrow’s schools. *American Educational Research Journal*, *27*(4), 581–588.

Klein, P. D. (1999). Reopening inquiry into cognitive processes in writing-to-learn. *Educational Psychology Review*, *11*(3), 203–270.

Klein, P. D. (2006). The challenges of scientific literacy: From the viewpoint of second-generation cognitive science. *International Journal of Science Education*, *28*(2/3), 143–178.

Mayer, R. E. (2000). What is the place of science in educational research? [Research News and Comment]. *Educational Researcher*, *29*(6), 38–39.

Munby, H. (2003). Educational research as disciplined inquiry: Examining the facets of rigor in our work [Guest editorial]. *Science Education*, *87*(2), 153–160.

Osborne, J. (2007). In praise of armchair science education. *E-NARST News*, *50*(2). Retrieved from http://www.narst.org/news/e-narstnews_july2007.pdf

Pellegrino, J. W., & Goldman, S. R. (2002). Be careful what you wish for: You may get it: Educational research in the spotlight [Comment]. *Educational Researcher*, *31*(8), 15–17.

Phillips, D. C. (2006). A guide for the perplexed: Scientific educational research, methodolatry, and the gold versus platinum standards. *Educational Research Review*, *1*(1), 15–26.

Purcell-Gates, V., Duke, N. K., & Martineau, J. A. (2007). Learning to read and write genre-specific text: Roles of authentic experience and explicit teaching. *Reading Research Quarterly*, *42*(1), 8–45.

Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher*, *34*(5), 25–31.

Rowe, M. B. (1974a). Relation of wait-time and rewards to the development of language, logic, and fate control: Part II – Rewards. *Journal of Research in Science Teaching*, *11*(4), 291–308.

Rowe, M. B. (1974b). Wait-time and rewards as instructional variables, their influence on language, logic, and fate control: Part I - Wait-time. *Journal of Research in Science Teaching*, *11*(2), 81–94.

Saul, E. W. (2004). Introduction. In E. W. Saul (Ed.), *Crossing borders in literacy and science instruction: Perspectives on theory and practice* (pp. 1–9). Newark, DE: International Reading Association & National Science Teachers Association.

Union of Concerned Scientists (2007, January 3). *Scientists' report documents ExxonMobil's tobacco-like disinformation campaign on global warming science* [Press release]. Retrieved from http://www.ucsusa.org/news/press_release/ExxonMobil-GlobalWarming-tobacco.html

Windschitl, M. (2005). The future of science teacher preparation in America: Where is the evidence to inform program design and guide responsible policy decisions? [Guest Editorial]. *Science Education*, *89*(4), 525–534.

Yore, L. D., Hand, B., Goldman, S. R., Hildebrand, G. M., Osborne, J., Treagust, D. F., et al. (2004). New directions in language and science education research. *Reading Research Quarterly*, *39*(3), 347–352.

Part IV
Statistics, Research Methods, and Science Literacy

Chapter 13
Multilevel Modeling with HLM: Taking a Second Look at PISA

John O. Anderson, Todd Milford, and Shelley P. Ross

The purpose of this book is to provide a synthesis of thought and practice in research in literacy and science education intended to lead to evidence-based results and generalizations that will serve as a foundation for public policy and informed curriculum, teaching, and assessment practices in education. The Gold Standard of educational research funding in the United States can be viewed as a response to the general dissatisfaction with the utility of educational research; this federal mandate fosters a shift of educational research toward positivist empirical research approaches, such as random controlled trials (RCTs). There is an expectation of greater generalization and policy relevance as Gold Standard research is conducted and reported. It should be noted this dissatisfaction is not confined to the United States. An international response to this dissatisfaction with educational research—systematic reviews of educational research such as the Campbell Collaboration (Campbell Collaboration, n.d.), the What Works Clearinghouse of the US Office of Education (US Institute of Education Sciences, n.d.), and the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) in the United Kingdom (EPPI-Centre, n.d.)—has also identified the general dearth of rigorous empirical research that can support meaningful generalization of research findings.

The deliberations about qualitative–quantitative approaches to educational research over the last 25 years have established parallels for quantitative data considerations (reliability, validity, significance, objectivity) that consider dependability, credibility, believability, and confirmability of information (Howe & Eisenhart, 1990; Husen, 1988; Lather, 1992; Phillips, 2005; Pring, 2000). The enactment of these considerations has produced a diverse collection of qualitative, quantitative, and mixed-methods research in literacy and science education. Inspection of these research studies during the 2nd Island Conference revealed concerns about the clarity

J.O. Anderson
University of Victoria

T. Milford
University of Victoria

S.P. Ross
University of Alberta

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

of constructs involved in the research, measurements of these constructs, interpretative frameworks and scoring rubrics, data scales (nominal, ordinal, internal, ratio), and statistical analyses of these data. Procedural rigor, clarity, mining and secondary analysis of existing databases and information resources, generalization of research results, research ethics, and mixed-methods and other innovative approaches were identified as issues to enhance research quality (see Yore & Boscolo, Chap. 2).

The description and measure of constructs starts with (a) clear understanding and articulation of the target ideas and relevant behaviors; (b) transparent, justified, and verified interpretative frameworks (scoring rubrics, warrants of data as evidence based on established theoretical backings, etc.); and (c) the identification of appropriate and powerful analysis techniques. Many measurements and statistics are robust and accommodate limited violations of the underlying assumptions about data types, homogeneity, etc.; however, awareness of these assumptions and the recognized violations are essential for quality research. Some qualitative researchers do not acknowledge the value of using established, consistent, and sustained data collection and interpretation to serve as anchors and markers for cross-study analyses and generalization of the collective results related to a specific problem space and similar research questions. Furthermore, many quantitative researchers do not seem to be aware of the newest data collection and interpretation techniques used in mathematical statistics and social sciences.

This chapter introduces the reader to a broad spectrum of methodologies that are aligned with rigorous scientific empirical research and are fully compatible with the research standards embraced by both the research and policy communities. These methods include data mining, graphical analysis and representation, item response theory (IRT) test analysis, and multilevel modeling—methods that recognize the inherent complexity of educational data and also foster accessible and representative reporting of results. This chapter also introduces the reader to multilevel modeling of educational assessment data from large-scale international studies that goes beyond the political statement of rankings based on simplistic interpretation of achievement results. More specifically, we provide a description of the use of hierarchical linear modeling (HLM) with data from the Programme for International Student Assessment (PISA, Organisation for Economic Co-operation and Development [OECD], 2001).

13.1 Background

Large-scale assessments of student performance can provide a window into broadly defined concepts of student achievement in relation to some of the correlates of learning—such as student background, attitudes, and perceptions—and perhaps school and home characteristics. Given the scale of public funding for education and the importance of the literacy domains of science, mathematics, reading, and problem solving in terms of human capital in our so-called knowledge economy, it is of both interest and importance to identify consistent relationships between student and school characteristics and student achievement in these domains.

These characteristics include student traits such as gender, strategies for cognition, personal attitudes, and self-perceptions, and school traits such as size, school climate, and academic focus. If we better understand these relationships and the extent to which they are accessible through policy, it is more likely that the effectiveness of schools could be enhanced. Furthermore, this work can have theoretical significance in that enhancing understanding of relationships amongst the literacy domains and the student–home–school level correlates of learning can lead to the articulation of insights and relational patterns necessary for the development of theoretical frameworks and scaffolds for enhanced learning.

This chapter describes one avenue to investigating these relationships—multilevel modeling of student achievement data that is coupled to student and school variables—by means of a brief description of the assessment program and the data it generates, an introduction to analytic approaches, a summary of some preliminary results, and a consideration of potential research opportunities. It should be noted that another analytic pathway for investigating and modeling complex data is structural equation modeling (SEM) (Byrne, 2001; Kline, 1998). SEM explicitly models the relational patterns and structures associated with both observed and latent variables in a multivariate dataset (see Nieswandt & McEneaney, Chap. 10). A number of SEM studies within the science education research domain have been reported over the past two decades (George & Kaplan, 1998; Marsh & Yeung, 1998; Reynolds & Walberg, 1991; Yore, Craig, & Maguire, 1998).

Analyses that are conducted subsequent to, and often independent of the initial program that generated the data, are termed secondary data analysis (SDA). SDA can extend the analyses based upon previous research and can capitalize on alternate research perspectives to mine the data for new knowledge of the subject of interest (Hakim, 1982). These datasets constitute one of the most underutilized resources in education, and they exist for a broad spectrum of jurisdictions (school–college–university, districts, state–province, national–federal, and international) of the educational system. A cost–benefit analysis would reveal that full value of these extensive assessment programs has yet to be achieved. Large-scale assessment programs, such as PISA, provide the researcher with an opportunity to work with data that have been collected from rigorously designed samples of students and schools within well-defined educational jurisdictions. Data collection is conducted on a large scale as the name suggests, and administrative procedures are well documented and standardized. The achievement measures are well designed and described, and the scoring is designed and administered to minimize inconsistencies and error. Other advantages of SDA lie in (a) the cost savings of time, money, and personnel (Kiecolt & Nathan, 1985) and (b) the effective use of quality data beyond the initial focus on international and jurisdictional comparisons of mean performance.

The datasets made available from large-scale assessment programs are typically nationally or internationally representative. These datasets are usually designed to represent a specific population, such as the 15-year-olds targeted by the PISA study. The high response rates generated by the administrative protocols implemented in these programs mean that researchers can assume good representation of the target population and can generalize their findings across that same target population when interpreting their results (Hofferth, 2005).

Additionally, for the PISA dataset, there is careful quality control at all points in the data collection and analysis. The nature of the items and instruments is fully documented so that researchers can better understand the nature of the variables in the dataset (OECD, 2003a). The achievement test items and the questionnaire items for each PISA cycle are carefully field-tested before the final test booklets and questionnaires are created (OECD, 2003b). Test administrators at each testing location are trained, and students and test administrators are asked quality control questions to ensure that proper procedures are followed at each testing location. Test scorers are also systematically trained before the scoring to facilitate accurate and consistent scoring of student responses, and there are additional inter-rater reliability checks on scoring (OECD, 2003b). Before release, the test items and the achievement scores are analyzed using IRT approaches (see Froelich, Chap. 14) to develop estimates of student achievement with explicit error bands calculated and reported. To allow meaningful estimates of national, state, and district performance, sampling weights are calculated and included in the datasets to be used in generating population statistics from the sample data. Further, some large-scale programs are designed for longitudinal data collection and subsequent analyses investigating change over time; for example, studies such as the Berkeley and Oakland studies or the National Longitudinal Survey of Youth (Brooks-Gunn, Phelps, & Elder, 1991) and the National Longitudinal Survey of Children and Youth in Canada (Statistics Canada, n.d.).

There are disadvantages and limitations to SDA. Obtaining the data and preparing it for analysis can take more time than researchers may expect (Anderson, Monseur, & Cartwright, 2006; Rogers, Anderson, Klinger, & Dawber, 2006)—although this time is minimal compared to the time it would take to actually collect the data. More substantive challenges in SDA are the limitations imposed by the specifics of the sample of students and the variables operationalized and measured. The instruments and procedures used in collecting data in these large-scale programs are predetermined by design, which can create a problem for researchers who wish to examine a particular variable (Hyman, 1972). Researchers need to match the research question to the available data rather than collecting data that answers a research question. If the original operational definition of the variable used in the large-scale program differs from the current researcher's definition or if the sample of individuals is not the same to which the researcher wants to generalize, then the subsequent research may be unduly constrained. It is incumbent upon the researcher to fully understand the variable definitions and measures used in a large-scale assessment program before embarking on a SDA. Definitions of science literacy may differ between those used in the literacy and science education communities (Norris & Phillips, 2003; Yore, Pimm, & Tuan, 2007) and those used in large-scale assessment programs. Fortunately these descriptions are made available as part of the resources generated by most programs (OECD, 2002, 2003b).

13.2 Programme for International Student Assessment

One large-scale, internationally representative dataset that is of particular interest to literacy and science education researchers is the Programme for International Student Assessment; it was commissioned by the OECD in the late 1990s. The first assessment was conducted in 2000 with subsequent surveys coming every 3 years thereafter. In 2000, 43 countries participated; in 2003, 41 countries participated; and in 2006, 57 countries were involved (OECD, 2007) with a minimum of 150 schools sampled from each participating country and 5,250 students from each country (Turner & Adams, 2007). PISA is an age-based survey that assesses literacy of 15-year-old students. PISA uses the term *literacy* to encompass a broad range of competencies relevant to coping with adult life. The achievement domains targeted concurrently are the literacies associated with reading, mathematics, and science with some attention paid to problem solving. These competencies were based on relevance and applicability to adult life with no specific linkage to curricula of the participating countries—as is the case for the Trends in International Mathematics and Science Study (TIMSS & PIRLS International Study Center, n.d.). Along with the achievement estimates in each discipline-specific literacy (mathematics, science, reading, problem solving) for students in the OECD countries and other nations participating in these studies, information is collected on student attitudes and perceptions related to schooling, home background variables, and school information. Further information about the school is collected via a questionnaire administered to principals of each participating school. These datasets offer researchers the opportunity to investigate relationships amongst the correlates of learning and achievement and do so from an internationally comparative perspective.

PISA focuses on the knowledge and skills students have learned at school in the context of situations and challenges that call for application of that knowledge (Turner & Adams, 2007). It aims to measure whether young people at the end of compulsory schooling are well prepared to meet the challenges of contemporary life (McQueen & Mendelovits, 2003). As further noted by Turner and Adams (2007):

> PISA assess the extent to which students can use their reading skills to understand and interpret various kinds of written material that they are likely to meet as they negotiate their daily lives; the extent to which students can use their mathematical knowledge and skills to solve various kinds of mathematic-related challenges and problems; and their scientific knowledge and skills to understand, interpret and resolve various kinds of scientific situations and challenges. (p. 238)

In this way, PISA has not built its survey "in terms of mastery of the school curriculum, but in terms of important knowledge and skills needed in future life" (OECD, 2003a, p. 14), which is a critical difference to other international assessment programs during these times of content-specific reforms. Additionally, the PISA survey measures more than just literacy performance; both demographics and background information on students and schools are collected. The availability of this information at the PISA website (OECD, n.d.) and the test results can be combined to produce rich, connected datasets that facilitate extended analysis. General results are published by the OECD in a comprehensive international report that presents literacy outcomes within the major assessment domains (Turner & Adams). Additionally, differences among countries with respect to demographic variables are described with efforts to expand on the relationships between these variables and student cognitive outcomes.

PISA is administered in a 3-year cycle, each year measuring student performance in three major domains of literacy: reading, mathematics, and science; problem solving was added in 2003. However, each year one domain is the focus of the assessment in that the performance measure for that domain (e.g., science literacy in 2006) comprises two thirds of the items administered in the assessment program; and the focus of description and analysis is on that domain.

However, PISA is not without its detractors; and they come from a variety of perspectives. Goldstein (2004) found fundamental flaws in ensuring the *comparability* of meaning for test scores across diverse educational systems and cultures. He argued that multidimensionality of achievement measures should be considered in the statistical analysis of multilevel data such as PISA, which—given the scoring and analysis approach that is based upon a 1-parameter IRT model (for a full description of the scoring and analysis approach, see PISA 2003 Technical Report, OECD, 2003b, pp. 122–136)—confines development and analysis of the performance measure to a one-dimensional model. He suggested that PISA would be enhanced if it were designed as a longitudinal study as opposed to the current cross-sectional design; he believes that outcomes should not be used primarily as a ranking vehicle but more as a way to explore country differences in terms of culture, curricula, and school organization.

Dohn (2007) looked to the fundamental question of PISA 2003 (*How does initial schooling prepare students for participating in after-school life*?) and argued that, although the question is reasonable on a superficial level, assessing achievement on individual parts of mathematical literacy could fail to reflect a student’s more general mathematical ability. She raised concerns as to whether a single question, or even a series of questions, can capture the complexity of mathematical literacy. Dohn further concluded that PISA is actually a relatively reliable instrument for assessing a student’s knowledge and skills but the only real-life situation that PISA accurately measures is the PISA test situation itself. Prais (2003) questioned the value of PISA for country-specific educational policy since the achievement measures are not specifically designed for curricular outcomes but rather the more broadly based competencies of everyday life. However, investigating the concurrent performance outcomes from PISA in the domains assessed and relating these outcomes and measures to national curriculum could derive substantive policy development and implementation benefits. A further point Prais made was that the age of the students sampled excludes many students due to dropout, graduation, or other forms of attrition; therefore, the study was biased toward slower-maturing students.

Although there are these criticisms of PISA and other large-scale assessment programs, PISA's design and technical specifications meet the requirements of a good study, as laid out by Beaton, Postlewaite, Ross, Spearritt, and Worl (1999). To summarize the characteristics of PISA in relation to Beaton and colleagues' criteria:

- The aims of the PISA study were clearly presented as well as the rationale for going beyond the traditional curricula measures.
- The study is transparent in design and employed a variety of individuals and international organizations in a highly iterative process.
- The target population was identified for the purpose of collecting data on students at the end of formal education.
- Sampling procedures were clearly articulated and appropriate for representative national sampling.
- Item construction—although not linked to curriculum—went through an extensive selection process, and items were field trialed and further analyzed via various Rasch (a 1-parameter IRT analysis) fit statistics.
- Data were collected via a national project manager in each participating country and analyzed to allow the ability measures to be linked to the variety of background and demographic variables collected.
- Finally, the results were reported within each assessment domain emphasizing the profile of student responses in each country.

(Interested readers are encouraged to review the organization and administration of the PISA project as described by Turner and Adams (2007) for a more extensive discussion of the above points. Similarly, other PISA publications are open to all interested parties through the OECD website.)

PISA meets many of the Gold Standards of educational research, such as rigorous sampling design, well-developed objective measures of student achievement, and collection of data related to student and school traits. In addition, the data are hierarchical in structure, which is not uncommon for educational datasets. This means that multilevel modeling (e.g., HLM) is a suitable analytic approach.

13.3 Hierarchical Linear Modeling

Hierarchical linear modeling (Raudenbush & Bryk, 2002) is a regression-based analysis that explicitly incorporates into the analysis the hierarchical structure common to many educational datasets; in our case (PISA), students are nested within schools within countries. The data required for these analyses consist of both achievement (performance) and personal measures of students (level 1) and measures of school traits for each school (level 2) attended by the students.

At level-1, HLM allows us to describe the linear relationships of literacy achievement to student characteristics, such as gender, socioeconomic status (SES), student motivations, attitudes toward self, or attitudes toward school. This can be represented as the familiar linear regression equation, for example, modeling mathematics achievement (Math for student *i* in school *j*) with student gender, SES, and motivation:

$$\mathrm{Math}_{ij} = \beta_{0j} + \beta_{1j}\mathrm{Gender}_{i} + \beta_{2j}\mathrm{SES}_{i} + \beta_{3j}\mathrm{Motivation} + \mathrm{error}_{1j} \qquad (13.1)$$

Here each student's mathematics score is modeled as the intercept (β_{0j} – roughly similar to the mean mathematics score; in this case, for each of *j* schools) plus the weight (β_{1j}) associated with gender plus the weighted (β_{2j}) SES-level for that student plus the weighted (β_{3j}) motivation score plus individual error. However, unlike multiple regression, it must be noted that the weights are subscripted by *j*, signifying that a weight (e.g., the β_{1j} for gender) is calculated for each of the *j*-schools in the dataset. So if the weight for gender is 1.3 for a particular school and males are coded as 0 and females as 1, then on average females score 1.3 points more than males in that school.

HLM explicitly models variation in the gender relationships across schools and evaluates whether the variation is 0 or not. This can be done for every coefficient (the βs) in Eq. 13.1; a second set of regression equations is developed, which are termed the level-2 models. For example, in modeling the intercept (β_{0j} – which can be thought of as the conditioned school mean mathematics scores), not only is school variation in the intercept modeled (the error term – $error_{0j}$) but school-level traits, such as school size and an index of teacher morale, can be incorporated into the equation:

$$\mathrm{B}_{0j} = \gamma_{00} + \gamma_{01}\mathrm{School\ size}_{j} + \gamma_{02}\mathrm{Teacher\ morale}_{j} + \mathrm{error}_{0j} \qquad (13.2)$$

Here the school intercept is modeled with a level-2 intercept (γ_{00} – which is constant for all schools in the dataset) plus, in this example, a weighted (γ_{01}) measure of school size plus a weighted (γ_{02}) measure of teacher morale plus a school-level error term. This models the average school mathematics score as a function of the overall average mathematics score, school size, and teacher morale. HLM then tests the significance of the residual error to evaluate variation in school mean mathematics scores (the intercepts – β_{0j}s) once mathematics achievement has been conditioned (in this equation) on school size and teacher morale. If the error variance is significant, it can be interpreted to mean that there is still significant variation in the average school scores after conditioning on school size and teacher morale; whereas a nonsignificant error variance term suggests that once school size and teacher morale are accounted for, there is no significant variation in mean scores from one school to another.

Likewise, the gender, SES, and student motivation slopes or gradients in the student-level Eq. 13.1 can be modeled with school-level variables. This modeling of slopes is unique to multilevel modeling: modeling of relationships. For example, it may be that at the student level (level-1) SES is significantly and positively related to mathematics achievement (in our example, this would mean that γ_{20} in Eq. 13.3 is significant and positive). But it may be the case that there are substantial differences between schools for this relationship (the β_{2j} slopes in Eq. 13.1). One school may have a steep positive slope suggesting that student home background has a strong relationship to achievement, whereas another school may have a near zero SES slope suggesting that the school is more equitable in relation to student SES; it may also mean that there is little variation in student SES within that school.

HLM analysis explicitly estimates and evaluates these relationships for each school and in doing so provides the researcher with the opportunity to model the school slope variation with school traits. For example, if the SES slopes (β_{2j} in Eq. 13.1) vary significantly across schools, they can be modeled with school traits, such as measures of school academic focus or teacher morale:

$$B_{2j} = \gamma_{20} + \gamma_{21}\text{Academic focus}_j + \gamma_{22}\text{Teacher morale}_j + \text{error}_{2j} \qquad (13.3)$$

If student SES is highly related to mathematics achievement (a significant level-2 intercept: γ_{20}) and teacher morale has a negative relationship (γ_{22}) to this slope, it suggests that schools with higher levels of teacher morale (according to the perceptions of the school principal) will tend to be more equitable (lower SES slopes – the β_{2j} for that school) in terms of student SES. This finding would suggest that teacher morale moderates the relationship of student SES to achievement; in schools with high teacher morale, student SES is not as strongly related to student achievement as in schools with lower teacher morale. A policy implication could be that, if steps are taken to enhance teacher morale, SES equity could be positively influenced. Further, by explicitly modeling school-level error (the error_{2j} term in Eq. 13.3), we can evaluate—assuming we account for teacher morale and academic focus—any significant variation in the SES slopes remaining. If so, what other school traits could be influential in this relationship?

Another fundamental outcome of HLM analyses is the intraclass correlation coefficient generated by running an unconditioned model—the so-called null model (Raudenbush & Bryk, 2002). This statistic is an index of the proportion of variance in the outcome measure that can be accounted for by level-2 units. The results from PISA 2003 (Table 13.1) show that on average 35% of the variance in mathematics achievement can be attributed to schools. However, there is a broad range of values across countries—from 4% for schools in Iceland to over 60% for the schools of The Netherlands. The variation in intraclass correlations suggests structural differences in the ways school characteristics are related to student performance. Although both Iceland and The Netherlands are relatively high-performing countries (in the top 10% of national mean mathematics achievement), in Iceland and Finland (the top-performing country in PISA 2003) school differences account for almost no variation in student mathematics achievement. This is not the case in The Netherlands where the nature of the schools, which by design are structurally distinct with academic and vocational tracks, is more strongly related to student achievement. This example demonstrates how the measurement and modeling of school traits can lead to better understanding of educational performance as indexed by mathematics achievement.

Table 13.1 Intraclass correlation coefficients (ICC)—PISA 2003

Country	ICC	Country	ICC
ISL	0.042	THA	0.374
FIN	0.048	MEX	0.388
NOR	0.070	KOR	0.415
SWE	0.108	LIE	0.418
POL	0.127	TUN	0.426
DNK	0.132	URY	0.433
CAN	0.168	SVK	0.435
IRL	0.171	BRA	0.445
NZL	0.180	IDN	0.454
MAC	0.185	FRA	0.459
ESP	0.196	HKG	0.471
AUS	0.212	CZE	0.523
LVA	0.223	ITA	0.527
GBR	0.223	JPN	0.537
USA	0.263	AUT	0.553
RUS	0.307	TUR	0.560
LUX	0.317	BEL	0.562
CHE	0.334	DEU	0.581
PRT	0.341	HUN	0.586
GRC	0.363	NLD	0.626
YUG	0.364		
		Mean	**0.345**

13.4 Some Results from PISA 2003

The PISA 2003 assessment focused on mathematics literacy from the perspective that students can use the knowledge and skills they have learned and practiced at school when presented with situations in which that knowledge is relevant (Turner & Adams, 2007). Mathematical literacy for OECD/PISA is defined as:

> Mathematical literacy is an individual’s capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgments and to use and engage with mathematics in ways that meet the needs of that individual’s life as a constructive, concerned and reflective citizen. (OECD, 2003a, p. 24)

At the University of Victoria, we have been exploring the possibilities inherent in combining student background and demographic information with achievement data from this large sample of students across many countries. Five HLM studies with data derived from PISA 2003 have been conducted thus far in our laboratory; all examined the relationship between achievement and student characteristics (at level-1) and school characteristics (at level-2). All of the studies investigated relationships to mathematics achievement; and one study compared models not only across five countries but across the domains of mathematics, science, and reading. Although each study examined different combinations of these variables, there are common variables across several of the studies. We summarize these variables in Tables 13.2 (student-level characteristics) and 13.3 (school-level characteristics) and further summarize the findings for all of the countries examined in all studies in Table 13.4. Additionally, we describe the overall findings for each of the studies in the section that follows.

Goh (2006) examined the student- and school-level correlates of mathematics literacy for Canadian students with a particular focus on students' intrinsic motivations, perceptions of teacher support, and perceptions of student–teacher relations while controlling for gender and SES. The initial null model, which separates the variability of the outcome into within- and between-school components, yielded an intraclass correlation coefficient of 0.17—indicating that about 17% of the total variability in mathematics literacy could be attributed to schools (i.e., the group level); significant variation in school means across Canada was reported. Through a further series of HLM analyses, contextual student-level variables, demographic controls, and school-level variables were added to the models to allow for simultaneous analysis of student- and school-level factors.

The student-level variables used in the models were students' intrinsic motivation, perceptions of teacher support, and perceptions of student–teacher relations in the school (Table 13.2). The school-level variables used in the models were teacher morale, school autonomy, teachers' participation, and use of assessments (Table 13.3). Students' intrinsic motivation, perceptions of teacher support, and perceptions of student–teacher relations in the school were found to be significantly positive predictors of mathematics literacy among 15-year-old Canadian students. Together, they explained up to 17% of the student-level variance in mathematics literacy, with intrinsic motivation accounting for the majority of variance. These variables were next entered into the level-2 equations to model between-school variability in mathematics literacy while controlling for contextual student-level and demographic variables. These models explained between 30% and 34% of the estimated between-school variance. Overall, the results of Goh's (2006) study uncovered the significant and positive impact of intrinsic motivation, teacher support, and student–teacher relations in the school on students' mathematics literacy within Canadian schools.

Gu (2006) extended some of the questions in Goh's study and continued the focus on students' mathematics achievement in relation to their beliefs about themselves and the school environment—all within a hierarchical structure. However, she rationalized that an international perspective was possible and potentially meaningful because PISA is based both on international cooperation and comparison. She took this study one step further and compared the relevant HLM results between two countries involved in PISA: Canada and Hong Kong-China (note that Hong Kong is labeled as *Hong Kong-China* in PISA data collection and analyses). As she theorized that the contexts of different countries affected these relationships,

Table 13.2 Student variables from PISA 2003 used in the level-1 HLM models

Composite	Selected individual items
Mathematics self-concept	*I am just not good at mathematics* *I get good marks in mathematics* *I learn mathematics quickly*
Mathematics self-efficacy	(*How confident are you about…*) *Using a train timetable to work out how long it would take to get from one place to another* *Calculating how much cheaper a TV would be after a 30% discount* *Calculating how many square meters of tiles you need to cover a floor* *Understanding graphs presented in newspapers*
Teacher support	(*How often do these things happen in your mathematics lessons?*) *The teacher shows an interest in every student's learning* *The teacher gives extra help when students need it* *The teacher helps students with their learning*
Disciplinary climate	(*How often do these things happen in your mathematics lessons?*) *Students don't listen to what the teacher says* *There is noise and disorder* *Students cannot work well*
Intrinsic motivation	*I enjoy reading about mathematics* *I look forward to my mathematics lesson* *I do mathematics because I enjoy it* *I am interested in the things I learn in mathematics*
Student–teacher relations	*Students get along well with most teachers* *Most teachers are interested in students' well-being* *Most of my teachers really listen to what I have to say* *Most of my teachers treat me fairly*
Socioeconomic status [for PISA, the variable is called the Economic Social and Cultural Status]	*Parental occupation* *Parental education* *Home educational and cultural resources* *Family wealth*
Family structure	*Who usually lives at home with you?* *Mother? Father? Female guardian? Male guardian? Other?*
Immigration background	*In what country were*(*was*): *Your mother born?* *Your father born?* *You born?*
Instrumental motivation	*Mathematics is an important subject for me because I need it for what I want to study later on* *I will learn many things in mathematics that will help me get a job*
Performance orientation [labeled "preferred learning environment (competitive versus cooperative)" in PISA]	*I would like to be the best in my class at mathematics* *I try very hard in mathematics because I want to do better in the examinations than others* *I make a real effort in mathematics because I want to be one of the best* *In mathematics I always try to do better than the other students in my class*

Table 13.3 School variables from PISA 2003 used in the level-2 HLM models

Composite	Selected individual items
Student-related factors affecting the school climate (principals' perception)	(*In your school, to what extent is the learning of students hindered by*:) *Disruption of classes by students* *Students skipping classes* *Students lacking respect for teachers*
Teacher-related factors affecting the school climate (principals' perception)	(*In your school, to what extent is the learning of students hindered by*:) *Teachers' low expectations of students* *Teacher absenteeism* *Poor student-teacher relations* [Note: reversed scoring]
Teacher morale and commitment (principals' perception)	*The morale of teachers in this school is high* *Teachers work with enthusiasm* *Teachers take pride in this school*
Students' morale and commitment (principals' perception)	*Students enjoy being in school* *Students work with enthusiasm* *Students value academic achievement* *Students do their best to learn as much as possible*
School autonomy and teacher participation (principals' perception)	(*In your school, who has the main responsibility for*:) *Selecting teaches for hire* *Firing teachers* *Formulating school budget* *Setting teachers' starting salaries* *Deciding which courses are offered*
Use of assessments (principals' perception)	(*Generally in your school, how often are students assessed using*:) *Standardized tests* *Teacher-developed tests* *Teacher judgmental ratings* *Student portfolios* *Student assignments and projects*
School socioeconomic status	Derived from the aggregate average for participating students in the school
Proportion of girls	Principal-supplied value
Proportion of non nuclear families	Principal-supplied value
Proportion of non native students	Derived from the aggregate average of numbers of participating students not born in the country

she explored relationships between students' beliefs about themselves in mathematics and their mathematics achievements in both countries. Specifically, she sought to examine the student- and school-level correlates of mathematics literacy for Canadian and Hong Kong-China students.

The student-level variables used in the models were mathematics self-concept, mathematics self-efficacy, teacher support, and disciplinary climate (Table 13.2).

Table 13.4 Descriptives for countries used in the HLM models (all values unweighted)

Country	Students in sample (N)	Schools in sample (N)	Mean mathematics score	Intraclass correlation for general mathematics score
Canada	27,953	1,066	532	0.17
HongKong-China	4,478	145	550	0.47
United States	5,456	262	483	0.26
United Kingdom	9,535	361	*	0.22
Japan	4,707	144	534	0.54
Korea	5,444	149	542	0.41

*The school-level response rate was not high enough to allow for a comparison, so no score is given.

The school-level variables that were used in the level-2 models were student behaviors related to school climate, teacher behaviors related to school climate, student morale, and teacher morale (Table 13.3). Preliminary analyses showed a marked difference between the two countries in terms of the variance in mathematics performance that could be attributed to schools, which was estimated by the intraclass correlation coefficient (Raudenbush & Bryk, 2002). The intraclass coefficient was 0.17 for Canada and 0.47 for Hong Kong-China, meaning that 17% of variance in mathematics performances can be attributed to Canadian schools whereas for Hong Kong almost half of the variance in mathematics performance can be attributed to schools. Through a further series of HLMs, contextual student-level variables, demographic controls, and school-level variables were added to the models to allow for simultaneous analysis of student- and school-level factors and subsequent comparison between the two countries.

In the Canadian student-level model, students who had higher mathematics self-efficacy and mathematics self-concept were predicted to perform better in mathematics. School-level student morale and disciplinary climate were positively associated with average school scores. Surprisingly, teacher support was negatively related to average school scores—meaning that as the amount of teacher support reported by students increased there was a decrease in predicted mathematics scores. The Hong Kong-China students who had higher mathematics self-efficacy and mathematics self-concept were predicted to perform better in mathematics. Students' morale, behaviors, and disciplinary climate were positively associated with average school scores while teacher support was again negatively related to average school scores. Comparisons between the two countries based upon these models uncovered other similarities: the relationship between achievement and mathematics self-concept and mathematics self-efficacy, students' morale and the disciplinary climate within schools were both significantly and positively related to mathematics achievement, and teacher support was significantly and negatively related to school average scores.

However, differences also existed between the two countries in the final models. Hong Kong-China had a higher overall average score than Canada by about 20 points. Furthermore, self-efficacy and self-concept slopes in Canada were steeper than Hong Kong-China, meaning that self-efficacy and self-concept have a stronger relationship to achievement in Canada than in Hong Kong-China. However, student morale averaged for each school significantly reduced the effect of self-efficacy on mathematics achievement in Hong Kong-China but not in Canada—meaning that for schools with higher average student morale, the relationship of self-efficacy to achievement was reduced. Teachers' morale significantly enhanced the self-concept slope in Hong Kong-China but not in Canada.

The primary aim of Gu's (2006) study was to examine the relationships among students' beliefs about themselves in mathematics, learning environment at school, and mathematics achievement (at student and school levels) in Canada and Hong Kong-China. At the student level, mathematics self-concept and mathematics self-efficacy were identified as being significantly and positively related to mathematics achievement for both countries. Similarly, at the school level and across both countries, students' morale, students' behaviors, and the disciplinary climate were positively associated with average school scores while teacher support was negatively related to average school scores in mathematics. However, it was uncovered that Canada has stronger relationships between students' self-beliefs (concept and efficacy) in mathematics and their mathematics performance than Hong Kong-China. But school learning environment—as measured by teacher support, disciplinary climate, students' and teachers' behaviors, and students' and teachers' morale—had more effect on school mathematics achievement in Hong Kong-China than in Canada.

Hsu (2007) also compared Canada and Hong Kong-China but investigated the student- and school-level variables of mathematics achievement for 15-year-old students focused on student demographic characteristics and student gender. She rationalized that since both countries had similar backgrounds—high overall mathematics achievement, a large immigrant population, and former British colonies—they would represent meaningful comparisons of Western and Eastern cultures. She added student- and school-level variables to the models to allow for simultaneous analysis of student- and school-level factors and subsequent comparison between the two countries. The student-level variables used were SES, gender, family structure, and immigration background. The school-level variables used were school SES, proportion of girls, proportion of nonnuclear families, and proportion of nonnative students.

It should be noted that SES was measured in PISA by means of a composite variable developed from the educational and vocational levels of parents and an index of cultural possessions. Nonnuclear families were defined as those families not consisting of both a mother and a father. It should also be noted that the variable *native/nonnative* refers specifically to whether the student was born in the country or not; it does not refer to aboriginal or indigenous status, as used within some countries.

Findings pointed to student SES as being significantly and positively associated with mathematics achievement at the student level in both countries. The positive impact of SES on mathematics achievement was smaller in Hong Kong-China than in Canada (SES explains 1% versus 11% of the within-school variance in mathematics achievement, respectively). Nonnative students were predicted to have significantly lower mathematics performance than native students in both countries; however, in both countries, first-generation students (those students born in-country to parents who immigrated to the country) outperformed the other two groups (native and nonnative). At the school level, having a higher proportion of girls in the school predicted a significant increase in the school average mathematics achievement only in Canada; but the proportion of students from nonnuclear families in a school predicted a significantly negative impact on school average mathematics achievement in both Canada and Hong Kong-China. Finally, having a higher proportion of nonnative students within a school was associated with significantly lower overall school average mathematics achievement in Hong Kong-China whereas it did not significantly influence Canadian overall school average mathematics achievement.

Ram (2007) investigated the effects of student- and school-level variables on the mathematics achievement of 15-year-old students in Canada and Japan. Specifically, she sought answers to the following questions: (a) Do gender differences exist in mathematics achievement? (b) Is there a relationship between student-perceived teacher support and mathematics achievement? (c) Are there significant relationships between mathematics achievement and SES? (d) Are there significant relationships between mathematics achievement and principals' perceptions of teachers' behavior related to school climate? The intraclass correlation of 0.17 for Canada and 0.54 for Japan denoted that schools accounted for over half the variance in mathematics performance of Japanese students whereas schools accounted for less than 20% of mathematics performance variance of Canadian students. The final HLMs, which included all significant student- and school-level variables, differed mainly in the level of student-perceived teacher support: mathematics achievement scores were significantly related to teacher support in Canada but not in Japan.

Ram (2007) found that in Canada males outperformed females in mathematics achievement. Students' economic, social, and cultural status (ESCS)—the SES-variable within the PISA framework—was found to be a significant predictor of average mathematics. Students who reported having a higher SES did better on the mathematics assessment than those who reported having a lower SES. In Japan, similar results were found: males outperformed females on mathematics achievement; and students with a high ESCS achieved higher results when compared to students with a lower SES. However, the influence of SES on mathematics achievement in Canada was found to be almost five times larger than for Japan. Unlike Canada, Japan did not show any significant relationship between the levels of student-perceived teacher support and mathematics achievement. Interestingly, however, the relationship between mathematics achievement and students' perceptions of teacher support in Japan was shown to significantly vary from school to school.

In the most comprehensive analysis to come out of this research program to date, Ross (2008) examined the relationship between student achievement in mathematics, science, reading, and problem solving and the following student variables: instrumental motivation, intrinsic motivation, self-efficacy, and performance orientation. The level-2 variables examined were students' perceptions of teacher support, principals' perceptions of teacher factors to school climate, and student morale and commitment. The findings were compared amongst six countries: Canada, the United States, the United Kingdom, Hong Kong-China, Japan, and Korea. The proportion of variance in mathematics performance that was accounted for by schools (i.e., intraclass correlations) differed substantially across the countries studied: 0.17 for Canada, 0.22 for the United Kingdom, 0.26 for the United States, 0.47 for Hong Kong-China, 0.54 for Japan, and 0.42 for Korea.

The final HLMs for all countries, which included all significant student- and school-level variables, indicated at level-1 that increased mathematics self-efficacy scores predicted higher achievement scores in all countries and across the three domains of mathematics, science, and reading. With a few exceptions (i.e., mathematics scores in Japan and all domains in Korea), increased performance orientation predicted a decrease in average scores across all literacy domains—meaning that as students reported level of performance orientation (sample item: *I would like to be the best in my class in mathematics*) increases, there was a related decrease in mathematics scores. At level-2, principals' perception of student morale was significant for all countries across all domains.

For the mathematics domain at level-1, intrinsic motivation was a significant predictor of increased score for all but two countries: for the United Kingdom, intrinsic motivation predicted a decrease in scores, and intrinsic motivation was not significant in the US HLM model. For Canada only, instrumental motivation was a significant positive predictor. At level-2, only principals' perception of student morale appeared in all models for all countries, with increases in perceived student morale associated with increased mathematics performance. For Japan, teacher support at the school level (as an aggregate of level-1 student perception) was significant and positive—meaning that as the average level of perceived teacher support increased in the school, there was an increase in average mathematics score for the school. For the United Kingdom, principals' perception of teacher factors related to school climate was significant and positive—meaning that as the perceived teacher contribution to school climate increased, there was a related increase in mathematics performance.

For the reading literacy domain, intrinsic motivation was significant for all Western countries where an increase in intrinsic motivation was associated with a decrease in reading score. Intrinsic motivation was not significant in any of the HLM models for the Eastern countries; only performance orientation and mathematics self-efficacy appeared at level-1 for those countries' models for reading literacy. For Canada and the United States, instrumental motivation was significant and predicted an increase in reading scores. At level-2, only principals' perception of student morale appeared in all models for all countries. For Canada and the United Kingdom, principals' perception of teacher factors related to school climate was significant and positive—meaning that as the perceived teacher contribution to school climate increased there was a related increase in reading performance. For Japan only, teacher support (as an aggregate of level-1 student perception) was significant and positive—meaning that as the average level of student perceived teacher support increased there was an associated increase in school mean reading performance.

For the science literacy domain, intrinsic motivation was significant for the United Kingdom and the United States (where it predicted a decrease in science scores) and Hong Kong-China (where it predicted an increase in science scores). For Canada and the United States, instrumental motivation was significant and predicted an increase in science scores. At level-2, only principals' perception of student morale appeared significant in all models for all countries. For Canada and the United Kingdom, principals' perception of teacher factors related to school climate was significantly related to increase in school-level achievement. For Japan only, teacher support (as an aggregate of level-1 student perception) was significant and increased the slope of the intercept.

Finally, for the problem-solving domain, intrinsic motivation was significant for Japan, Hong Kong-China, and Korea where an increase in problem-solving scores was predicted. For Canada and the United States, instrumental motivation was significant, and an increase in problem-solving scores was predicted. At level-2, only principals' perception of student morale appeared in all models for all countries. For the United Kingdom, principals' perception of teacher behaviors related to school climate was significant, which increased the slope of the intercept. For Japan only, teacher support (as an aggregate of level-1 student perception) was significant, which increased the slope of the intercept.

The results of these HLM studies of PISA 2003 data (Table 13.5) should be interpreted with care and caution. The use of large international databases has some inherent difficulties. Although the sample size is large and powerful, the questionnaire items may be too general to uncover important underlying constructs that exist in authentic educational contexts (i.e., the potential for inappropriate aggregation of data). Additionally, secondary data analysis of large national datasets does not allow causal interpretation of the school effects. That is, the HLM analyses on these students are correlational and nonexperimental. Finally, the school-level variables were derived entirely from questionnaires completed by school principals; this presents several concerns when interpreting these data. Generalization from a single source of information for each school may fail to completely capture the multidimensional nature of the factors. In this case, principals may not be the best source of information about teachers' morale and commitment; bias is possible; and teachers may keep their true beliefs, feelings, and actions hidden from their school. One issue with secondary data analysis is that the nature of variables is determined by the program designers and administrators and fixed into the dataset. This may lead to problems of relating findings to theoretic underpinnings of the research and necessitates that the researcher has a firm, comprehensive understanding of the operationalization of constructs measured and the nature of the resulting variables.

Table 13.5 Comparison of HLM theses

	Research questions	Intraclass correlation	Results
Goh (2006)	Do Canadian students' intrinsic motivations, perceptions of teacher support, and perceptions of student–teacher relations—while controlling for gender and socioeconomic status—influence mathematics literacy?	0.17, indicating that about 17% of the total variability in mathematics literacy could be attributed to schools	Students' intrinsic motivation, perceptions of teacher support in the mathematics classrooms, and perceptions of student–teacher relations in the school are significant positive predictors of mathematic literacy The models explain between 30% to 34% of the estimated between-school variance School variables such as teacher morale and commitment, school autonomy, teachers' participation, and frequency of assessment are statistically nonsignificant in all three models Significant school variation persists after the variables have been added Schools with higher mean student–teacher relations and higher mean intrinsic motivation are more equitable for students
Gu (2007)	Do student- and school-level correlates of mathematics literacy with a particular focus on students' mathematics self-concept, mathematics self-efficacy, teacher support, disciplinary climate, student behaviors, teacher behaviors, student morale, and teacher morale differ for Canadian and Hong Kong-China students?	0.17 for Canada 0.47 for Hong Kong-China	Mathematics self-efficacy and mathematics self-concept are significant positive correlates influencing students' performance for both Canada and Hong Kong-China School learning environment has more effect on school mathematics achievement in Hong Kong-China than in Canada Significant school variation persists after the variables have been added Effects of school characteristics on students' mathematics achievement in Canada are more homogeneous than in Hong Kong-China
Hsu (2007)	Do student- and school-level variables of mathematics achievement for 15-year-old students focused on student characteristics (socioeconomic status, family structure, immigrant status, and gender) differ across Canada and Hong Kong-China for mathematics achievement?	0.17 for Canada 0.47 for Hong Kong-China	Higher student socioeconomic status predicts higher individual mathematics literacy only in Canada Student gender (being female) and immigrant status (being non native) were significant negative predictors of student mathematics literacy in both countries As school mean socioeconomic status increased, school average mathematics scores increased in Canada and Hong Kong-China As school proportion of non nuclear families increased, school average mathematics scores decreased in Canada and Hong Kong-China

(continued)

Table 13.5 (continued)

	Research questions	Intraclass correlation	Results
			As school proportion of girls increased, school average mathematics scores increased in Canada As school proportion of non native students increased, school average mathematics scores decreased in Hong Kong-China
Ram (2007)	Do gender differences, student-perceived teacher support, socioeconomic status, and principals' perceptions of the school climate exist for mathematics achievement? What are the differences in findings between Canada and Japan?	0.17 for Canada 0.54 for Japan	Significant positive predictors of mathematics literacy were student gender (male advantage) in Canada and Japan, teacher support (in Canada but not in Japan), socioeconomic status (more influential in Canada than Japan) School variables such as principals' perception of teacher morale and commitment, student morale and commitment, teacher-related factors affecting school climate, and student-related factors affecting school climate were all significant predictors of mathematics literacy Effects of school characteristics on students' mathematics achievement in Canada are more homogeneous than in Japan
Ross (2008)	What are the relationships between achievement in mathematics, science, reading, and problem-solving and student instrumental motivation, intrinsic motivation, self-efficacy, and performance orientation, and students' and principals' perceptions of teacher support, school climate, and student morale and commitment? Subsequent findings were compared among Canada, the United States, the United Kingdom, Hong Kong-China, Japan, and Korea.	0.17 for Canada 0.22 for the United Kingdom 0.26 for the United States 0.47 for Hong Kong-China 0.54 for Japan 0.42 for Korea	For all countries, mathematics self-efficacy score predicted higher achievement scores for all domains Instrumental motivation was a significant predictor only for Canada and the United States For all countries and domains where performance orientation was significant, it predicted lower achievement scores For all countries and across all domains, principals' perception of student morale was significant For Canada and the United Kingdom, principals' perception of teacher behaviors was significant for some domains In Japan only, aggregate student reports of teacher support were significant for all domains

13.5 Conclusion

From the five studies reviewed above and those in other more general reviews (e.g., Anderson, Lin, Treagust, Ross, & Yore, 2007), the potential for meaningful, policy-relevant research is clear. For example, the impact of student background characteristics is consistently significant and positive, but the magnitude of the relationship varies from one country to another, and within countries it can vary significantly from one school to another (Hsu, 2007). This means that a simplistic view of the impact of SES on achievement is not a universal given but rather should be carefully evaluated for each specific educational context. Should significant school variation be identified, it would be a useful policy to redirect analyses to the identification of those schools that are most equitable (the flattest SES gradients) and attempt to determine if there are any systematic, reproducible school traits associated with these flatter slopes. The data required for such analysis may very well lie within the datasets that generated the initial finding of between-school variation.

Another field of findings associated with HLM centers on the influences of school traits on student-level correlates of learning. For example, both Goh (2006) and Ross (2008) found that student motivation is positively related to achievement, which is a finding consistent with general expectations and with the research literature. However, it was shown that this relationship can be enhanced by the general level of student–teacher relationships within the school. So although student motivation can have a positive influence on achievement, the relationship varies from one school to another; and this variance is dependent to an extent upon an element of school climate, as perceived by the school principal.

Multilevel modeling reflects the complexities of social organizations in which systems are nested within larger organizational units (e.g., schools within districts within states/provinces). Traditionally, social systems have been represented with mechanistic analogies. This modeling has been premised on the instrumental expectation that the results will be useful for monitoring schools, evaluating programs, formulating policy, and implementing school change. But, as Lindblom (1968, 1992) has pointed out time and again, the desire that models of complex social systems such as public education have an instrumental use remains an elusive dream. Models of complex social systems are likely to be, at best, enlightening—allowing incrementally expanding understandings of complex and dynamic systems such as public schools (Kennedy, 1999). It may be more productive and understandable to model educational systems more biologically to capture the dynamic, nonrule-driven ecosystems involving students, parents, teachers, and administrators. HLM offers means to describe and eventually better understand not only the complexities but the schools themselves. Rather than mask the complexities by relying solely on overly simplistic descriptors such as mean scores, multilevel modeling can (a) augment the results with descriptions of patterns of relationships between student-level and school-level correlates and (b) estimate the extent of school-to-school variation in these relationships. This would allow the policy maker to judge whether broad-scale or context-specific approaches to school and system policy and practice are the most reasonable routes to pursue.

The potential for multilevel analysis within the context of secondary data analysis of large-scale assessment programs is in the early stages, but the promise is substantial. Findings generated from these models can lead to better and more nuanced understandings of school performance and can also be used to guide data targeting in future rounds of assessment programs such as PISA or large-scale, systematic, experimental studies. The approach engenders an evidence-based focus for both the research and policy communities. Multilevel modeling in the social sciences is and will continue to be a meaningful approach to the analysis of evidence for research on educational performance since so much of what we study is multilevel. Failure to recognize the multivariate and hierarchical structure of our datasets and use of appropriate analytic approaches can lead to serious problems (Luke, 2004). The work with datasets generated from PISA will continue with results from the 2006 cycle, which had a focus on science literacy. A new element of PISA 2006 is the development of items to capture student attitudes toward science, and these items are contextualized with the performance items in the assessment (OECD, 2006). This expansion of information collected should allow for further enhancement of our models and lead to better understandings of student literacy performance.

Acknowledgments The funding for the work at the University of Victoria has come from the National Science and Engineering Research Council of Canada, which supports the *Centre for Research in Youth, Science Teaching & Learning* (*CRYSTAL–Pacific*) Project, and from the Canadian Educational Statistics Council and the Social Sciences and the Humanities Research Council, which supports the *Correlates of Learning Outcomes* (*COLO*) Project.

References

Anderson, J. O., Lin, H.-L., Treagust, D. F., Ross, S. P., & Yore, L. D. (2007). Using large-scale assessment datasets for research in science and mathematics education: Programme for International Student Assessment (PISA). *International Journal of Science & Mathematics Education*, *5*(4), 591–614.

Anderson, J. O., Monseur, C., & Cartwright, F. (2006, April). *Procedures and issues associated with the data and analysis of PISA 2003 thematic research*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Beaton, A. E., Postlewaite, T. N., Ross, K. N., Spearritt, D., & Worl, R. M. (1999). *The benefits and limitations of international educational achievements studies*. Retrieved April 22, 2008, from http://unesdoc.unesco.org/images/0011/001176/117629e.pdf

Brooks-Gunn, J., Phelps, E., & Elder, G. H. (1991). Studying lives through time: Secondary data analyses in developmental psychology. *Developmental Psychology*, *27*(6), 899–910.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications and programming*. Mahwah, NJ: Lawrence Erlbaum.

Campbell Collaboration (n.d.). *Homepage*. Retrieved October 3, 2007, from http://www.campbellcollaboration.org

Dohn, N. B. (2007). Knowledge and skills for PISA – Assessing the assessment. *Journal of Philosophy of Education*, *41*(1), 1–16.

Evidence for Policy and Practice Information and Co-ordinating Centre. (n.d.). *Homepage*. Retrieved June 20, 2008, from http://eppi.ioe.ac.uk/cms/

George, R., & Kaplan, D. (1998). A structural model of parent and teacher influences on science attitudes of eighth graders: Evidence from NELS: 88. *Science Education*, *82*(1), 93–109.

Goh, M. (2006). *A multilevel analysis of mathematics literacy: The effects of intrinsic motivation, teacher support, and student-teacher relations*. Unpublished master's thesis, University of Victoria, Victoria, British Columbia, Canada.

Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 319–330.

Gu, Z. (2006). *Students' beliefs about themselves, learning environment at school, and achievement*. Unpublished master's thesis, University of Victoria, Victoria, British Columbia, Canada.

Hakim, C. (1982). *Secondary analysis in social research*. Boston: Allen & Unwin.

Hofferth, S. L. (2005). Secondary data analysis in family research. *Journal of Marriage & Family*, *67*(4), 891–907.

Howe, K. R., & Eisenhart, M. (1990). Standards for qualitative (and quantitative) research: A prolegomenon. *Educational Researcher*, *19*(4), 2–9.

Hsu, J. C. (2007). *Comparing the relationships between mathematics achievement and student characteristics in Canada and Hong Kong through HLM*. Unpublished master's thesis, University of Victoria, Victoria, British Columbia, Canada.

Husen, T. (1988). Research paradigms in education. In J. P. Keeves (Ed.), *Educational research, methodology and measurement: An international handbook* (pp. 17–20). Toronto, Ontario, Canada: Pergamon Press.

Hyman, H. (1972). *Secondary analysis of sample surveys: Principles, procedures, and potentialities*. Toronto, Ontario, Canada: Wiley & Sons.

Kennedy, M. M. (1999). Infusing educational decision making with research. In G. J. Cizek (Ed.), *Handbook of educational policy* (pp. 54–80). San Diego, CA: Academic Press.

Kiecolt, K. J., & Nathan, L. E. (1985). *Secondary analysis of survey data*. Newbury Park, CA: Sage.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford.

Lather, P. (1992). Critical frames in educational research: Feminist and poststructural perspectives. *Theory into Practice*, *31*(2), 87–99.

Lindblom, C. E. (1968). *The policy-making process*. Englewood Cliffs, NJ: Prentice Hall.

Lindblom, C. E. (1992). *Inquiry and change: The troubled attempt to understand and shape society*. New Haven, CT & New York: Yale University Press & Russell Sage Foundation.

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage.

Marsh, H. W., & Yeung, A. S. (1998). Longitudinal structural equation models of academic self-concept and achievement: Gender differences in the development of math and English constructs. *American Educational Research Journal*, *35*(4), 705–738.

McQueen, J., & Mendelovits, J. (2003). PISA reading: Cultural equivalence in a cross-cultural study. *Language Testing*, *20*(2), 208–224.

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, *87*(2), 224–240.

Organisation for Economic Co-operation and Development. (2001). *Knowledge and skills for life: First report from the OECD programme for international student assessment*. Paris: Author.

Organisation for Economic Co-operation and Development. (2002). *PISA 2000 technical report*. Paris: Author.

Organisation for Economic Co-operation and Development. (2003a). *The PISA 2003 assessment framework – mathematics, reading, science and problem solving: Knowledge and skills*. Paris: Author.

Organisation for Economic Co-operation and Development. (2003b). *PISA 2003 technical report*. Paris: Author.

Organisation for Economic Co-operation and Development. (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006*. Paris: Author.

Organisation for Economic Co-operation and Development. (2007). *What PISA assesses*. Paris: Author.

Organisation for Economic Co-operation and Development. (n.d.). *PISA Homepage*. Retrieved June 30, 2008, from http://www.pisa.oecd.org/pages/0,2987,en_32252351_32235731_1_1_1_1_1,00.html

Phillips, D. C. (2005). The contested nature of empirical educational research (and why philosophy of education offers little help). *Journal of Philosophy of Education*, *39*(4), 577–597.

Prais, S. J. (2003). Cautions on OECD's recent educational survey (PISA). *Oxford Review of Education*, *29*(2), 139–163.

Pring, R. (2000). The 'false dualism' of educational research. *Journal of Philosophy of Education*, *34*(2), 247–260.

Ram, A. (2007). *A multilevel analysis of mathematics literacy in Canada and Japan: The effects of sex differences, teacher support, and the school learning environment*. Unpublished master's thesis, University of Victoria, Victoria, British Columbia, Canada.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis* (2nd edn.). Thousand Oaks, CA: Sage.

Reynolds, A. J., & Walberg, H. J. (1991). A structural model of science achievement. *Journal of Educational Psychology*, *83*(1), 97–107.

Rogers, W. T., Anderson, J. O., Klinger, D. A., & Dawber, T. (2006). Pitfalls and potential pitfalls of secondary data analysis of the Council of Ministers of Education, Canada, national assessment. *Canadian Journal of Education*, *29*(3), 757–770.

Ross, S. P. (2008). *Motivation correlates of academic achievement: Exploring how motivation influences academic achievement in the PISA 2003 dataset*. Unpublished doctoral dissertation, University of Victoria, Victoria, British Columbia, Canada.

Statistics Canada. (n.d.). *National longitudinal survey of children and youth for ages 16–17*. Retrieved April 22, 2008, from http://www.statcan.ca/english/kits/microdata/microdata.htm

TIMSS & PIRLS International Study Center. (n.d.). *Homepage*. Retrieved July 11, 2008, from http://timss.bc.edu/

Turner, R., & Adams, R. J. (2007). The Programme for International Student Assessment: An overview. *Journal of Applied Measurement*, *8*(3), 237–248.

United States Institute of Education Sciences. (n.d.). *What Works Clearinghouse: Homepage*. Retrieved June 20, 2008, from http://ies.ed.gov/ncee/wwc/

Yore, L. D., Craig, M. T., & Maguire, T. O. (1998). Index of science reading awareness: An interactive-constructive model, test verification, and grades 4–8 results. *Journal of Research in Science Teaching*, *35*(1), 27–51.

Yore, L. D., Pimm, D., & Tuan, H.-L. (2007). The literacy component of mathematical and scientific literacy. *International Journal of Science & Mathematics Education*, *5*(4), 559–589.

Chapter 14
Methods from Item Response Theory: Going Beyond Traditional Validity and Reliability in Standardizing Assessments

Amy G. Froelich

In determining the effectiveness of educational interventions, the Gold Standard requires the use of tests and assessments of proven validity. Messick (1989) defined validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores” (p. 13). Education researchers wishing to evaluate the effectiveness of educational interventions and programs under the Gold Standard must either develop and validate their own tests and assessments or use ones developed and validated by others. As a result of the No Child Left Behind federal legislative mandate for Grades K-12 in the United States (NCLB, 2002), educational research through intervention programs that improve student learning in mathematics, reading, and science education in Grades K-12 have one natural test of interest: the standardized examination used in the state for determining student proficiency status and school and district proficiency rates. Local school personnel and state education professionals are particularly interested in research showing improvements in student performance on these high-stakes tests. Other standardized assessments that can be used to show the effectiveness of an educational program or intervention are the National Assessment of Educational Progress (NAEP, US National Center for Education Statistics, n.d.), the ACT®, (ACT, n.d.), and the SAT® (College Board, n.d.).

However, the use of state NCLB tests and these other assessments is precluded in many situations. For example, the educational program or intervention may be targeted at a subject area not covered by these assessments, such as history or study in a foreign language. Even if the subject area is in mathematics, reading, or science, the goal of the intervention may not align with the underlying curriculum and goals of the NCLB tests in the subject area. For example, programs focusing on the development of problem-solving skills in mathematics may have different goals than the curriculum tested on the NAEP or the NCLB state assessment. These assessments would not be good measures of the effectiveness of this type of intervention program.

A.G. Froelich
Iowa State University

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education,*
© Springer Science + Business Media B.V. 2009

In the situation where educational researchers need separate assessments from these large-scale standardized tests, the issue of validity looms large. In establishing the validity of an assessment, the focus is on gathering different types of supporting evidence (Kane, 2006). First, a substantive analysis of the assessment should be conducted. This analysis is based on the important aspects of the subject under study and the development of the theory of learning of this subject. Content area specialists play an important role in the substantive analysis of an assessment by ensuring alignment with the specifications of the subject and the goals of the educational program. This type of analysis is generally referred to in the literature as *content validity*. Second, establishing the validity of an assessment should also be based on considerations external to the assessment. These external considerations include alignment with other standardized assessments measuring similar constructs or lack of alignment with standardized assessments measuring different constructs, or both. This type of analysis is generally referred to in the literature as *criterion validity*.

This chapter deals with the third aspect of establishing the validity of an assessment—an internal study of the properties of the assessment in the population of interest. This type of analysis is generally referred to in the literature as *construct validity*. Construct validity is concerned with an analysis of a test and its items or questions. What construct does the test actually measure overall? Does each item on the test measure the same construct, or do certain items appear to be measuring different constructs in the population of interest? What are the properties of these items (i.e., how difficult are the items and how well do they discriminate among test takers)? How much error of measurement is present in the test? In essence, construct validity seeks to answer the questions: What is the test measuring? How does it work? (Borsboom, 2006).

14.1 Construct Validity and Reliability

One typically used means of studying the above questions is the calculation of the internal consistency or internal reliability of an assessment based on Cronbach's alpha statistic (Cronbach, 1951). The reliability of a measurement is defined as the correlation of values from repeated measurements using the same instrument over a fixed and short amount of time (Traub, 1994). Reliability of measurements is a concern in many other fields outside of education; it plays a particularly important role in many engineering applications. However, in educational testing, people learn from and have memories of their experiences, making it impossible to produce a true estimate of the reliability of an educational assessment based on this definition. Methods have thus been developed (including Cronbach's alpha statistic) to estimate the reliability of an assessment given the limitations present in educational testing. A good overview of these methods can be found in Traub and in Haertel (2006).

The problem with the focus on reliability as a measure of the internal structure of an assessment is that it provides a possible answer to only one main question:

How much error of measurement is present in the test? An analysis of the reliability of an assessment does not consider or estimate statistically the overall construct being measured. Reliability analyses cannot provide answers to the questions at the item level, such as: Does each item on the test measure the same construct, or do certain items appear to be measuring different constructs in the population of interest? The analysis of the reliability of an assessment allows the researcher to determine an estimate for the correlation between scores on the assessment of a single individual if he or she were given the same assessment twice.

14.2 Inferences and Actions Based on Assessments

Almost all educational tests have a common goal: the determination of some level of ability or achievement of the individuals taking the test. Typically, this ability level is determined by a total test score or, in the case of multiple-choice examinations, by the number correct score. We expect test takers with higher scores on the test to have more ability or achievement on the constructs being measured on the test than test takers with lower scores.

Two assumptions are implicit in the use of the total test score as a measure of ability or achievement: the total test score provides a stochastic ordering of the test takers on ability or achievement, and the standard error of measurement of the test is small. Stochastic ordering means that individuals receiving a higher test score will have *on average* a higher ability than individuals receiving a lower test score. The assumption of the stochastic ordering of examinees by test performance is, therefore, a probabilistic one: the ordering of ability through the total test score could be incorrect for individual test takers. An individual with a higher-level of ability could receive a lower test score than an individual with a lower-level of ability. To help control for error in the ordering of individual test takers on ability, the assessment should also have a small standard error of measurement. One way to estimate the standard error of measurement is by using an estimate of the reliability of the assessment, such as Cronbach's alpha (Traub, 1994). This estimate assumes a constant standard error of measurement for all test takers.

Clearly, stochastic ordering of the test takers by total test score and the standard error of measurement of the assessment are important considerations to establishing the validity of an assessment. Verification of these two assumptions provides "empirical evidence [to] support the adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1989, p. 13) and thus should be a part of establishing the validity of an assessment. The problem then becomes showing empirically that the assumption of stochastic ordering is reasonable given the available data collected on the assessment and determining the standard error of measurement of the ability estimate obtained for each test taker. In the rest of this chapter, methods from Item Response Theory (IRT, Lord & Novick, 1968) are introduced. These methods will be used to examine the assumption of the stochastic ordering of the ability of test takers with test scores and to estimate the standard error of measurement across the range of abilities.

14.3 Introduction to Item Response Theory

In the terminology of IRT, test questions are referred to as test items and people completing the test are referred to as examinees. Let U_i be the response of a randomly sampled examinee on the *i*th item of a test, and let $\mathbf{U}^{(n)} = (U_1, U_2,\ldots,U_n)^T$ denote the response pattern of a randomly sampled examinee to an *n* item test. Although it is possible to study other types of items, most standardized assessments use multiple-choice questions with just one correct answer. This chapter will focus on these dichotomous items, so $U_i = 1$ if the examinee answers item *i* correctly, and $U_i = 0$ if the examinee answers item *i* incorrectly.

IRT assumes examinee responses to test items depend upon both the characteristics of the items themselves and upon the (possibly multidimensional) latent examinee random ability vector Θ. The probability a randomly sampled examinee with ability vector Θ = θ answers item *i* correctly is given by the conditional probability:

$$P_i(\theta) = P(U_i = 1 \,|\, \Theta = \theta). \tag{14.1}$$

The model in Eq. 14.1 is referred to as the item response function (IRF) of item *i*.

Two assumptions are generally made about the form of the IRF defined in Eq. 14.1. First, the latent variable probability model is assumed to be monotone increasing coordinatewise in θ for each item *i*. This assumption implies that examinees with higher-ability levels have a higher probability of answering the item correctly. Second, the latent variable probability model is assumed to be locally independent. A latent variable model is locally independent if:

$$P_i(\mathrm{U}^{(n)} = \mathrm{u}^{(n)} \,|\, \Theta = \theta) \prod_{t=1}^{v} P(U_i = u_i \,|\, \Theta = \theta) \tag{14.2}$$

holds for all θ and all possible response patterns $\mathbf{u}^{(n)}$. Thus, examinee responses to test items are independent conditioned upon the value of the latent examinee ability vector Θ. Some models for the IRF have been developed that relax the assumption of local independence (van der Linden & Hambleton, 1997); but in the great majority of IRF models, including the most commonly used models, these two assumptions are made.

14.4 Assessing Dimensionality

The dimensionality of a test is then traditionally defined as the minimum number of dimensions *d* of the latent vector Θ required to produce a locally independent and monotone increasing latent variable IRF model (Lord, 1980; Stout, 1990). In other words, the latent ability vector Θ contains *all* dimensions or abilities that affect examinee performance on the test items, many of which could in fact exist only on a very small number of test items. Finding the value of *d* requires the verification of Eq. 14.2 for all 2^n possible response patterns for a dichotomous test and for all values of the *d* dimensional latent ability vector θ.

Humphreys (1985) and Stout (1987, 1990), along with other psychometricians, have argued the dimensionality of a test should be defined instead as the number of *dominant* dimensions on a test. So, dimensions existing only on a very small number of test items should not be considered true dimensions on the test. To define the number of dominant dimensions, the assumption of local independence is replaced with the assumption of weak local independence. A latent variable model displays weak local independence if, for each of the $n(n\text{-}1)/2$ item pairs (i,l) and for every θ (Eq. 14.3),

$$Cov(U_i, U_l = \left| \Theta = \theta \right) = 0. \tag{14.3}$$

Therefore, the basis of the definition of the number of dominant dimensions on a test is the item pair conditional covariances.

Under this framework, Zhang and Stout (1999a) defined the properties of these conditional covariances for a large class of multidimensional models for the IRF in Eq. 14.1. Each item in this class of models can be represented geometrically (Ackerman, 1996; Reckase, 1997), as illustrated in Fig. 14.1. In this representation, each item is a vector whose direction is the ability or composite ability best measured by the item. For example, Item 1 measures θ_1 alone. While Item 2 measures both θ_1 and θ_2, it best measures a composite ability that lies between the two main abilities. A set of items, such as the entire test, can also be represented by a vector. This vector, called the direction of best measurement of the test (Zhang & Stout) and denoted by Θ_{TT} in Fig. 14.1, can be thought of as the *ability* composite best measured by the test as a whole. While this unidimensional composite is usually not estimated directly by any procedure, Θ_{TT} is used to stand for its approximation by measures such as the total test score.

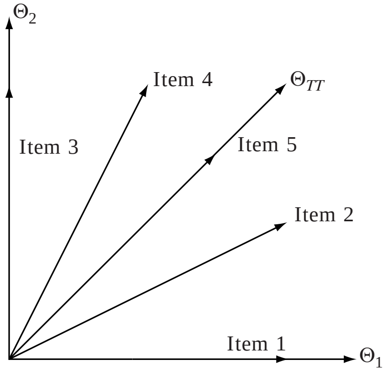


Fig. 14.1 Graphical representation of multidimensional test items

The principal result from Zhang and Stout (1999a) is that information about the multidimensional structure of a test can be recovered simply by finding the item pair conditional covariances based on the unidimensional Θ_{TT}. For example, Fig. 14.1 represents a $d = 2$ dimensional test. Items with vectors on the same side of the direction of best measurement of the test Θ_{TT} (e.g., Items 1 and 2 or Items 3 and 4) will have positive conditional covariances when conditioned on Θ_{TT}. Items with vectors on the opposite side of Θ_{TT} (e.g., Items 1 and 3 or Items 2 and 4) will have negative conditional covariances when conditioned on Θ_{TT}. Finally, the conditional covariance between items for which one or both of their vectors lie in the same direction as Θ_{TT} (e.g., Item 5 with any other item) will have a zero conditional covariance. The magnitudes of these conditional covariances are related to the closeness of the items' directions of best measurement to each other (along with their closeness to Θ_{TT} and the length of the discrimination vector). Zhang and Stout also showed by using vector projections that similar results apply to higher $d > 2$ dimensional tests.

Under the assumption of a monotone, locally independent, unidimensional ($d = 1$) model for the IRFs, the examinee ability θ is stochastically ordered by the total test score (Sijtsma & Molenaar, 2002). However, if violations of any of these three assumptions are found in the test data, the stochastic ordering of examinee abilities θ by the total test score does not hold. Thus, the determination of the true dimensional structure of the test data becomes an important consideration to establishing the validity of an assessment.

In practice, the conditional covariance values and the direction of best measurement of the test that are the basis of dimensionality analysis are estimated from the examinee by item test data. Three procedures and associated programs are available to study different aspects of the dimensional structure of an examination through these estimated conditional covariance values. In the sections below, each of these three programs is described. An example of their use on a large-scale standardized assessment can be found in Stout and colleagues (1996).

14.4.1 HCA/CCPROX

One method for exploring the possible dimensionality of an assessment is HCA/CCPROX (Roussos, Stout, & Marden, 1998). This is an agglomerative hierarchical cluster analysis method; each item starts as a separate cluster, and at each stage the two clusters with the smallest distance between them are joined together. At the final stage, all of the items are joined together in a single cluster. The proximity measure used for judging the distance or proximity between a pair of items is based on the estimated conditional covariance of the two items, conditioned on the direction of best measurement of the test estimated by a measure of the total test score. Any two items that measure close to one another when compared with the direction of best measurement of the test will have a small proximity value, and any two items that measure far away from one another when compared with the direction of best measurement of the test will have a large proximity value. Each stage of the cluster analysis combines one item with another item, or item cluster, based on the value of the proximity between that item and the other item, or item cluster. In this way, items formed into clusters early in the cluster analysis are more dimensionally alike than items formed into clusters later in the cluster analysis.

The program HCA/CCPROX is completely exploratory in nature. No determination is made on which clustering stage contains the correct clustering of the test items by dimension, nor does HCA/CCPROX determine a stopping point for the cluster analysis. For these reasons, HCA/CCPROX is seen as a first step in exploring the dimensional structure of an assessment.

14.4.2 DETECT

The second step in exploring the dimensional structure of test data is the program DETECT. The DETECT procedure (Kim, 1994; Zhang & Stout, 1999b) is designed to find the optimal partitioning of test data, assuming the test items cluster around a given number of composite abilities. Even when the test data do not have this structure, the number of clusters found in the final partition by the DETECT procedure should approximate the dimensionality of the test. In DETECT, the basic idea is to find the partitioning of the test items so that items within the same partition will have a positive conditional covariance with each other and a negative conditional covariance with items from any other cluster. The optimal partitioning will then maximize the value of the estimated DETECT index, which is based on these conditional covariances. Ideally, all possible partitionings of the items to clusters would be used to determine the partitioning that maximizes the estimated DETECT index value. However, in practice, the procedure uses a genetic algorithm to search for these partitions based on the initial item clusterings provided by the HCA/CCPROX program.

Unlike HCA/CCPROX, DETECT estimates the dimensional structure of the test data by returning both the number of clusters of the test items and the division of the test items to these clusters. However, like HCA/CCPROX, this procedure is exploratory in nature. No determination is made to whether the clusters found by this procedure are actually dimensionally distinct.

14.4.3 DIMTEST

The DIMTEST procedure (Froelich, 2008; Nandakumar & Stout, 1993; Stout, 1987; Stout, Froelich, & Gao, 2001) is a hypothesis test designed to test for the dimensional distinctiveness of two clusters of items. In the DIMTEST literature, these two clusters are called the assessment subtest (AT) and the partitioning subtest (PT). The null hypothesis of the DIMTEST procedure is that AT and PT can be fit by a unidimensional, locally independent, monotone latent variable IRF model. Under the assumptions of a locally independent and monotone IRF model, the alternative hypothesis becomes that AT and PT are dimensionally distinct; they measure two different dimensions.

Thus, clusters found as a part of the HCA/CCPROX and DETECT procedures can be tested to determine if they are truly dimensionally distinct. If AT constitutes a carefully chosen set of items thought to be dimensionally different from the rest of the test items and PT is the rest of the test items, DIMTEST can be used to test for the unidimensionality of the entire test. In simulation studies, Froelich (2008) and Froelich and Habing (2008) have found the DIMTEST procedure has a Type I error rate around the nominal rate of α and very high power for detecting departures from unidimensionality.

14.5 Item Reduction: Assessing the Strength of the Resulting Scale(s)

By applying the three dimensionality programs HCA/CCPROX, DETECT, and DIMTEST, the dimensional structure of the test in the population of interest can be estimated given the test data. While the use of these programs can indicate the overall dimensionality of the test data, they were not designed to look at the fit of individual items to the monotone, locally independent, unidimensional IRF model. Even test data determined through the use of the DIMTEST procedure to be unidimensional could contain a small number of items that do not fit this IRF model. The inclusion of these items on the test negates the stochastic ordering of the examinee abilities by the total test score.

Mokken scaling (Mokken, 1971; Molenaar & Sijtsma, 2000) is a method for studying the fit of individual items to this monotone, locally independent, unidimensional model for the IRF. For a collection of test items displaying these three assumptions, the following properties of the (unconditional) covariance are true:

- $Cov(U_i,U_l) > 0$ for all item pairs *i* and *l*.
- Define $R_{(i)}$ as the rest score, the number correct score on all items excluding item i. Then $Cov(U_i,R_{(i)}) > 0$ for all items i.

Based on these two properties, the scalability coefficient of an item pair $H_{i,l}$ is defined in Eq. 14.4 as:

$$H_{i,l} = \frac{Cov(U_i,U_l)}{Cov_{\max}(U_i,U_l)} \tag{14.4}$$

where Cov_{max} is the maximum possible covariance between the two dichotomous items U_i and U_l. Thus, $H_{i,l}$ will be greater than 0 for a monotone, locally independent, unidimensional IRF model for items *i* and *l* and will have maximum value 1. In the same way, a scalability coefficient of an item *i* is defined in Eq. 14.5 as:

$$H_i = \frac{Cov(U_i, R_{(i)})}{Cov_{\max}(U_i, R_{(i)})} \tag{14.5}$$

and the scalability of a test or subtest containing *n* items is defined in Eq. 14.6 as:

$$H = \frac{\sum_{i=l}^{n} Cov(U_i, R_{(i)})}{\sum_{i=l}^{n} Cov_{\max}(U_i, R_{(i)})}. \tag{14.6}$$

Under a monotone, locally independent, unidimensional IRF model, both H_i and H will be greater than 0 and will have maximum values of 1. It can also be shown that the scalability of the test H is bounded below by the minimum scalability coefficient of the test's items and bounded above by the maximum scalability coefficient of the test's items ($\min(H_i) < H < \max(\mathrm{H}_i)$).

A Mokken scale (Sijtsma & Molenaar, 2002) is then defined as a set of items where:

- $H_{i,l} > 0$ for all item pairs *i* and *l*,
- $H_i \geq c > 0$ for a given constant *c*, which implies that
- $H \geq c > 0$ for the same constant *c*.

The constant *c* determines the strength of the resulting scale. When $0.3 \leq c < 0.4$, the scale is referred to as a weak scale. When $0.4 \leq c < 0.5$, the scale is referred to as a medium-strength scale; and for $c \geq 0.5$, the scale is called strong (Sijtsma & Molenaar). Tests producing a constant $c < 0.3$ are usually considered unacceptable for use as a scale, making $c = 0.3$ the practical lower bound for a Mokken scale. Items are generally deleted from the test one at a time through an analysis of the values of $H_{i,l}$ for all item pairs, the values of H_i for each item, and the value of H for the test in order to produce a set of items satisfying the requirements of a Mokken scale.

14.6 Estimating the Model Parameters

Thus far, other than the triple assumptions of a monotone, locally independent, unidimensional latent variable model for the IRF, no other assumptions have been made about the form of the IRF given in Eq. 14.1. Generally, when a set of items is found to satisfy these three conditions (through the methods in the Sects. 14.4 and 14.5), the form of this IRF is assumed to be logistic in nature as described by the three-parameter logistic model (Eq. 14.7, Birnbaum, 1968):

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]} \tag{14.7}$$

where θ is the unidimensional ability of a randomly selected examinee. The item parameters are the item discrimination parameter (a_i), the item difficulty parameter

(b_i), and the item pseudo-guessing parameter (c_i). The pseudo-guessing parameter (c_i) is the lower asymptote of the IRF and is thought of as the probability that an examinee with very low ability will answer the item correctly. The examinee with very low ability may or may not actually answer the item by guessing; thus, the attachment of *pseudo* to the parameter's name. The item difficulty parameter (b_i) is the location of the inflection point of the curve on the ability scale θ. The probability $P_i(\theta)$ at this point b_i is, therefore, $(1+c_i)/2$. Easier items will have a smaller value of b_i while more difficult items will have a larger value for b_i. Finally, the item discrimination parameter a_i is proportional to the slope of the curve at the inflection point $\theta = b_i$. Items with higher discrimination values do a better job of discriminating between examinees with abilities around the inflection point than items with smaller discrimination values. Figure 14.2 contains three different IRFs with varying item parameters (a, b, c). The different values of the parameters were chosen to illustrate different possible IRFs that can be obtained using Eq. 14.7.

For an *n*-item test with *N* examinees, the set of 3*n* item parameters and *N* examinee ability parameters can be estimated using several different techniques based on maximizing a likelihood function. The most commonly used technique is called the marginal maximum likelihood procedure (Bock & Aitkin, 1981) and is implemented in the software program BILOG (Zimowski, Muraki, Mislevy, & Bock, 2007). The procedure consists of two phases. In Phase I, the item parameters for each of the *n* test items are estimated using a distribution of examinee abilities $g(\theta)$ in place of the individual examinee ability parameters θ. The distribution $g(\theta)$ is chosen to match the distribution of abilities thought to occur in the population of interest. In educational testing, $g(\theta)$ is usually set to the standard normal distribution. In Phase II, the values of the item parameters are fixed at the estimated values found in Phase I; and estimates of each examinee's ability θ is found through a maximum likelihood

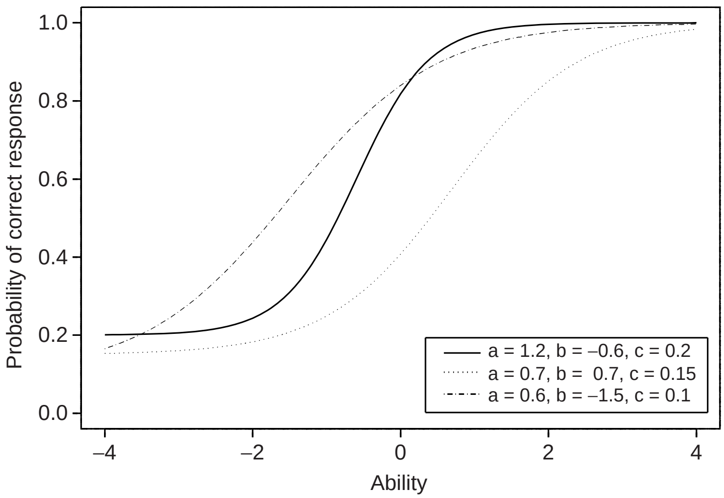


Fig. 14.2 An example of three-item response functions

procedure. It is not required that Phase I and Phase II estimations be conducted on the same test dataset. Often, the Phase I estimation procedure is used to estimate the item parameters of a standardized test, particularly of pretest or preoperational items. As these items become operational (used to estimate examinee abilities), the item parameters are fixed at the values found in previous Phase I investigations. Even without this process of pretesting items, an analysis of a standardized assessment can benefit from a Phase I estimation of the item parameters without conducting the second phase of examinee ability estimation.

Given the set of estimated item parameters for an *n*-item test, an estimate of the IRF $\hat{P}_i(\theta)$ for each item *i* from Eq. 14.7 can be found. We can use this estimated IRF to calculate the observed item information function (Eq. 14.8):

$$I_i(\theta) = \frac{(\hat{P}_i'(\theta))^2}{(\hat{P}_i(\theta)(1-(\hat{P}_i(\theta))} \tag{14.8}$$

where $\hat{P}_i(\theta)$ is the derivative of the estimated IRF at the value θ. The item information function gives the range of θ values where the item provides the most *information* about the examinee ability level. Examples of the item information functions for the three items in Fig. 14.2 are given in Fig. 14.3.

These item information functions are then summed to give the test information function $I(\theta) = \sum_{i=l}^{n} \mathrm{I}_t(\theta)$. The test information function, therefore, gives the range of θ values where the test provides the most information. Figure 14.4 depicts a test information function obtained from estimated item parameters from a 25-item Armed Services Vocational Aptitude Battery (ASVAB) Auto Shop test (Mislevy & Bock, 1984).

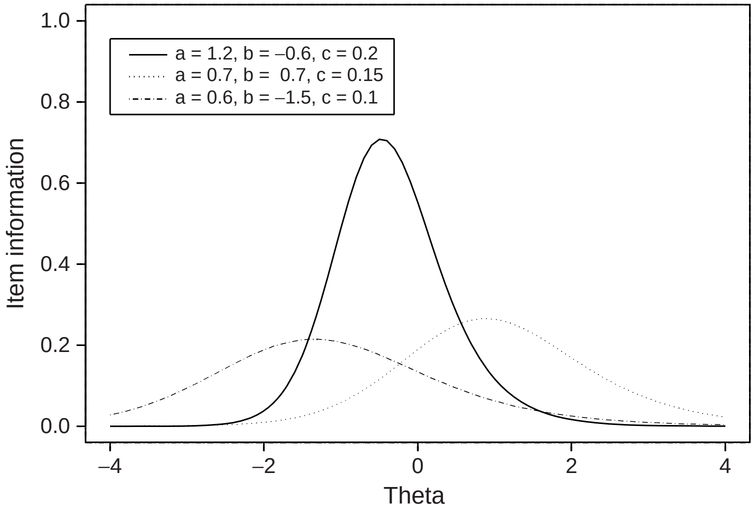


Fig. 14.3 An example of three-item information functions

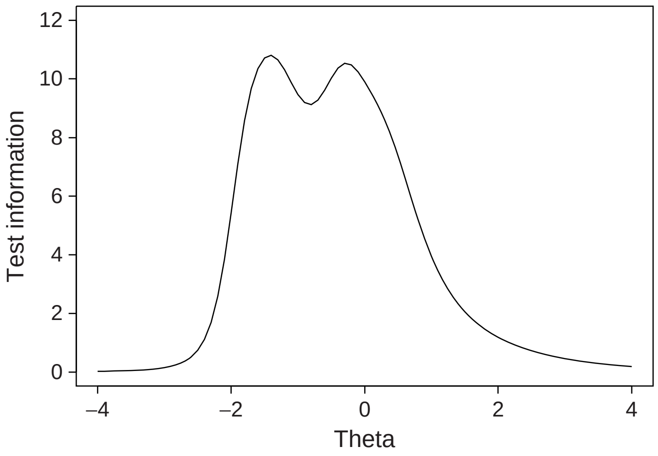


Fig. 14.4 An example of a test information function

In the Phase II estimation of examinee abilities, the square root of the inverse of the observed test information function $I(\theta)$ is the standard error of the estimator of θ (Lord, 1984). Thus, the standard error of measurement of the test will be lower for abilities θ with higher test information function values $I(\theta)$, and the standard error of measurement of the test will be higher for abilities θ with lower test information function values $I(\theta)$.

The most appropriate test information function for a given standardized assessment will be determined by the purpose of the assessment. If an assessment is used to classify the performance of examinees throughout the ability scale, the test information function should be constant or flat across the entire ability scale. In this way, the standard error of the estimated abilities is equal across all ability levels. However, if an assessment is used to primarily classify students into two or more groups based on their estimated abilities, the test information function should have a peak around the ability values where the classifications are made. In this way, the standard error of the estimated abilities is smaller in the crucial range of examinee abilities where the classification decisions are made and larger in the range of examinee abilities not affected by these decisions.

Therefore, researchers should match the intended use of the assessment with this test information function $I(\theta)$. If the test is being developed as a general measure to indicate the effectiveness of an educational program or intervention or to measure ability–achievement in a particular subject, a constant test information function is desired. In this application, a peaked test information function leads to a loss of desired information about examinees in a certain range or ranges of abilities. Likewise, if a test is being developed as a certification examination or for some other high-stakes decision, a constant test information function leads to a loss of desired information about examinees with abilities around the range of classification.

In essence, some of the test items are wasted in this case since they provide information about examinee abilities ultimately not affected by the classification.

If a mismatch occurs between the test information function of the standardized assessment and its intended use, care should be taken to develop additional items or remove items from the test, or both, in order to change the test information function to a more desired form. If additional items are added, they should first be tested by the methods in Sects. 14.4 and 14.5 of this chapter in order to determine if the test including the new items can be fit by a monotone, locally independent, unidimensional IRF model. These items can then be used to study whether the test information function is a better fit for the intended use of the standardized assessment.

14.7 Conclusions

The methods detailed in this chapter are just several steps of many that researchers should undertake in order to validate their standardized assessment. It is important to conduct a full content analysis of the field and area covered by the assessment and to look at the theoretical developments in learning theory applied to that field and area. These analyses should not be thought of as separate analyses from the methods described above. While statistical analyses can estimate the dimensional structure of an assessment or can help determine problems with individual test items, the analyses themselves cannot explain these findings. Thus, psychometricians with experience in educational measurement and IRT, educational researchers with experience in learning theory, and content area specialists must work together to fully analyze and validate these assessments.

References

Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, *20*(4), 311–329.

ACT. (n.d.). *Homepage*. Retrieved July 11, 2008, from http://www.act.org/

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425–440.

College Board. (n.d.). *About the SAT*. Retrieved May 15, 2008, from http://www.collegeboard.com/student/testing/sat/about.html

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Froelich, A. G. (2008). *A new bias correction method for the DIMTEST procedure*. Unpublished manuscript, Iowa State University at Ames.

Froelich, A. G., & Habing, B. (2008). Conditional covariance-based subtest selection for DIMTEST. *Applied Psychological Measurement*, *32*(2), 138–155.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th edn., pp. 65–110). Westport, CT: American Council on Education & Praeger.

Humphreys, L. C. (1985). General intelligence: An integration of factor, test, and simplex theory. In B. J. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications* (pp. 201–224). New York: John Wiley & Sons.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th edn., pp. 17–64). Westport, CT: American Council on Education & Praeger.

Kim, H. R. (1994). *New techniques for the dimensionality assessment of standardized test data*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Linden, W. J., van der, & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement*, *21*(3), 239–243.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd edn., pp. 13–103). New York: American Council on Education.

Mislevy, R. J., & Bock, R. D. (1984). *Item operating characteristics of the Armed Services Aptitude Battery, Form 8A* (Technical Report N00014-83-C-0283). Washington, DC: Office of Naval Research.

Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research*. The Hague, The Netherlands: Mouton.

Molenaar, I. W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows*. Groningen, The Netherlands: iecProGAMMA.

Nandakumar, R., & Stout, W. F. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, *18*(1), 41–68.

No Child Left Behind Act of 2001. Pub. L. No. 107–110, 115 Stat. 1425. (2002).

Reckase, M. D. (1997). A linear logistic model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement*, *35*(1), 1–30.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589–617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*(2), 293–325.

Stout, W. F., Froelich, A. G., & Gao, F. (2001). Using resampling to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Dujin, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 357–375). Dordrecht, The Netherlands: Springer.

Stout, W. F., Habing, B., Douglas, J., Hae Rim, K., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*(4), 331–354.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications*. Thousand Oaks, CA: Sage.

United States National Center for Education Statistics. (n.d.). *NAEP: The nation's report card*. Retrieved July 11, 2008, from http://nces.ed.gov/nationsreportcard/

Zhang, J., & Stout, W. F. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika*, *64*(2), 129–152.

Zhang, J., & Stout, W. F. (1999b). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*(2), 213–249.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2007). BILOG-MG3 [computer software]. Mooresville, IN: Scientific Software International. Available from http://www.ssicentral.com/irt/index.html

Chapter 15
Confounding in Observational Studies using Standardized Test Data: Careful Disentanglement of Statistical Interpretations and Explanations

Mary C. Meyer

Standardized testing of public school students has been, and continues to be, a focal point of the political side of education reform. The use of standardized test results to set policies and identify *problem* schools requires some understanding of what factors contribute to high or low overall scores. Many large datasets at the state and national levels contain information about mean scores (i.e., average class size, poverty level, teacher salary, etc.); it is tempting to create statistical models, draw cause-and-effect conclusions, and perhaps set policy based on statistically significant relationships observed in these data; for example, National Assessment of Educational Progress (NAEP, US National Center for Education Statistics, n.d.); Trends in International Mathematics and Science Study, and Progress in International Reading Literacy Study (TIMSS & PIRLS International Study Center, n.d.); Programme for International Student Assessment (PISA, Organisation for Economic Co-operation and Development, n.d.). However, many examples of confounding, that is, the apparent associations between the variables that change depending on which covariates are selected, can be found in these associations. In this chapter, some results using standardized testing data are presented, which demonstrate by example the difficulties inherent in making conclusions or comparisons based on observational data and disentangling second- and third-order influences on these relationships. First, some statistical terminology is reviewed, and some simplified, fabricated examples are presented to illustrate concepts. Next, a dataset containing scores for the Illinois Standardized Achievement Test (ISAT, Illinois State Board of Education, n.d.), taken by Grade 8 students in Illinois public schools, is used to demonstrate confounding relationships. Finally, the Scholastic Achievement Test (SAT, College Board, n.d.) scores by state are used to show some misleading rankings of states' average scores.

M.C. Meyer
Colorado State University

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

15.1 Background

Statistical models attempt to describe and quantify relationships between variables. In the models presented in this chapter, there is a response variable (sometimes called dependent variable) and at least one predictor variable (sometimes called independent or explanatory variable). When investigating a possible cause-and-effect type of relationship, the response variable is the putative effect and the predictors are the hypothesized causes. Typically, there is a main predictor variable of interest; other predictors in the model are called covariates. Unknown covariates or other independent variables not controlled in an experiment or analysis can affect the dependent or outcome variable and mislead the conclusions made from the inquiry (Bock, Velleman, & De Veaux, 2009).

A *p* value (*p*) measures the statistical significance of the observed relationship; given the model, *p* is the probability that a relationship is seen by mere chance. The smaller the *p* value, the more confident we can be that the pattern seen in the data is not random. In the type of models examined here, the R^2 measures the proportion of the variation in the response variable that is explained by the predictors specified in the model; if R^2 is close to 1, then almost all the variation in the response variable has been explained. This measure is also known as the multiple correlation coefficient.

Statistical studies can be grouped into two types: experimental and observational. In an experiment, the researchers set the value of the main predictor variable of interest; in an observational study, the variable is simply noted. For example, suppose a health researcher wanted to know the effects of oat bran consumption on blood pressure. In an experiment, the subjects would be assigned levels of oat bran; but in an observational study, the subjects would be asked about their usual oat bran consumption levels. In assigning cause-and-effect conclusions, it is important to determine whether the data come from an experiment or an observational study.

Confounding occurs when a relationship between the response and the main predictor of interest is altered when a covariate is taken into account. This can happen when the covariate is related to both the response variable and the main predictor; it is a potential problem with observational studies. To continue the previous example, suppose people who usually eat a lot of oat bran also exercise more regularly and consume less salt. If an association between oat bran and blood pressure is found in the observational study, we do not know whether it is really driven by the exercising or the salt levels. In the experiment, people who exercise more and those who use less salt are likely to be roughly evenly distributed among oat bran groups; therefore, confounding is less likely.

To illustrate these concepts further and demonstrate the phenomenon of confounding, we will use two simple, fabricated examples. First, suppose it is observed that, in the population of elementary school children, those students with larger feet have higher reading abilities. In this observational study, the response variable is the reading score of the child and the main predictor of interest is the child's foot length. The dataset (paired measures of reading and foot length for each elementary school child in a specified population) is shown in Fig. 15.1 (left), with the response variable plotted against the predictor variable. Each dot represents a child's data pair, and a definite pattern is seen in the plot. The relationship can be assessed with simple linear regression (fitting a line to the data), as shown by the line in plot (a). This correlation is highly statistically significant, with $p < .00001$; further, the R^2 is 0.812, indicating that 81.2% of the variation in reading score is explained by foot length. Impressive!

However, before we can conclude that having big feet *causes* kids to be smarter, we should note that the children observed represent a range of ages (5–10 years old). In Fig. 15.1 (right), the ages of the children are used as plot characters for each subgroup of students. It is easy to see that 5-year-olds (represented by circles) have both smaller feet and lower reading abilities, compared with 10-year-olds (inverted triangles). If we perform a multiple regression using both age and foot length as variables to predict reading ability, we find that the foot length variable is no longer statistically significant ($p > .25$). The fit within age groups shown in plot (b) indicates that the relationship between reading score and foot length is fairly flat, reflecting the large p value for this relationship. We say that age was confounding the relationship between reading score and foot length because we reach a different conclusion about the effect of foot length on reading when the age variable is accounted for in the model. The very strong statistical significance seen in the simple linear regression of the total dataset not aggregated by age was not the result of a direct relationship between reading and foot length; it was caused by two other age-dependent relationships: reading with age and foot length with age. A small p value for the linear regression indicates that the pattern is not random, but it does not necessarily reflect a direct cause-and-effect phenomenon.

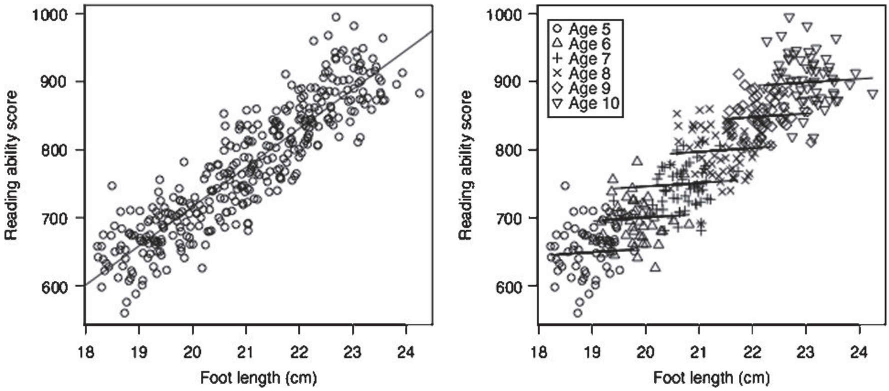


Fig. 15.1 Fictitious example of the relationship between reading score and foot length for elementary school children ($N = 350$). The simple linear regression in plot **(***left***)** produces a very significant association between reading ability and foot size. However, when the age variable is accounted for in the model, we see that foot size is no longer a significant predictor of reading ability ($p > .25$). The fit to the model with two predictor variables is shown in plot **(***right***)**

The foot length example was fabricated to be obviously silly. No one would seriously claim that having bigger feet is a cause of higher reading ability. However, more believable cause-and-effect claims are often made with as little scientific merit. If the predictor variable were head circumference instead of foot length, a cause-and-effect conclusion might not be as readily questioned but would be equally invalid. One famous example of a serious mistake made by attributing cause and effect to a relationship with observational evidence was refuted in recent years. Many observational studies showed that postmenopausal women who were taking hormone replacement therapy (HRT) had lower rates of heart disease than women who did not. It became usual to prescribe HRT to women as a heart-disease preventive although income and insurance status are possible confounders. When experiments were performed, they showed that the effects of HRT on otherwise healthy women seemed to be detrimental rather than otherwise. An editorial in the *Journal of the American Medical Association* proclaimed about the results of one set of clinical trials: “Experimentation Trumps Observation” (Petitti, 1998). Clearly, users of observational and experimental research results need to be critical readers with a healthy skeptical and emotional disposition characteristic of scientific literacy (Yore, Pimm, & Tuan, 2007).

Let us now consider a fictitious example from education: suppose Burbville has three secondary schools—two standard and the other called an academy. In the last 5 years, students at the academy have attained substantially higher SAT scores and more college scholarships compared to those students at the standard schools. Using the descriptive statement as the basis of a tentative relationship (hypothesis) would make the response variable the student’s SAT score, and the main predictor of interest would be the categorical variable school attended. Many people point to differences in curriculum as the cause of the difference in SAT performance and why the academy wins awards. However, the assignment of students to schools is not random. Many parents wait in line for days to put their children on the waiting list for the better school, and students have to write essays for admission. Perhaps if the variables of parental involvement in education and student motivation were to be accounted for in the assessment of performance by school, attendance at the academy would no longer be a significant predictor of SAT score. In other words, a competing tentative relationship (hypothesis) might be that *any* school having students with high parental involvement and high student motivation does better on the SAT, and the academy simply has more of these students.

If the assignment of students to Burbville schools had been random, the data would be experimental. Students with high motivation and parental involvement would be roughly equally distributed among schools. Other possibly important sociocultural factors, such as family income, neighborhood, etc., would also be roughly equally distributed in a random assignment. In the experimental setting with random assignments if an outcome is significantly different, a cause-and-effect conclusion is warranted. However, in an observational study, confounding is almost always an issue, especially if the subjects self-select into the groups. The assignment of group might depend on a variable that also is related to the outcome measure; in our example, student motivation is related to both enrollment and SAT score. If a confounding variable can be accounted for in the analysis, a better picture of relationships between variables emerges (see Anderson, Milford, & Ross, Chap. 13). However, it is difficult to account for all possible confounding factors; and it is an unfortunate truism, especially in education, that the most important variables are the hardest to measure.

Let us now turn to two real examples of confounding using standardized testing data—ISAT scores from Illinois and SAT rankings by state. The purpose of these discussions is not to analyze the performance of Illinois schools using the ISAT data or to make comparisons between states but to (a) use these data to present examples illustrating the difficulties inherent in drawing conclusions from observational data and many other similar datasets, and (b) increase the awareness and criticality of literacy and science education researchers to such potential outcomes in secondary analyses.

15.2 Illinois Standardized Testing Data: Class Size, Socioeconomic, and Ethnicity Predictors of Percent Passing the ISAT

Standardized testing data for schools in Illinois and many other states are available to the general public. The ISAT, for example, is given to all Grade 8 students in Illinois; it tests the children in reading, writing, and mathematics. The percent of students passing the test is recorded for each public school in the state. Other school-level measures are available, such as pupil/teacher ratio, percent low-income students, proportion of ethnicities, minutes per day of mathematics instruction, etc. These data may be used to describe schools and school districts with unusually high or low passing percentages or to determine relationships between test performance measures and various characteristics of the school. (The interested reader should explore the Web site http://www.isbe.net/assessment/isat.htm for further information.)

Statistical computing packages are readily used to quantify relationships between variables and to obtain measures of statistical significance, such as a *p* value. There are a wide variety of statistical programs that provide graphical display of data and the results of analyses. Most commonly used statistical packages (e.g., SYSTAT, www.systat.com, and SPSS, www.spss.com) have comprehensive sets of graphical displays with editing capabilities. There are also free downloads available—one of the most widely used is R, which is a language and environment for statistical computing and graphics (R Project for Statistical Computing, n.d.). R provides a wide variety of statistical (including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, and clustering) and graphical techniques. One of its strengths is the ease with which well-designed, publication-quality plots can be produced, including mathematical symbols and formulae where needed.

However, although software provides tables of statistical results, interpretation of the results is often not as straightforward as it might seem. If the results of standardized tests are to play a role in educational policymaking, it is important to determine the critical predictors of passing rates. For example, if it is determined that smaller class sizes are related to higher scores, increasing funding to increase the number of teachers at the school might be considered. The data may also be used to compare schools so that those with lower performance might be targeted for changes and those with acceptable or higher performance perhaps used as models.

Let us start with the question of class size. Many educators believe that children in smaller classes learn better (Glass, 1982; Smith & Glass, 1980). To the extent that *better learning* is reflected in success on standardized tests, we can attempt to assess this claim using the ISAT data. A scatter plot of the percent of students passing the ISAT against the average Grade 8 class size in their school is shown in Fig. 15.2. Each point represents a data pair (percent passing, average class size) for a school in Illinois ($N = 3{,}134$), and a general downward trend is apparent if not sharp ($p < .0001$). The scatter is large, indicating that the correlation is not strong. The R^2 is only 0.03, indicating that about 3% of the variation in percent passing the ISAT is explained by the average class size; 97% is unexplained. However, the statistical significance is strong, reflected in the very small p value for the downward slope. Analysis of large datasets can yield a small p value indicating statistical significance, but the small R^2 is an indication that practical significance is in question. (Technical note: To correct for heteroskedasticity, all regression models discussed here are weighted by the number of students taking the test.) It is tempting to conclude that larger class sizes are among the causes of poor test performance.

Examination of the scatter plot in Fig. 15.2 reveals that the points fall roughly into two clumps (noted by overprint of data points): one big clump at the top half of the plot

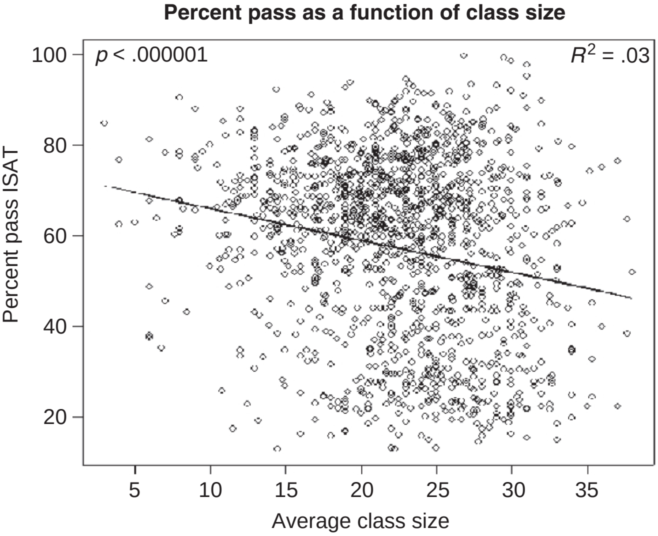


Fig. 15.2 The relationship between percent of Grade 8 students passing the ISAT and average class size is shown. Each point represents a school in Illinois. The line is calculated using least-squares regression, weighted by the estimated number of students taking the examination. While only 3% of the variation in percent pass is explained by average class size, the downward slope is highly significant ($p < .000001$)

and a smaller one beneath, slightly shifted to the right. The lower clump has both larger class sizes, on average, and lower percent ISAT pass, compared to the higher clump. Within each clump, the downward trend is not apparent. Perhaps the overall trend is created by the positions of the two clumps; but within the clumps, the trend does not exist. This would indicate that the statistically significant relationship between percent ISAT pass and class size is confounded by whatever is causing the clumps.

Investigation into the lower data points reveals that these schools are mostly in the city of Chicago. Anyone familiar with Illinois could guess that an analysis of education data ought to recognize that conditions in Chicago tend to be different from the rest of the state, and a statistical model ought to allow for these differences. In other contexts, one might wish to model the size of the city or rural versus urban school districts; in some settings, it might be important to model class or religious ethnicities. In general, investigators may use their judgment in selecting possibly important predictor variables for the analyses.

15.2.1 Socioeconomic Effects and the ISAT Scores

Shown in Fig. 15.3 is the same scatter plot as in Fig. 15.2 but with Chicago schools marked as triangles. Most of the bottom points (low-performing schools) are Chicago schools, but there is some overlap between the groups as well. Regression lines fit to the two sets of data, utilizing separate regression analyses, now show

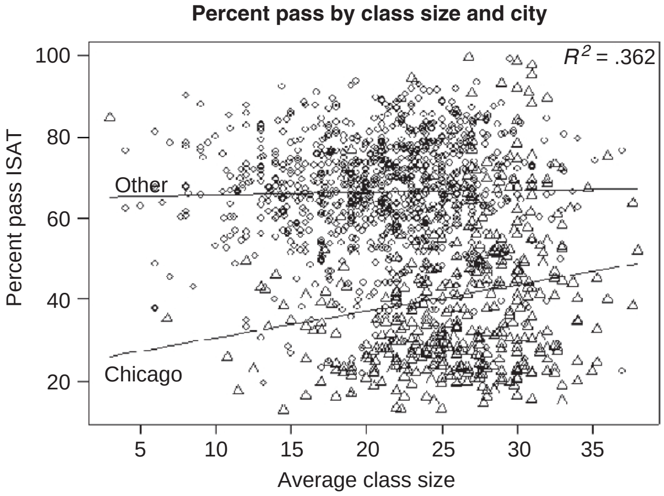


Fig. 15.3 The relationship between percent of students passing the ISAT and average class size is demonstrated in a scatter plot, with Chicago schools marked as triangles. The lines are fit to the data using weighted least squares, with the bottom line representing the Chicago schools

a different relationship between percent passing the ISAT and average class size. We see that for Chicago schools the percent pass now *increases* significantly (p < .0001) with class size; while for the other schools, the percent pass does not significantly change with class size (p > .10). The two lines explain 36.2% of the variation in percent pass although the Chicago variable is responsible for most of the explained variation. This is determined by the fact that the R^2 value for the model with only the Chicago variable is almost as large as that for both variables together. The explanatory capability of the average class size for predicting the percent passing the ISAT is still very low.

Of course, we should not conclude that larger class sizes are beneficial for Chicago students. There are other possible confounders, and visual inspection suggests that the line for the Chicago schools does not seem to fit the data very well. The pattern of triangles in the scatter plot suggests that within the Chicago schools two clumps exist: the first in the upper-right corner (large class sizes, high ISAT performance) and the second distributed across the bottom half of the plot (mixed class sizes, lower ISAT performance). This upper clump seems to be pulling the line upwards; and perhaps if this clump is identified, the relationship between percent passing the ISAT and average class size will change yet again.

In Fig. 15.3, the triangles are, on average, dramatically lower than the circles, indicating that Chicago schools have, on average, much lower percent passing scores. Let us quantify the difference and see if we can find explanations for it based on available covariates. The plot in Fig. 15.4 shows percent pass with schools divided into two groups: those in Chicago and those in the rest of the state. The points are jittered (scattered randomly side to side) to better indicate the distributions of the scores in each group. The averages in each group are marked with horizontal lines. The difference is immediate: percent pass is considerably lower, on average,

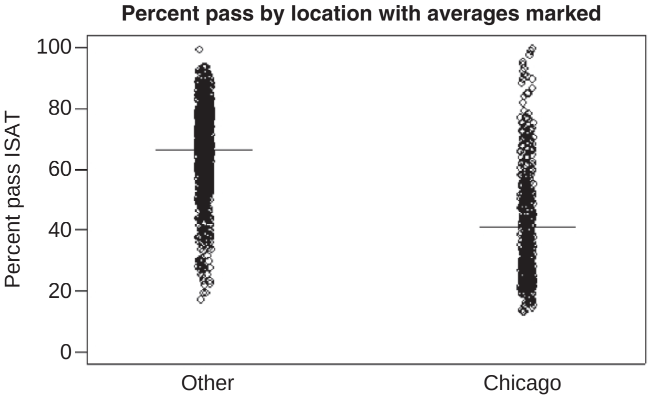


Fig. 15.4 Percent pass in Chicago schools compared with other schools. The horizontal lines indicate average percent pass for each group

in Chicago schools (68% versus 41%). How should Illinois politicians and educators react to this dramatic difference? Should the school system in Chicago be completely overhauled? Should the administrators be fired?

An investigation into the reasons why Chicago schools have low percent pass might reasonably start with a look at poverty rates (Bradley & Corwyn, 2002). If Chicago has more underprivileged children, this might be related to differences between percent pass because we would expect children of higher socioeconomic status to perform better on these tests (Entwisle & Alexander, 1988).The overall socioeconomic status of the children in the individual schools can be measured by the variable percent low income in the ISAT dataset.

Figure 15.5 shows two dramatic relationships of low income with the two variables (location and ISAT performance) in the previous plot. First, in the left plot we see that most Chicago schools have very high proportions of low-income students. In fact, average proportion of low income for Chicago is 85%, compared to the average for other schools of only 28%. Second, in the right plot we see that the percent of students passing the ISAT decreases steeply as percent low income in the school increases.

Because percent low income is a strong predictor of percent pass, and Chicago schools have much higher proportions of low-income students, a fairer assessment of the Chicago schools' ISAT performance should control for the low-income variable. The relationship between percent passing the ISAT and percent low income is shown in Fig. 15.6, with Chicago schools and other schools plotted separately. While the decrease in percent pass with rising percent low income looks roughly linear in the other schools (left), there appears to be some curvature in the relationship for the Chicago schools (right). In particular, it appears that Chicago schools do comparatively well, on average, where percent low income is in the middle range of the scale. The least-squares line for the other schools and the least-squares

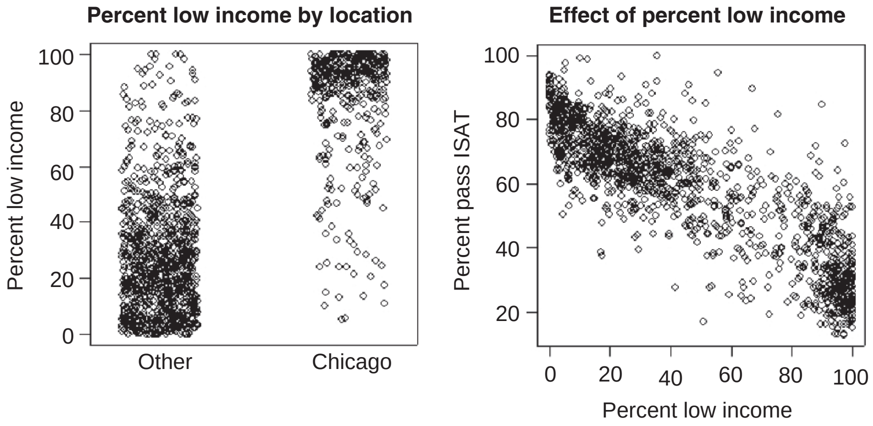


Fig. 15.5 The relationships of percent low income with the location variable and the ISAT percent passing variable. Both relationships are dramatic, indicating that a fair comparison of the performance of Chicago schools must take into account the proportions of low-income students

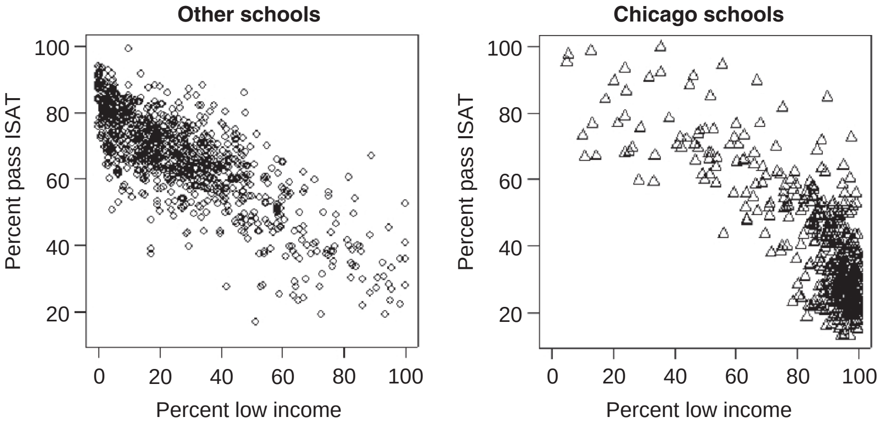


Fig. 15.6 Percent pass plotted against percent low income, with Chicago schools shown separately from other schools. While there are more Chicago schools with high proportions of low-income students, we see that for moderate-to-low percent low income the percent passing the ISAT appears to be higher, on average, for Chicago schools

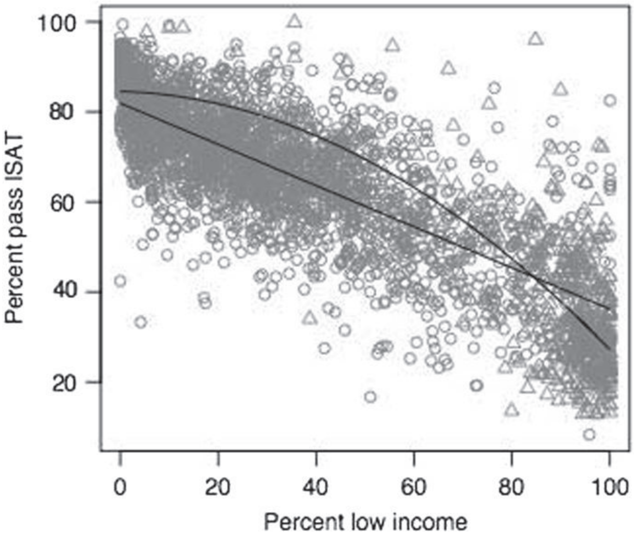


Fig. 15.7 Percent passing the ISAT is plotted against percent low income, with triangles representing Chicago schools. Also shown are the least-squares linear fit to the other schools and the least-squares quadratic fit to the Chicago schools. Given a percent low income in the middle range, the percent pass for Chicago schools is significantly higher

parabola for the Chicago schools are shown in Fig. 15.7, where the plots and fits to the data are superimposed. While Chicago schools have substantially lower passing rates, on average, they also have a lot more poverty. When poverty is controlled for (i.e., included in the statistical model), Chicago schools have *higher* percent pass, on average. In particular, for schools with medium poverty levels, the Chicago effect is greater.

The class-size issue can now be revisited by creating a statistical model with three predictors: average Grade 8 class size, the Chicago variable, and the percent low-income variable. We find that class size is no longer a significant predictor of the students' ISAT performance (percent passing). This is another example of confounding—many of the Chicago schools with medium-to-low levels of poverty have large class sizes so that the poverty level variable was confounding the relationship between class size and percent passing for the Chicago schools.

15.2.2 Minority Students and ISAT Standardized Tests

Differences in performance on standardized tests for different ethnicities have received attention from the media and from school districts. In particular, black and Hispanic students score lower, on average, than white students; and Asian students tend to have, on average, higher scores than white students (Kao & Thompson, 2003).To illustrate this effect for the Illinois data, the percent pass ISAT is plotted against percent black students and percent Hispanic students in the top two plots of Fig. 15.8. The top-left plot shows that schools in Illinois are still quite segregated; many have either almost all black students or very low percentage of black students. There is a steep decline in percent passing the ISAT across the range, indicated by

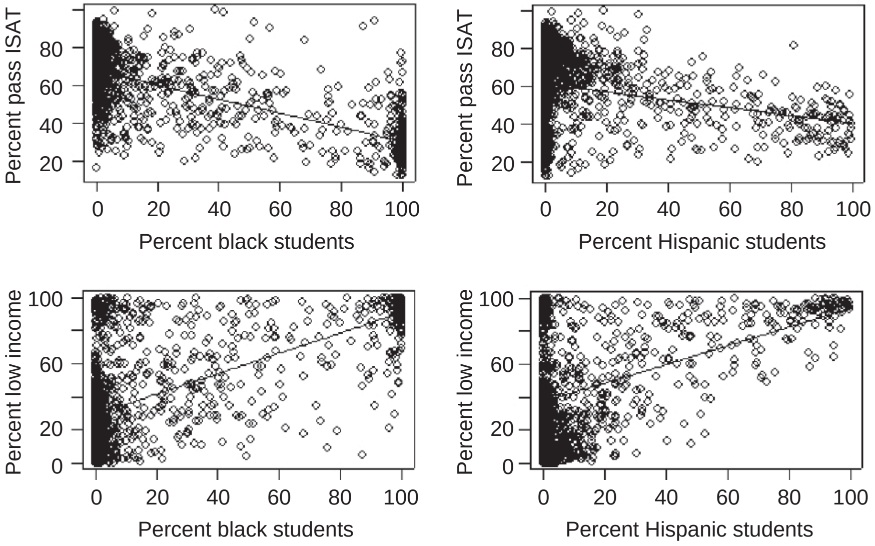


Fig. 15.8 The top two plots illustrate the decline in percent pass as the percentages of black and Hispanic students increase. The lower two plots show that there is a strong relationship between percent minorities and percent low income. The superimposed lines are the weighted least-squares fits to the data; although a linear relationship might not be the best representation, it illustrates the trend

the line shown superimposed on the plot. A similar trend, not quite so dramatic, is seen in the plot of percent pass by percent Hispanic students in the school. These trends are highly statistically significant, but what do they mean?

Can the decline of percent pass with increasing minorities be explained by percent low income? The bottom two plots of Fig. 15.8 suggest a strong relationship between both percent black students and percent Hispanic students with percent low income. The lines shown are weighted least-squares fits, which indicate the overall trend; although a linear model does not appear to be correct or complete, statistical analyses about the trend are not reported. The two bottom plots strongly suggest that percent low income is a possible confounding variable in the relationship between percent pass and percent of the two ethnicities.

A multiple regression with ISAT percent pass as the response variable and both percent low income and percent black or Hispanic will provide a better assessment of the relationship between percent passing and percent black or Hispanic students in the presence of the percent low-income variable. The data points for three variables would have to be plotted in three dimensions, and the model fits a plane to the scatter plot rather than a line. The data and fitted model are hard to display visually, and the results displayed as computer output are not as readily interpreted. To present analysis of these results in a way that allows for intuitive understanding, the concept of *residuals* of a linear model is introduced (Bock et al., 2009).

To demonstrate what residuals are and how they can be used, we look again at the relationship between percent passing and percent low income, shown in Fig. 15.9 with the linear trend superimposed. Points above the line represent schools with

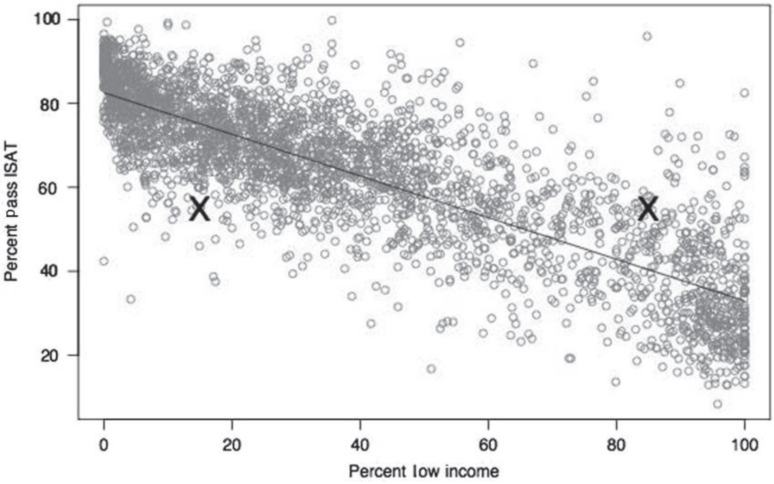


Fig. 15.9 Percent of students passing the ISAT plotted against percent low-income students, with the weighted least-squares regression line superimposed. The X on the left marks 15% low income and 55% passing the test, while the X on the right marks 85% low income and 55% passing the test

higher-than-average percent pass, *given the percent low-income students*. Similarly, points below the line represent schools with lower-than-expected percent pass, taking into account the percent low income. To illustrate, two points are marked with large Xs. The X on the left marks a school with 15% low income and 55% passing the test. Compared to other schools with about 15% low-income students, the 55% passing is a very low score. This school is at the bottom of the performance range for its income value. On the other hand, a 55% passing is quite high compared to most schools with about 85% low-income students. The X on the right marks a school doing quite well for the socioeconomic conditions at the school. This school has above-average performance for its income value. While the X on the left represents a school with a disappointingly low percent pass, the X on the right demonstrates unexpectedly high performance in spite of the fact that the two schools had the same average score.

The vertical distance from the point to the line is called a residual, with positive sign if the point is above the line and negative sign if the point is below the line. Thus, the school marked with the X on the left has a negative residual, and the X on the right marks a school with a positive residual. When comparing the performance of two schools, the comparison of these residuals could be said to be fairer than a comparison of the raw scores. The residuals are the schools' score, *corrected for percent low income*.

The plots in Fig. 15.10 show the residuals (percent passing ISAT corrected for percent low income) plotted against percent black and percent Hispanic students with the least-squares fits superimposed. We notice that the residual values range from approximately −10% to 40%, rather than −15% to 100% in Fig. 15.9. The trends are much less steep than those in the plots of raw scores, the slope is negative for black students, and the slope is positive for Hispanic students. The positive slope for Hispanic students indicates that the percent pass is increasing with increasing percent Hispanic students when percent low income is controlled for in the model. In other words, among schools with a given percent of low-income students, the

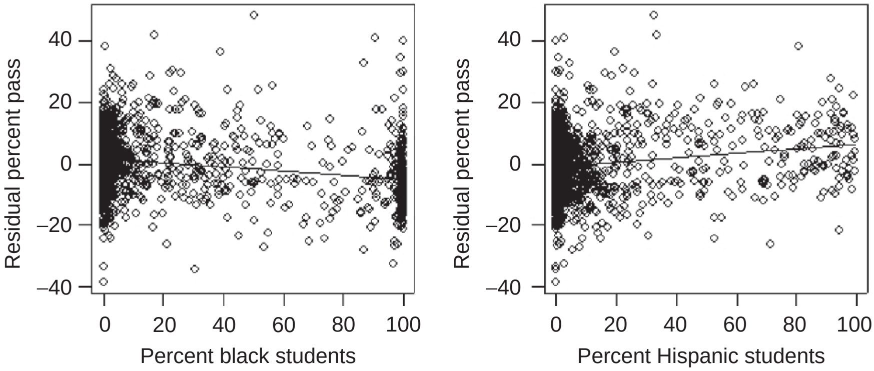


Fig. 15.10 The residual percent passing ISAT (corrected for percent low income) is plotted against percent black and percent Hispanic students, with the trends superimposed

percentage of some minority students does not substantially affect the percent passing the ISAT. The apparent relationship between test performance and proportion of minority students is confounded by percent low income.

15.2.3 Discussion

Learning, schooling, and teaching are amongst the most common experiences in modern democratic societies, but they are also amongst the most complex processes encountered by people (Bellamy & Goodlad, 2008). Being commonplace tends to encourage the use of simple models to illustrate relationships and afford explanations. Clearly, this can be misleading. Several instances of confounding using the ISAT data have been illustrated. In each case, an apparent relationship changes drastically when competing influences are disentangled and a covariate is included in the model. Each instance of confounding has a logical explanation, and the model with more covariates is more representative of the underlying relationships than the alternative with fewer variables. In the larger model, the relationship between the response and main predictor of interest is *corrected for* the effects of the covariates. In different language, the covariates are *controlled for* when assessing the relationship between the response and main predictor of interest.

While it is not certain that the larger model is not being affected by other confounders, there is confidence that such a model controlling for important covariates is in a sense better than one that does not. If these data will be used to inform policy, a better understanding of the relationships is important. For example, should minority students be targeted for special programs or, instead, should low-income students be targeted? The model with both predictors tells us that the percent low income is by far the stronger predictor of the school's percent passing the ISAT. We also found that the dramatically lower scores for Chicago schools are explained by substantially higher poverty levels for children in these schools. The fact that for a fixed poverty level Chicago schools have *higher* passing rates, on average, should prompt a scrutiny of the deleterious effects of poverty on learning—instead of trying to find problems with the teachers and administrators of Chicago schools.

15.3 Ranking States by Average SAT Scores

The SAT® (College Board, n.d.) is one of two standardized tests taken by secondary school students who hope to go to college or university in the United States. Postsecondary institutions use the scores as part of determining which students will be enrolled. The other standardized test used by American colleges and universities is the ACT® (n.d.). Each year the average SAT scores by state are ordered—and the rankings published in newspapers, often generating editorial columns in states near the top and near the bottom. The average scores rankings by state for a recent year are shown in Table 15.1. The range of average scores is quite large: the average score for North Dakota of 1,107 is 263 points higher than for South Carolina of 844, and Iowa at 1,099 is 245 points higher than Georgia at 854. Why are there such huge spreads? Are the students in North Dakota and Iowa smarter or more motivated than the students in South Carolina or Georgia? Should Oregon (rated 25th) and Arizona (rated 26th) be considered average states when it comes to education? Remember, Ockham’s Razor applied to statistics—the tendency to use simple models to illustrate relationships and explanations—may be problematic. A statistical analysis can help us find relationships between scores and state attributes to further the understanding of what the rankings really tell us.

Several readily available variables are obvious predictors of average SAT scores: poverty measures, average teacher salary, average student/teacher ratios, state education spending, percent of eligible students taking the examination (PESTE), etc. It turns out that the strongest predictor of average SAT score is the last of these variables. The strong relationship between average SAT score and PESTE is shown in Fig. 15.11. Average SAT score drops dramatically over the range of smaller percentages taking the examination, then levels out, and possibly increases again. A quadratic trend would seem to be a very reasonable choice for a model.

The explanation for this relationship is that states where the larger universities require only the ACT will have low percentages of students taking the SAT. Therefore, a nonrandom, situation–choice factor is at play. Those students that do take the SAT are typically applying to schools out of state, and it seems reasonable to conclude that many of these would be stronger students. Hence, average SAT scores decrease strongly with PESTE, to about 60–65%, after which there appears to be a slight increasing trend. If states with very high percentages taking the SAT have also a higher percentage of motivated students, this would explain the increase in average SAT score at the high end of the range of PESTE. It seems that a fairer comparison of states would have to account for PESTE in the ranking. The PESTE for top-ranking North Dakota is 5% while the PESTE for second-lowest-ranked Georgia is 65%, close to the vertex (lowest point) of the parabola.

Table 15.1 Average SAT scores by state for a recent year, ranked from highest to lowest (ranking, state, and score)

1. ND 1,107	11. TN 1,040	21. KY 999	31. NV 917	41. TX 893
2. IA 1,099	12. MS 1,036	22. CO 980	32. MD 909	42. NY 892
3. MN 1,085	13. MI 1,033	23. ID 979	33. CN 908	43. FL 889
4. UT 1,076	14. AL 1,029	24. OH 975	34. MA 907	44. HI 889
5. WI 1,073	15. OK 1,027	25. OR 947	35. CA 902	45. RI 888
6. SD 1,068	16. LS 1,021	26. AZ 944	36. VT 901	46. IN 882
7. KS 1,060	17. NM 1,015	27. WA 937	37. NJ 898	47. PA 880
8. NE 1,050	18. MT 1,009	28. NH 935	38. DE 897	48. NC 865
9. IL 1,048	19. AK 1,005	29. AK 934	39. ME 896	49. GA 854
10. MO 1,045	20. WY 1,001	30. WV 932	40. VI 896	50. SC 844

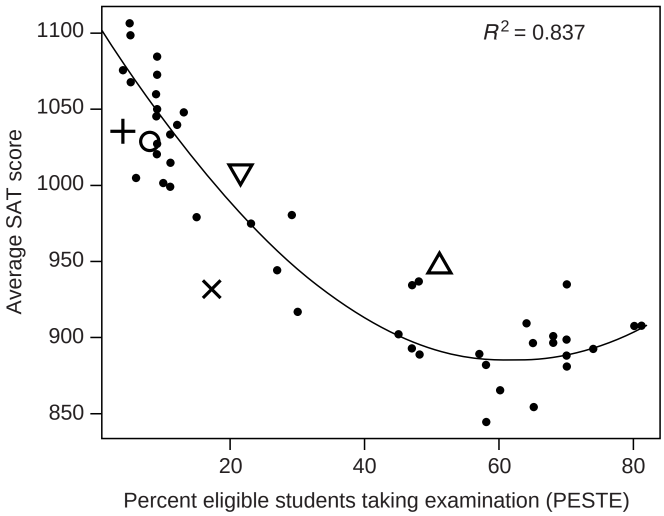


Fig. 15.11 The average SAT score plotted against percent of eligible students taking the examination (PESTE). Each dot represents a state. The weighted least-squares quadratic fit to the data is superimposed on the scatter plot of average SAT score against PESTE. The circle indicates Alabama, the plus is Mississippi, the triangle is Oregon, the upside-down triangle is Montana, and the X is West Virginia

The quadratic trend shown in Fig. 15.11 results from all regression analyses being weighted by an estimate of the number of students taking the examination. The R^2 statistic for the fit is 0.837, meaning that 83.7% of the variation in average SAT score by state is explained by the quadratic relationship with PESTE. Only 16.3% of the variation is explained by other factors.

Earlier, the residual score was used to disentangle the influence of a confounding variable in a target relationship. States above the curve have higher-than-expected SAT average, given their PESTE, and states below are lower than expected when PESTE is accounted for. The residual score for a state is the vertical distance from its point to the parabola with a positive sign for points above the parabola and negative sign for points below. States with positive residuals have higher SAT scores than other states with similar PESTE. For example, Mississippi's score of 1,036 (plus sign in Fig. 15.11) is below the parabola, but Montana's score of 1,009 (inverted triangle in Fig. 15.11) is above the parabola. Mississippi's score was lower than others at that PESTE level so did poorly compared with its PESTE peers, but Montana's average score was higher than the expected score for its PESTE group. Maybe a fairer ranking of states' SAT performances would use the residuals from the PESTE relationship rather than the raw scores.

Table 15.2 shows the rankings of the states using the residuals (i.e., the rankings once PESTE is accounted for). Note that the range of residual scores is less than half of the range for the raw scores. This indicates that there is less variability over the states once the effect of PESTE is removed from the model. The order of states

Table 15.2 Rankings of states using residuals from the quadratic fit of SAT score on PESTE (ranking, state, and residual)

1. OR +56	11. MD +25	21. HI +4	31. TX −4	41. NV −27
2. NH +47	12. IL +24	22. CT +3	32. UT −5	42. LA −27
3. WA +42	13. VT +15	23. OH +2	33. SD −6	43. GA −30
4. MN +37	14. KS +12	24. NE +2	34. FL −6	44. ID −34
5. AK +37	15. VI +12	25. CA +1	35. PA −8	45. KY −37
6. ND +33	16. DE +11	26. RI 0	36. AZ −12	46. SC −41
7. CO +32	17. NJ +10	27. NY 0	37. NC −19	47. WY −41
8. MT +27	18. ME +10	28. IN −3	38. NM −21	48. MS −45
9. IA +25	19. TN +10	29. MI −3	39. OK −21	49. AR −62
10. WI +25	20. MA +4	30. MO −4	40. AL −25	50. WV −70

Note: The average SAT score and the residual are shown for each state. For example, the average SAT score for Oregon was 56 points higher than the expected average score, given its PESTE.

has also changed quite dramatically with Oregon (marked with a triangle in Fig. 15.11) now on top. This state has a large percentage of students taking the examination as well as a very good score compared to other states with similar PESTE. North Dakota (the point on the top left of Fig. 15.11) drops only to ranking 6; its average SAT score is still high even when compared to other states with similar low PESTE value. West Virginia (marked with an X on Fig. 15.11) lands on the bottom in spite of a fairly respectable raw average score; similarly, the relatively high raw average SAT score for Arkansas is no longer impressive when compared with other schools with such a low PESTE. Alabama (circle in Fig. 15.11) and Mississippi (plus sign in Fig. 15.11) fall to ranking 40 and 48 from 16 and 14, respectively. Both states have SAT averages that are below what is expected, given the low PESTE value.

Are the new rankings *fair*? It could be argued that any analysis of standardized test scores must take into account the effect of poverty versus affluence. We would expect more affluent states to have, on the whole, higher SAT averages. It might not be considered fair to compare the SAT average for a relatively affluent state to that for a state with a high poverty level and more families with fewer resources. Figure 15.12 shows the residuals from the previous quadratic regression plotted against the poverty rate indicated by the percent of families living at or below the official poverty level. A decreasing trend is apparent, showing that, on average, less affluent states have lower SAT scores after PESTE is accounted for. Perhaps a ranking of states that takes into account poverty percentage as well as PESTE would result in fairer comparisons. Note that the percentage of families living at or below the poverty level varies between a little less than 8% (New Hampshire and Minnesota) to more than 20% (Louisiana).

A weighted multiple regression of SAT average on PESTE and percent of families living below the poverty level produces an R^2 of 0.888, indicating that 88.8% of the variation in SAT average can be explained by PESTE and poverty level. Both predictors are highly significant ($p < .0001$). The residuals from the more comprehensive model may be used to compare state scores, with the effects of these

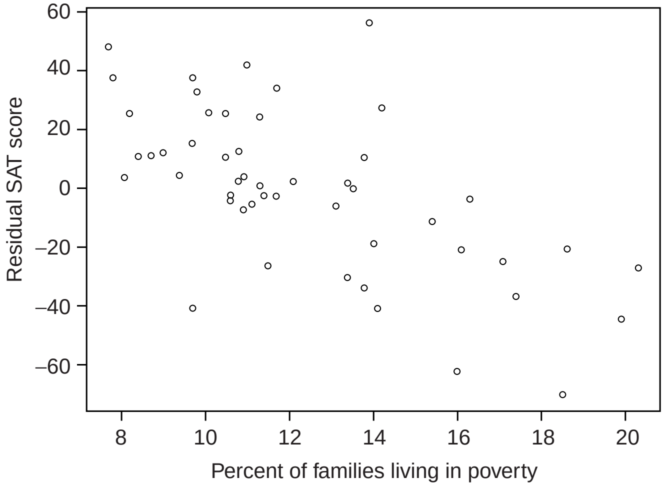


Fig. 15.12 The residuals from the quadratic regression of average SAT score on PESTE, plotted against the percent of families living in poverty. A roughly linear decreasing trend is apparent

Table 15.3 Rankings by SAT score with effects of percent taking examination and percent poverty removed (ranking, state, adjusted residual)

1. OR +59	11. MN +13	21. KS +2	31. AL −6	41. SD −13
2. WA +31	12. TN +13	22. CA +1	32. OK −7	42. UT −13
3. NH +31	13. NY +12	23. RI 0	33. FL −8	43. KY −17
4. MT +30	14. TX +9	24. VI −2	34. NE −8	44. GA −24
5. ND +29	15. MD +7	25. DE −2	35. MO −9	45. ID −32
6. AK +20	16. LA +7	26. NJ −3	36. PA −10	46. SC −34
7. IL +15	17. VT +6	27. AZ −4	37. MS −10	47. NV −37
8. CO +14	18. ME +5	28. CT −4	38. MI −10	48. WV −47
9. WI +14	19. NM +5	29. HI −5	39. NC −12	49. AR −47
10. IA +14	20. MA +2	30. OH −5	40. IN −12	50. WY −56

Note: For each state, the raw average SAT score, the PESTE, the percent of families living in poverty, and the residual from the multiple regressions are shown.

two important predictors removed. States with positive residuals have higher-than-expected SAT averages given the poverty level and PESTE for that state. States with negative residuals have lower-than-expected SAT averages compared to other states with similar PESTE and poverty percentage.

The rankings of states using the residuals of the multiple regression are shown in Table 15.3. The order of the states is not changed drastically from Table 15.2, but there are some interesting observations. Alabama and Mississippi, both with higher-than-average poverty levels, move up to ranks 31 and 37 from 40 and 48, respectively. Oregon stays at the top as the poverty level is also higher than average in this state. Wyoming moves to the bottom; the poverty level is low in this state. When Wyoming is compared with others in the same poverty level and PESTE range, the average SAT score is low although the raw score is higher than average. Wyoming's original ranking was 29 (Table 15.1). This analysis has produced new state rankings of SAT scores with the effects of the PESTE variable and a measure of state affluence taken into account. These are arguably fairer than the original rankings, but there are other variables that may change the state rankings further if they are included in the regression model.

15.4 Closing Remarks

As mentioned in the introductory chapter of this book, the quality Gold Standard for educational research should encourage new methods and evolve from observational designs to experimental ones. Researchers and policymakers should use caution in interpreting data as even the most straightforward comparisons can be misleading. Accounting for covariates—sometimes called lurking variables for their unknown existence in the shadows—can change the nature of relationships between variables, and including more covariates in a statistical model gives a truer and more comprehensive representation of significant effects. Finally, it is important to keep in mind that the variables we can readily measure are not likely to be the most important predictors of performance. As Albert Einstein (Einstein, n.d.) pointed out, "Not everything that can be counted counts, and not everything that counts can be counted." Results from statistical studies can be useful additions to the wisdom of experienced and thoughtful educators but only if used wisely and thoughtfully.

References

ACT. (n.d.). *Homepage*. Retrieved July 11, 2008, from http://www.act.org/

Bellamy, G. T., & Goodlad, J. I. (2008). Continuity and change in the pursuit of a democratic public mission for our schools. *Phi Delta Kappan, 89*(8), 565–571.

Bock, D. E., Velleman, P. F., & De Veaux, R. D. (2009). *Intro stats* (3rd ed.). Boston, MA: Pearson Education.

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology, 53*(1), 371–399.

College Board. (n.d.). *About the SAT*. Retrieved May 15, 2008, from http://www.collegeboard.com/student/testing/sat/about.html

Einstein, A. (n.d.). *quotesmuseum Homepage*. Retrieved July 11, 2008, from http://www.quotes-museum.com/author/Albert%20Einstein/1925

Entwisle, D. R., & Alexander, K. L. (1988). Factors affecting achievement test scores and marks of black and white first graders. *The Elementary School Journal, 88*(5), 449–471.

Glass, G. V. (1982). Meta-analysis: An approach to the synthesis of research results. *Journal of Research in Science Teaching, 19*(2), 93–112.

Illinois State Board of Education. (n.d.). *Student assessment*. Retrieved May 18, 2008, from http://www.isbe.net/assessment/isat.htm

Kao, G., & Thompson, J. S. (2003). Racial and ethnic stratification in educational achievement and attainment. *Annual Review of Sociology, 29*(1), 417–442.

Organisation for Economic Co-operation and Development. (n.d.). *PISA Homepage*. Retrieved June 30, 2008, from http://www.pisa.oecd.org/pages/0,2987,en_32252351_32235731_1_1_1_1_1,00.html

Petitti, D. B. (1998). Hormone replacement therapy and heart disease prevention: Experimentation trumps observation [Editorial]. *Journal of the American Medical Association, 280*(7), 650–652.

R Project for Statistical Computing. (n.d.). *Homepage*. Retrieved July 11, 2008, from http://www.r-project.org/

Smith, M. L., & Glass, G. V. (1980). Meta-analysis of research on class size and its relationship to attitudes and instruction. *American Educational Research Journal, 17*(4), 419–433.

TIMSS & PIRLS International Study Center. (n.d.). *Homepage*. Retrieved July 11, 2008, from http://timss.bc.edu/

United States National Center for Education Statistics. (n.d.). *NAEP: The nation's report card*. Retrieved July 11, 2008, from http://nces.ed.gov/nationsreportcard/

Yore, L. D., Pimm, D., & Tuan, H.-L. (2007). The literacy component of mathematical and scientific literacy. *International Journal of Science and Mathematics Education, 5*(4), 559–589.

Chapter 16
Predicting Group Membership using National Assessment of Educational Progress (NAEP) Mathematics Data

David A. Walker and Shereeza F. Mohammed

Since 1969 in the United States, the federally mandated National Assessment of Educational Progress (NAEP) has been used to assess the condition of student learning at the state and national levels, in particular subject areas, and in specific grades and/or ages (US National Center for Educational Statistics [NCES], n.d.-a). In the 1990s, academic achievement in various disciplines derived from NAEP scores were tracked by state and measured via the percentage of students at or above the levels established in a three-tiered scoring model: basic, proficient, or advanced (Hombo, 2003).

In concert with the No Child Left Behind Act of 2001 (NCLB, 2002), the US Congress decreed NAEP as the Nation's Report Card, to be used to indicate student-achievement score trends in academic areas such as mathematics, reading, science, and history among states in nationally representative samples (US NCES, n.d.-a). Results derived from NAEP data aggregated at the state level pertaining to student achievement—which in the latter years of its nearly 40-year existence have been employed as a means toward ascertaining accountability affiliated with high-stakes testing, such as NCLB—have yielded mixed interpretations and perceptions (see the American Educational Research Association, 2000, definition for a standard interpretation of the term *high-stakes testing*). For example, Hanushek and Raymond (2006) found that NAEP mathematics scores provided some positive evidence for accountability, whereas Amrein and Berliner (2002) determined the contrary with various state-level NAEP scores in certain academic areas. Darling-Hammond (2007) contended that NAEP, in areas such as mathematics, did not measure higher-order cognitive domains but instead "measures less complex application of knowledge" (p. 319). This lack of NAEP's application of knowledge via measuring—for instance, problem-solving skills—is linked to multiple-choice-type tests used for accountability purposes under NCLB that do not allow for the assessment of student achievement with the following question: "What can students do with what they have learned?" (p. 319).

D.A. Walker
Northern Illinois University

S.F. Mohammed
Florida Atlantic University

M.C. Shelley II et al. (eds.) *Quality Research in Literacy and Science Education,*

Yet, given its prominent role within the standardized testing and accountability movements in the United States, NAEP's use in the analysis of research pertaining to state-level student achievement is evident. As Dorn (2006) noted about a foremost function of NAEP data, "Research using the aggregate-level data has expanded both our knowledge of accountability's effects and the questions that are worth investigating" (p. 2). Hombo (2003) remarked that "NAEP is widely considered to be the gold standard of educational achievement survey assessment" (p. 62). Finally, Linn (2001) stated that "Comparative data are needed to evaluate the apparent gains [in student achievement]. The National Assessment of Educational Progress provides one source of such data" (p. 3). Thus, an intent of this research was to use aggregate-level NAEP mathematics data in a predictive discriminant analysis (PDA) model to investigate plausible measures that may assist in adjusting or customizing adequate yearly progress (AYP) targets to schools serving impoverished communities so that they may raise their level of achievement and meet established, high-stakes, accountability measures.

16.1 Elementary and Secondary Education Act and National Assessment of Educational Progress

Over the last 35 years, NAEP studies, as well as others pertaining to achievement testing, have associated lower performance to vulnerable sections of the population affected by poverty (Coleman, 1987). Over this same time period, the federal government used Title I programs (first launched in the War on Poverty) as part of the Elementary and Secondary Education Act of 1965 (ESEA, 1965) as aid for the disadvantaged (Jennings, 2001). In the early years of Title I programs during the 1960s and 1970s, there was little evidence that "the achievement gap between at-risk students and their more advantaged peers" was closing (Borman & D'Agostino, 2001, p. 49). This was due, in part, to ineffectual policy implementation. Consequently, Congress and the US Department of Education (US ED) were instrumental in evolving Title I interventions toward more effective implementations and evaluations (Borman & D'Agostino).

A policy information report from the Educational Testing Service on achievement gains from Grades 4 to 8 investigated growth using NAEP data (Coley, 2003). The study looked at two cohorts of Grade 4 students. The first, called the Class of 2000, was tracked by comparing their performance in 1992 in Grade 4 with the gains made in 1996 in their Grade 8 year. The second cohort was the Class of 2004, composed of students in Grade 4 in 1996 and Grade 8 in 2000. The 2004 cohort displayed an overall gain of 50 points in NAEP mathematics scores. The Class of 2000 also performed similarly, and there was no statistically significant difference with respect to growth between the two cohorts. However, an additional finding not only pointed to the gap between students eligible for free and reduced-price lunch and those who were not but also that the gains were less for the disadvantaged group (Coley). These studies suggest that many groups served by Title I programs (e.g., free and reduced-price lunch) showed small gains in achievement.

16.2 Accountability, No Child Left Behind, and Adequate Yearly Progress

According to the NCES publication *State Education Reforms* (US NCES, n.d.-b), by the 2004/05 academic year, 48 of the 50 states had mathematics assessments aligned to their state standards. Further, 46 states had mathematics assessments at the elementary, middle, and, secondary school levels, which shows the high number of compliant states using assessments to monitor student performance on their state standards as prescribed by the NCLB Act. However, NCLB also requires states, districts, and schools to be held accountable for student performance, especially for schools receiving Title I federal funding. In order to monitor performance, schools are to be evaluated on their AYP on statewide assessments as defined by individual states (US NCES). In many states, high-performing schools are rewarded with extra money and low-performing schools with assistance and a series of consequences over a 4-year period if no improvement is evident.

According to the NCLB Act, if a Title I school fails to attain AYP, the district identifies it for an improvement plan and technical assistance. During this time, parents can opt to either move their children to higher-performing schools or receive supplemental educational services. If no improvement is seen in the school's performance at the end of the second year, then the following corrective actions can occur: failure-related school staff being replaced, a new curriculum being implemented, management authority in the school being decreased, outside experts used to advise the school, extension of the school day or year, or the organizational restructuring of the school (US ED, 2003).

If no improvement occurs following 1 additional year, then changes in governance in accordance with state law proceeds. Such measures involve closing and then reopening the school as a public charter school, replacing the principal and staff, having a private management company operate the school, or the school being run directly by the state in a state takeover (US ED, 2003). During the 2003/04 academic year, 19,644 schools nationwide did not make AYP. Of that total, 11,008 schools were identified as low performing and in need of improvement by their districts. While most of these schools were located in California, almost all the states (98%) were affected by low-performing schools (US NCES, n.d.-b). By extension, these schools will also need to show improvement or face corrective action.

According to Slavin (2001), because performance correlates with socioeconomic status, schools serving impoverished neighborhoods may be targeted unfairly by policies where progress is determined largely by test scores. Instead, he suggested that growth indicators be added to the criteria for school improvement. In fact, several studies have investigated performance growth. From 1996 to 1999, the US ED contracted Policy Studies Associates to assess the Title I program with a study entitled *The Longitudinal Evaluation of School Change and Performance*. This study investigated the progress of Grades 3 to 5 students in 71 high-poverty schools. These students, who were recipients of Title I funding, were also eligible for free and reduced-price lunch. The study found that student performance on the mathematics portion of the Stanford Achievement Test-9 indicated some growth, but students were still performing below the national norm at the end of Grade 5 (US ED, 2001).

Ultimately, the aforementioned AYP results illustrate that almost all states have low-performing schools that may be vulnerable to the dire consequences of corrective action and restructuring as prescribed by the NCLB Act. However, studies have shown small achievement gain among Title I students. Since some states are using growth indicators as part of their AYP, the current study proposed to identify variables outside the classroom that would be viable classifiers of gain or no gain in NAEP mathematics scores.

16.3 Methods

16.3.1 Variables of Study

Previous research has used many of the variables listed below in studies that looked at the relationship between measures of resource inputs and student achievement (Chubb & Hanushek, 1990; Hanushek, 1996; Hedges, Laine, & Greenwald, 1994). The scope of this research involved independent and dependent variables that were measured at the state and/or federal levels and often outside the control domain of schools and classrooms. For example, the numerous independent variables considered for the study were categorized as resource inputs, which are usually derived from state and/or federal sources.

Initially, in this theorized model, there were 13 independent variables per the 35 states in the study that had both NAEP mathematics scores from 1992 and 1996: state revenue, instructional expenditures, support services expenditures, noninstructional expenditures, per pupil total expenditures, per pupil instructional expenditures, per pupil support services expenditures, per pupil noninstructional expenditures, percentage of students with individual education plans, percentage of students receiving limited English proficiency (LEP) services, Title VII state grant program funding for LEP students, percentage of students eligible for free and reduced-price lunch, and state median household income. Thus, given these variables, the early phases of this study focused on the more global research question: In terms of their relative contribution to classification accuracy, which of the variables of study, if any, predict if a state on the NAEP mathematics test from 1992 to 1996 is going to be a no-gain-score state or a gain-score state?

16.3.1.1 Sample

Multistage, sampling-based, quantitative NAEP data from the 1990s were used in this study because of their relationship in terms of historical context and apparent relevance within the perspective of federally based movements pertaining to standards, testing, and accountability in education. These standards-based movements were derived from federal initiatives such as the ESEA Act (1965), the Goals 2000: Educate America Act (1994), and the Improving America's Schools Act (1994). The NAEP data from the early to mid-1990s are relevant to use in this study given that the aforementioned accountability plans coalesced at about the same time when these data were collected. The cumulative effect of these national plans was an increased role of the federal government in educational accountability, standards, and testing—culminating in the NCLB Act of 2001, or the last reauthorization of the ESEA—and a paradigm shift in education research to a quantitative, data-driven emphasis termed *scientifically based research*.

16.3.1.2 Diagnostics

Upon initial diagnostic review to find the most parsimonious model, dimensionality reduction techniques were implemented via use of Statistical Package for the Social Sciences (SPSS). Properties such as variance inflation factor (VIF), tolerance, and eigenvalues within a collinearity diagnostic (i.e., at least one nonpositive value) were assessed. Of the initial 13 instructionally related and noninstructional variables considered, 8 were discarded from the original model due to very high VIF values (>10), which is an indicator of collinearity. Thus, in light of these multiple diagnostic measures, five variables (i.e., per pupil noninstructional expenditures, per pupil support services expenditures, percentage of students receiving LEP services, percentage of students eligible for free and reduced-price lunch, state median household income) were retained as components of a parsimonious, more credible model in terms of their relationship to the entities of interest, which are the 35 states whose 1992 and 1996 mathematics data were analyzed.

16.3.2 Research Questions

After the diagnostic phase of the study, two refined research questions emerged:

1. How accurately can group membership be predicted, in terms of no-gain scores and gain scores, from states' scores on the NAEP mathematics test from 1992 to 1996?

2. In terms of their relative contribution to classification accuracy, how well can state median household income and the percentage of students eligible for free and reduced-price lunch per state predict group membership?

16.3.3 Limitations

Because individual-level data, as the unit of analysis, were not available to the researchers, the final model used observational data aggregated at the state level, which is emblematic of research conducted on NAEP data (cf. Amrein & Berliner, 2002; Nichols, Glass, & Berliner, 2006; Schafer, Liu, & Wang, 2007). Further, using generalizability theory, Brennan (1995) found that aggregated scores tended to be more reliable than scores at the individual level. As well, Klein, Benjamin, Shavelson, and Bolus (2007) determined this inclination—that group mean scores were more reliable than scores from individuals—based on analyses derived from the Collegiate Learning Assessment. As Brennan noted, “Aggregation may well lead to a sizeable decrease in error variance” (p. 395).

To be sure, aggregate-level data may present some degree of threat to the research design, such as in the areas of ecological validity and sample exclusion bias (Dorn, 2006). Further, there is some caution in relation to the interpretation of results based on a hierarchical data structure in the areas of independence of observations and error (Osborne, 2000). For example, in terms of the latter, with concepts such as Huberty’s (1994) *I* statistic effect size measure, the totality of the proportional reduction in error would be greater at an individual or a district level than at a state level. However, concerning the challenge of the unit of analysis issue, Cronbach (1976) and Knapp (1977) noted that prominence should not be given to the fact that data are not available at certain levels of analysis. Instead, what should be of consequence is determining if the correct research question was posed at the right level of analysis. We believe the answer in this study is *yes*, where the research questions dealt with state median household income and the percentage of students eligible for free and reduced-price lunch—both of which are typically accumulated and apportioned at the state level.

Finally, it has been noted that collapsing a variable into a dichotomy (i.e., the dependent variable in this research) may yield a final variable of study that is less dependable (Cliff, 1987). However, dummy coding the dependent variable did not violate a normal distribution assumption, which is shown in Table 16.1, where the marginal distributions do not appear to be disproportionate (Schumacker, Mount, & Monahan, 2002).

Table 16.1 Linear external classification: Cross-validation L-O-O

	No gain	Gain	Total
No gain	14 (70.00%)	6 (30.00%)	20
Gain	4 (26.70%)	11 (73.30%)	15

Note: 71.43% of cross-validated grouped cases correctly classified.

16.4 Data Analysis and Results

Using the resampling cross-validation technique of the leave-one-out (L-O-O) rule or U method (Huberty, 1994; Lachenbruch & Mickey, 1968), the subsets of all possible variables were analyzed for the purpose of parsimony and to increase the cross-validation accuracy of the proposed model (Lieberman & Morris, 2004; Morris & Meshbane, 1995). Morris and Meshbane's FORTRAN program (Huberty) for an all-subset analysis to yield the best L-O-O hit rate for predictor selection, or $2^p - 1$ where p are the predictors, was conducted. Of the remaining five variables considered, three predictors were deleted that did not contribute to high predictive accuracy. Thus, only the percentage of students eligible for free and reduced-price lunch per state (FRELCH) and state median household income (AVGINC) were retained as components of a parsimonious model and were not correlated substantially ($r = -.396$); that is, there were five predictor variables for the two-group problem, which meant that there were 31 all-possible subset analyses (i.e., $2^5 - 1$). When the number of predictors in the best subset emerged, the maximum hit rate increased by 3.00% to 71.43% from the second best hit rate of 68.57% with three predictors; thus, parsimony with increased accuracy was achieved. Other variations within the all-possible subset analyses yielded a maximum hit rate range of 60.00% to 68.57%.

With the L-O-O method, it has been noted that a minimum sample size can be calculated as $N = 3kp$ or a large sample size of $N = 5kp$, where k = the number of groups and p = the number of predictors, and the 3 or 5 derived from the n/p ratio (Huberty, Wisenbaker, & Smith, 1987). Therefore, from the original five variables used in the all-possible subset analyses, $3 \times (2 \times 5) = 30$ or a minimum sample size of 30 is needed to use with the L-O-O method in this study. Multivariate normality of the data and equality of covariance matrices of the groups were met, with a normal-based rule establishing normality via a review of normal probability plots for data in each of the two groups (Huberty & Lowman, 1998). A significant degree of discrimination separating the two groups was confirmed, which indicated that the two variables of study, FRELCH and AVGINC, were the strongest for the separation on the construct of no-gain score or gain score.

Although it was assumed that the NAEP 1992 and 1996 mathematics tests were randomly parallel due to the fact that this test has a constant score scale across time and age, to heed to caution, the state raw mathematics scores from 1992 and 1996 were converted to z scores for the PDA model to put them into a common metric, with mean = 0 and standard deviation = 1. Based on the functions at group centroids, the mean discriminant score for the no-gain-scores group was −0.443 and the gain-scores group was −0.590. The average centroid value was 0.074 (i.e., −0.433, + 0.590 = 0.157/2). The cut point chosen for the two groups was 0, which is very close to the average centroid value of 0.074. Thus, states with z scores less than, or equal to, 0 were grouped as no-gain scores and coded as a 0, and states with z scores greater than 0 were grouped as gain scores and coded as a 1 (cf. Ananth & Kleinbaum, 1997; Schumacker et al., 2002).

The L-O-O rule (Huberty, 1994; Lachenbruch, 1967) was established as a bias-correction method for classification error rates. L-O-O took one subject out of the sample and developed a rule on the other 34 subjects and then took another subject out and developed a rule on the remaining subjects, and so on. This procedure was applied to all subjects in the sample so that rules were built on all 35 iterations.

For model accuracy based on the 35 states in this study, the no-gain-score group had 14 states or 70.0% (90% confidence interval (*CI*) for a binomial parameter = .49, .86; standard error (*SE*) = .10) classified accurately as hits, and 6 or 30.0% (*CI* = .14, .51) that were predicted as gain scores or misses. The gain-scores group had 4 states or 26.67% (*CI* = .10, .51; *SE* = .11) misclassified as no-gain score or misses and 11 or 73.33% (*CI* = .49, .90) that were predicted as gain scores or hits. In terms of total precision for all of the states, there was 71.43% accuracy (*CI* = .56, .84; *SE* = .08); that is, the model correctly classified over 71% of the states for the two-group problem, with a total group error rate estimate of 28.57% (*CI* = .16, .44).

When assessing each variable's contribution to the discriminant function, the standardized canonical discriminant function coefficients (weights) indicated that AVGINC's relative importance in predicting gain scores was 0.680, followed by FRELCH at −0.540. Predictor importance was also noted via a variable deletion method when AVGINC, for example, was taken out of the model, which produced the lowest hit rate for total group accuracy at 60.00% (cf. Huberty & Lowman, 1998). The order of the response variables' contribution toward predictive accuracy indicated how the predictor variables should be arranged. In terms of structure coefficients, the largest absolute correlation associated with the discriminant function was AVGINC at 0.861, followed by FRELCH at −0.768.

In regard to particular cases that may be fence-riders (i.e., subjects that were classified correctly but, when their probabilities were reviewed, confidence waned in terms of proper classification), the probability split between the highest group and second-highest group was established at .52/.48. Of the 35 subjects, 1 was deemed a fence-rider. Outliers were determined to be cases that had typicality probabilities less than 0.10; that is, although a subject was classified correctly with confidence, it appeared to be atypical of that group and hence garnered a low probability. Of the 35 subjects, four were estimated to be outliers. The fence-riders and the outliers were kept in the data and analyzed because omitting them may have inflated the hit rate of the model, which potentially could have yielded a model that was more accurate than in actuality.

Using a proportional chance criterion, Huberty's (1994) *Z* statistic was calculated from a FORTRAN program (J. D. Morris, personal communication, March 13, 2003) to determine if expected hit rates were exceeded (Eq. 16.1):.

$$Z = (o - e)/[e(n - e)/n]^{1/2} \tag{16.1}$$

o = observed frequency; e = expected frequency; n = number of subjects

This is a one-tailed test because there is little interest in whether the hit rate was significantly below expectation. The null hypothesis was that the hit rate is what would be expected by chance (e.g., .50 × 20 + .50 × 15 = 17.5). The alternative hypothesis was that the present hit rate is better than chance expectance. With an observed hit rate of 25, the Z of 2.42 ($p < .01$) for the total sample occurred because this hit rate was above expectation, which offers some evidence that the null should be rejected or that classification by the discriminant function resulted in more hits than random assignment by prior probabilities. This result was found, as well, when the Z value for each group was examined. The Z value for the gain-score group was quite large and statistically significant at 2.39 ($p < .01$), while the Z for the no-gain-score group was a more modest 1.16 and not statistically significant ($p > .05$). The reason this model appeared to be better than chance was that it was quite good at predicting the gain-score group, but it was less sufficient at predicting the no-gain-score group based on states' NAEP mathematics scores. The percentage improvement over chance for the gain-score group was 53.33% and for the no-gain-score group was 30.00%. The percentage of improvement over chance for the total sample was a substantial 41.67%. Thus, the classification of the two groups was exceptionally better, by over 41%, than would have been accomplished by chance.

To add to this argument from a different perspective and also to address the issue of the intermediate inequality of group sizes, the model was looked at via a maximum-chance criterion (max $[q_1, q_2]$) (Huberty, 1994). The maximum-chance criterion assigned all the subjects to the largest group for this study, the no-gain-score group, as a criterion for a hit rate better than chance. The Z value was 1.71 and statistically significant ($p < .05$), which meant that the model's hit rate was better than chance. Further, the percentage improvement over chance for the total sample was quite large, again, at 33.33%.

Huberty's (1994) effect size measure, the I statistic (Eq. 16.2), was calculated to determine:

$$\begin{aligned} I &= (1-e) - (1-o)/1-e \\ &= o-e/1-e \end{aligned} \tag{16.2}$$

the percentage correctly classified exceeding chance. The no-gain-score group had an $I = .400$, the gain score group had an $I = .466$, and the total model had an $I = .428$. Previous research (Huberty & Lowman, 2000) conducted on I indicated that these values should be regarded as having a high effect in terms of their ability to measure proportional reduction in error—meaning, for instance, that the total model had about 43% less misclassifications than would have occurred if just classified by chance.

16.5 Discussion

The nature of educational governance in the United States is uniquely decentralized, creating a loosely coupled system of federal, state, and district input that is steeped in local control. There has been a gradual movement from Goals 2000 in the 1980s to the Improving America Schools Act of 1994 that has increased the role of the federal government in accountability, standards, and testing. The NCLB Act exemplifies a top-down control of national and state standards, testing, and accountability, which often conflicts with state educational needs and preferences (Cooper, Fusarelli, & Randall, 2004). This potential limitation is a primary reason why federally based policy implementation often encounters many obstacles and requires a strategic approach to navigate the unique environment existing in every state.

To be sure, the problem of schools not meeting AYP and facing the resulting consequences affects each state, especially those areas with high proportions of need-based students, such as impoverished, LEP, or disabled students (Lyons & Algozzine, 2006). Therefore, an underlying purpose may be for NAEP to assist in confirming AYP, which states select as targets, to the ultimate goal of having 100% students attain proficiency by 2014. However, in the face of this very lofty target, the question remains as to whether Title I students, or all student groups for that matter, can reach proficiency by 2014. Simulation research (Lee, 2004) indicates that the answer leans toward *no*, with the additional predicament of "the risk of massive school failure due to unreasonably high AYP targets for all student groups" (p. 1).

The findings of this study indicate that median household income and the percentage of students eligible for free and reduced-price lunch can be used as alternative measures in adjusting or customizing AYP targets to schools serving impoverished communities. The use of alternative measures (e.g., adjusting the safe harbor provision percentage, measuring growth via value-added models, or employing effect size measures when analyzing student achievement scores) to meet AYP has been suggested by Linn (2003, 2005). These alternative measures may allow for AYP targets to be attained more realistically, permitting schools and their stakeholders to realize continuous improvement if incremental increase in expected growth is monitored and maintained. Consequently, Title I schools may attain realistic growth even if it is small when compared to their more advantaged peers, as has been illustrated by prior studies where financially disadvantaged schools tend to reach AYP at much lower rates than non-impoverished schools (Choi, Seltzer, Herman, & Yamashiro, 2007).

Finally, the Education Policy Studies Laboratory (2005) at Arizona State University noted that "the impact of poverty … on a school's ability to achieve AYP is not addressed [in NCLB and] … realistic standards linked to external expectations and grounded research" (p. 2) are not transpiring. Thus, via the current research, it is hoped that the use of median household income and the percent of students on free and reduced-price lunch—as alternative measures to select attainable AYP growth targets in addition to NAEP's role as the nation's report card—will serve to inform education policy in its data collection and use. These measures may then add to other indicators of student achievement to provide feedback to assist educators and legislators to fine-tune related policy instruments so that the ultimate goal of student proficiency can be attained realistically by all students.

Acknowledgement This chapter is published with permission from *The Educational Forum*, where a previous, truncated version of it appeared in *The Educational Forum* (2008).

References

American Educational Research Association. (2000). *AERA position statement on high-stakes testing in pre-K – 12 education*. Retrieved October 7, 2007, from https://www.aera.net/policyandprograms/?id = 378&terms = AERA + position + statement + concerning&searchtype = 1&fragment = False

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Educational Policy Analysis Archives, 10*(18). Retrieved from http://epaa.asu.edu/epaa/v10n18

Ananth, C. V., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology, 26*(6), 1323–1333.

Borman, G. D., & D'Agostino, J. V. (2001). Title I and student achievement: A quantitative synthesis. In G. D. Borman, S. C. Stringfield, & R. E. Slavin (Eds.), *Title I: Compensatory education at the crossroads* (pp. 25–57). Mahwah, NJ: Lawrence Erlbaum.

Brennan, R. L. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement, 32*(4), 385–396.

Choi, K., Seltzer, M., Herman, J., & Yamashiro, K. (2007). Children left behind in AYP and non-AYP schools: Using student progress and the distribution of student gains to validate AYP. *Educational Measurement: Issues and Practice, 26*(3), 21–32.

Chubb, J. E., & Hanushek, E. A. (1990). Reforming educational reform. In H. J. Aaron (Ed.), *Setting national priorities: Policy for the nineties* (pp. 213–248). Washington, DC: Brookings Institution Press.

Cliff, N. (1987). *Analyzing multivariate data*. San Diego, CA: Harcourt Brace Jovanovich.

Coleman, J. S. (1987). Families and schools. *Educational Researcher, 16*(6), 32–38.

Coley, R. J. (2003). *Growth in school revisited: Achievement gains from the fourth to the eighth grade*. Princeton, NJ: Educational Testing Service.

Cooper, B. S., Fusarelli, L. D., & Randall, E. V. (2004). *Better policies, better schools: Theories and applications*. Boston: Pearson/Allyn & Bacon.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design, and analysis*. Stanford, CA: Stanford Evaluation Consortium.

Darling-Hammond, L. (2007). The flat earth and education: How America's commitment to equity will determine our future [Third Annual Brown Lecture in Education Research]. *Educational Researcher, 36*(6), 318–334.

Dorn, S. (2006). No more aggregate NAEP studies? *Education Policy Analysis Archives, 14*(31). Retrieved from http://epaa.asu.edu/epaa/v14n31

Education Policy Studies Laboratory (2005, September 14). *Study predicts at least 85 percent of Great Lakes schools will be labeled 'failing' by 2014*. [Press release]. Retrieved from http://www.asu.edu/educ/epsl/EPRU/documents/EPSL-0509-109-EPRU-press.pdf

Elementary and Secondary Education Act of 1965. Pub. L. No. 89–10. (1965).

Goals 2000: Educate America Act of 1994. Title III. 20 U.S.C. 5801 § 302. (1994).

Hanushek, E. A. (1996). School resources and student performance. In G. Burtless (Ed.), *Does money matter? The effect of school resources on student achievement and adult success* (pp. 43–73). Washington, DC: Brookings Institution Press.

Hanushek, E. A., & Raymond, M. E. (2006). Early returns from school accountability. In P. E. Peterson (Ed.), *Generational change: Closing the test score gap* (pp. 143–166). Lanham, MD: Rowman & Littlefield.

Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). An exchange: Part I: Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher, 23*(3), 5–14.

Hombo, C. M. (2003). NAEP and No Child Left Behind: Technical challenges and practical solutions. *Theory into Practice, 42*(1), 59–65.

Huberty, C. J. (1994). *Applied discriminant analysis*. New York: Wiley.

Huberty, C. J., & Lowman, L. L. (1998). Discriminant analysis in higher education research. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 181–234). New York: Agathon Press.

Huberty, C. J., & Lowman, L. L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement, 60*(4), 543–563.

Huberty, C. J., Wisenbaker, J. M., & Smith, J. C. (1987). Assessing predictive accuracy in discriminant analysis. *Multivariate Behavioral Research, 22*(3), 307–329.

Improving America's Schools Act of 1994. Pub. L. No. 103–382, 108 Stat. 3518. (1994).

Jennings, J. F. (2001). Title I: Its legislative history and its promise. In G. D. Borman, S. C. Stringfield, & R. E. Slavin (Eds.), *Title I: Compensatory education at the crossroads* (pp. 1–24). Mahwah, NJ: Lawrence Erlbaum.

Klein, S., Benjamin, R., Shavelson, R. J., & Bolus, R. (2007). *The Collegiate Learning Assessment: Facts and fantasies* [White paper]. Retrieved October 15, 2007, from http://www.cae.org/content/pdf/CLA.Facts.n.Fantasies.pdf

Knapp, T. R. (1977). The unit-of-analysis problem in applications of simple correlation analysis to educational research. *Journal of Educational Statistics, 2*(3), 171–186.

Lachenbruch, P. A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics, 23*(4), 639–645.

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics, 10*, 1–11.

Lee, J. (2004). How feasible is adequate yearly progress (AYP)? Simulations of school AYP "uniform averaging" and "safe harbor" under the No Child Left Behind Act. *Education Policy Analysis Archives, 12*(14). Retrieved from http://epaa.asu.edu/epaa/v12n14

Lieberman, M. G., & Morris, J. D. (2004, April). *Selecting predictor variables in logistic regression*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Linn, R. L. (2001). Assessment and accountability (condensed version). *Practical Assessment, Research and Evaluation, 7*(11). Retrieved from http://PAREonline.net/getvn.asp?v = 7 &n = 11

Linn, R. L. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives, 11*(31). Retrieved from http://epaa.asu.edu/epaa/v11n31

Linn, R. L. (2005). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Education Policy Analysis Archives, 12*(33). Retrieved from http://epaa.asu.edu/epaa/v13n33/

Lyons, J. E., & Algozzine, B. (2006). Perceptions of the impact of accountability on the role of principals. *Education Policy Analysis Archives, 14*(16). Retrieved from http://epaa.asu.edu/epaa/v14n16/

Morris, J. D., & Meshbane, A. (1995). Selecting predictor variables in two-group classification problems. *Educational and Psychological Measurement, 55*(3), 438–441.

No Child Left Behind Act of 2001. Pub. L. No. 107–110, 115 Stat. 1425. (2002).

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Educational Policy Analysis Archives, 14*(1). Retrieved from http://epaa.asu.edu/epaa/v14n1

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research and Evaluation, 7*(1). Retrieved from http://www.pareonline.net/getvn.asp?v = 7&n = 1

Schafer, W. D., Liu, M., & Wang, H.-F. (2007). Content and grade trends in state assessments and NAEP. *Practical Assessment, Research and Evaluation, 12*(9). Retrieved from http://pareonline.net/pdf/v12n9.pdf

Schumacker, R. E., Mount, R. E., & Monahan, M. P. (2002). Factors affecting multiple regression and discriminant analysis with a dichotomous dependent variable: Prediction, explanation, and classification. *Multiple Linear Regression Viewpoints, 28*(2), 32–39.

Slavin, R. E. (2001). How Title I can become an engine of reform in America's schools. In G. D. Borman, S. C. Stringfield, & R. E. Slavin (Eds.), *Title I: Compensatory education at the crossroads* (pp. 235–260). Mahwah, NJ: Lawrence Erlbaum.

United States Department of Education. (2001). *The longitudinal evaluation of school change and performance in Title I schools* (Policy Studies Associates Contract No. EA 96008001). Rockville, MD: Author.

United States Department of Education. (2003). *No Child Left Behind, Accountability and adequate yearly progress (AYP)* [National Title I Directors' Conference 2003]. Retrieved June 5, 2008, from http://www.ed.gov/admins/lead/account/ayp203/edlite-index.html

United States National Center for Education Statistics. (n.d.-a). *NAEP overview*. Retrieved June 27, 2008, from http://nces.ed.gov/nationsreportcard/about/

United States National Center for Education Statistics. (n.d.-b). *State education reforms: Standards, assessment, and accountability*. Retrieved June 27, 2008, from http://nces.ed.gov/programs/statereform/saa.asp

Chapter 17
Incorporating Exploratory Methods using Dynamic Graphics into Multivariate Statistics Classes: Curriculum Development

Dianne Cook

Properly prepared and properly presented graphics often provide highly informative visualization of statistical analysis and results with education data. The applicability of graphical representations is shown, for example, through the use of repeated traces by van den Bergh et al. (see Chap. 20) in the discussion of time series analysis in this book. This chapter demonstrates how methods of statistical graphics can be applied to the study of science, literacy, and other areas of education research.

Multivariate data analysis is a course commonly offered in statistics undergraduate and graduate programs. There are many textbooks; most discuss ways to plot multivariate data in some form or another; usually a chapter is devoted to plotting methods. Unfortunately, most textbooks still focus on old methods, such as Chernoff faces, star plots, and Andrews curves, which have not been used seriously since the 1970s. A few textbooks give the topic some cursory attention and include material on static methods that are actively used, such as scatter plot matrices, trellis plots, and parallel coordinate plots. Not a single textbook discusses dynamic methods like multiple linked plots and tours. Yet these methods were invented in the early 1980s; they are commonly available in today's software and provide insight into data and theoretical concepts of multiple dimensions. Plots of multivariate data are very important—more so than for univariate data because multivariate data have more complexity and the distribution theory is less developed.

What is multivariate analysis? It is a suite of tools for describing and quantifying the relationship between multiple measured variables. Data having *p* variables and *n* cases is denoted in Equation 17.1 as:

$$X = [X_1\, X_2 \cdots X_p] = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \qquad (17.1)$$

D. Cook
Iowa State University

M.C. Shelley II et al. (eds.) *Quality Research in Literacy and Science Education,*

where X_{ij} is the element in the *i*th row and *j*th column, that is, the *i*th case and the *j*th variable. In this chapter, X_{ij} is typically real-valued. There may also be real-valued or categorical response variables $\boldsymbol{Y}_{n \times q}$ associated with $\boldsymbol{X}$, such as a vector of class labels for each data sample.

Multivariate data arise everywhere today. Here are some examples. Measurements recorded for handwritten digits might be used to generate a rule used by an electronic reader to recognize postal or mailing codes for automatic mail sorting. A spam filter can be improved by training a classifier on a sample of e-mail labeled by hand as spam or not. Music clips are clustered into groups to be loaded onto a digital music player based on their similarity on many measures of the sound. In educational studies, we might be interested in comparing performance of school districts on the basis of several measurements. A longitudinal study might record several student characteristics at many different times, allowing the examination of individual trends.

Many multivariate analyses are exploratory rather than confirmatory. Exploratory methods focus on discovery of structure in data—finding clusters, outliers, or nonlinear relationships—whereas confirmatory statistics test predetermined hypotheses. Graphics have a huge role in exploratory analyses because they provide a broad and largely unfettered view of data. Very few assumptions need to be made about the data to make a plot, and they are generally very helpful for getting some sort of overview of the whole dataset and data distribution. Even in confirmatory analyses, graphics have an important role in the initial analysis where assumptions are checked. Particularly for multivariate data, in recent years huge advances have been made in graphics research with new methods developed and applied successfully. The curriculum for multivariate analysis needs to be revised to reflect these changes. The advances make use of technology and enable analysts to interact with plots and make dynamic graphics.

This chapter describes some of the advances and makes suggestions about which material might be replaced with more current content. Several examples are provided where graphics has something to say about particular problems beyond what is possible with existing numerical methods. Pointers to the literature on graphics research are given for further reading.

17.1 Current Textbooks

This section describes material in current textbooks that is reasonably treated and also material that needs retiring.

17.1.1 Materials To Be Kept

Most textbooks describe the scatter plot matrix. Original discussions on the scatter plot matrix can be found in Tukey and Tukey (1981). The scatter plot matrix is a convenient way to lay out plots of all pairs of variables. The arrangement matches the correlation matrix, which is commonly used as a summary statistic for multivariate data. Figure 17.1 displays an example of scatter plot matrix for a dataset having six variables, and Table 17.1 shows a correlation matrix of the same six variables.

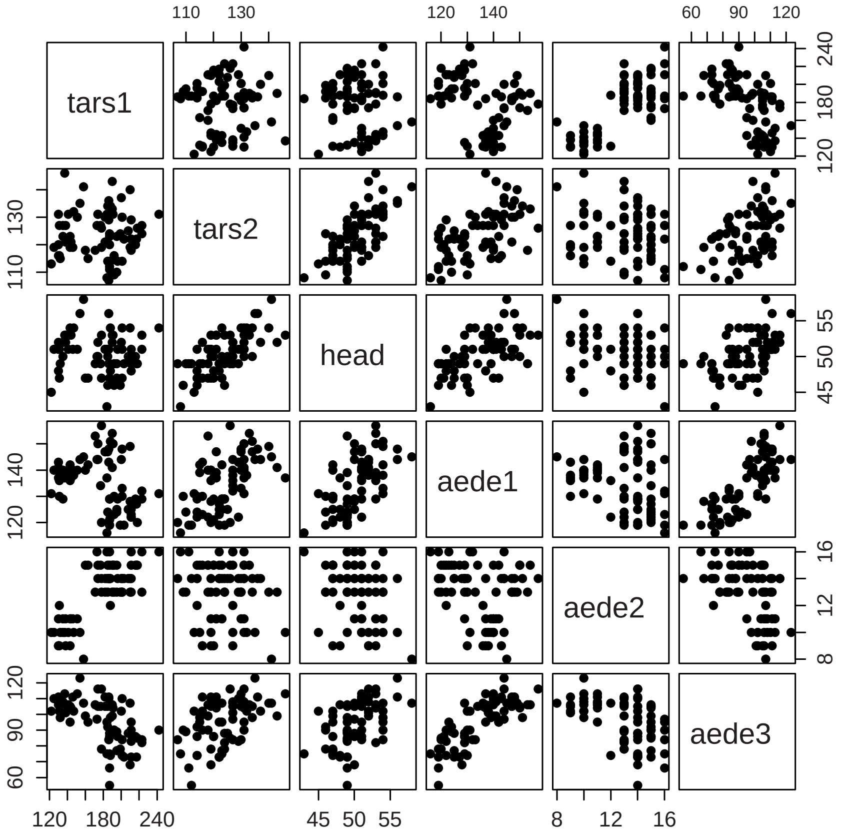


Fig. 17.1 Scatter plot matrix of the same six variables as shown in the correlation matrix in Table 17.1

Table 17.1 Correlation matrix of a dataset containing six variables

	tars1	tars2	head	aede1	aede2	aede3
tars1	1.00	0.03	−0.10	−0.34	0.78	−0.57
tars2	0.03	1.00	0.67	0.56	−0.12	0.49
head	−0.10	0.67	1.00	0.59	−0.31	0.52
aede1	−0.34	0.56	0.59	1.00	−0.25	0.78
aede2	0.78	−0.12	−0.31	−0.25	1.00	−0.48
aede3	−0.57	0.49	0.52	0.78	−0.48	1.00

A scatter plot can give more information about the association between pairs of variables than the correlation. Correlation is a measure of linear association between pairs of variables; consequently, the correlation matrix is a good, concise, summary statistic only for a select few multivariate datasets. These data are an example where the correlation matrix is not as revealing as a scatter plot matrix. The highest correlation between a pair of variables in our example data is 0.78 for tars1 and aede2. This corresponds to the 4th plot in the top row (and also the 4th plot in column 1 because the lower triangle of plots is a mirror copy of the upper triangle of plots). Correlation is not a good summary of the association between these two variables, as seen from the plot, because the main structure is clustering. Within each cluster, the correlation between the two variables is obviously much lower than 0.78. Between tars2 and head, the correlation (0.67) is a good summary of the association between these two variables because linear association is the main pattern (2nd plot in row 3 and 3rd plot in column 2) between these two variables. So generally, for multivariate data, the scatter plot matrix is a necessary summary statistic, and one should not only rely on the correlation matrix.

Multivariate textbooks usually describe profile plots, which are used to display measures collected repeatedly. The repeat variable, usually time, is displayed horizontally with measured data values vertically. The values for each subject, that is, the values in each row of data, are connected by line segments, as in Fig. 17.2. Profile plots are commonly used for repeated measures analysis or for exploring the interaction in ANOVA models. These types of plots are useful for multivariate data and are similar to the parallel coordinate plot, which are described in Sect. 17.2.

17.1.2 Materials in Need of Retirement

Some material presented in the textbooks is simply negligent in graphical value; for example, three-dimensional (3D) scatter plots (see Johnson, 1998; Johnson & Wichern, 2002) (Fig. 17.3, left). Three-dimensional graphics were developed to display the physical world and model scenes, and to create fragments of simulated reality for movies. Multivariate data can have any number of variables; therefore, visual methods cannot be limited by a 3D box. If there are only three variables, then a 3D plot, with the ability to rotate it, is critical; but the emphasis on 3D is misplaced.

Some material has simply been supplanted by new methods, for example, Andrews curves (Fig. 17.3, right) that display one-dimensional projections as a series of traces. One-line trace is generated by taking a row of the data matrix $x = (x_1\ x_2 \ldots x_p)$ and calculating (Eq. 17.2):

$$\frac{x_1}{2} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + \ldots, -\pi < t < \pi. \quad (17.2)$$

Andrews curves have been supplanted by parallel coordinate plots and tours, which are discussed in Sect. 17.2.

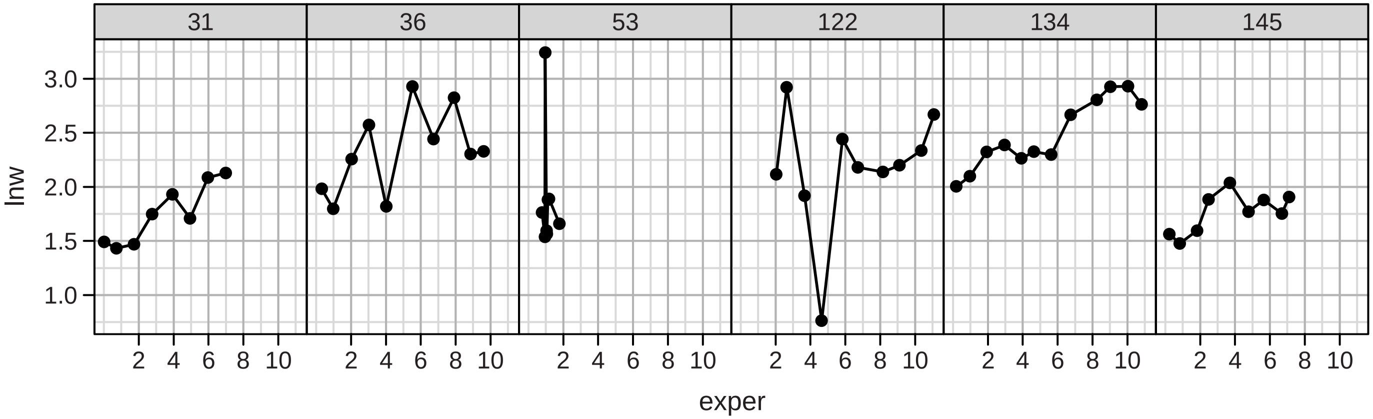


Fig. 17.2 Profile plots for a sample of subjects showing their log (wages) against workforce experience

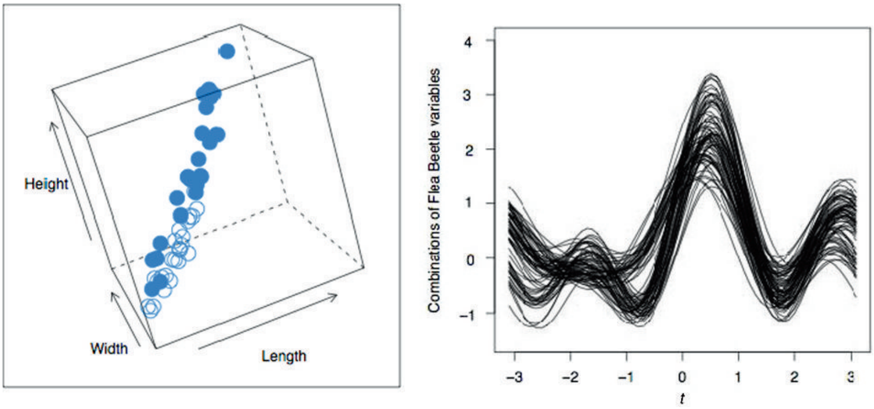


Fig. 17.3 Methods that should be retired: (*left*) 3D plots and (*right*) Andrews curves

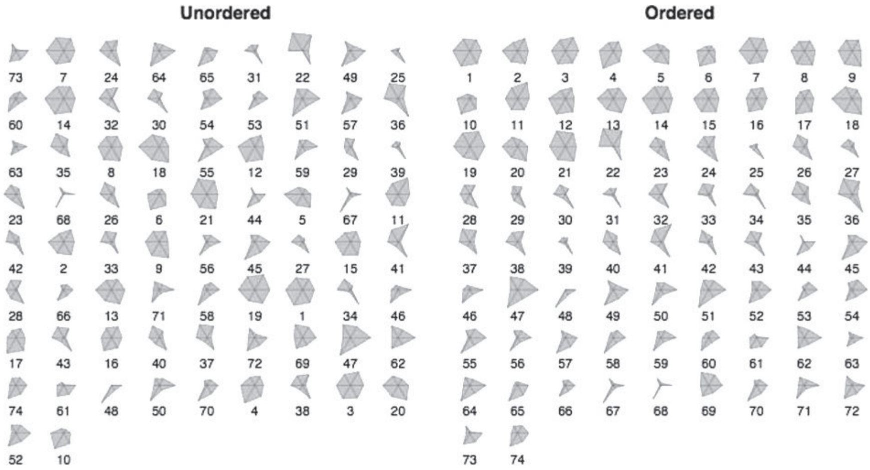


Fig. 17.4 Star plots of a dataset having 74 cases and 6 variables. The icons on the right have been reordered according to similarity of shape

17.1.3 Drop or Keep?

Icon plots code a vector of values into a single glyph. The most common examples are star plots, attributed by Friendly and Denis (2008) to date back to von Mayr (1877), and Chernoff faces (1973). All values for a case are represented by components of the icon. Figure 17.4 shows star plots of a dataset having 74 cases and 6 variables. The unordered icons in the left-side plot are simply displayed in random order. Some of the icons are roundish, and some have sharp points in one direction or another. Because there is no ordering of the icons, it is difficult to digest the patterns of similarity and variability in this field of icons. With a small amount of data and some organization of the icons, multivariate patterns can be seen. The ordered icons on the right have been organized roughly into three groups (1–21, 22–43, 44–74) by using another variable. Now it is possible to see these three groups as three different shapes. Icons 1–21 have large values of all six variables, resulting in a more rounded shape. Icons 22–43 have sharp points in the southeast direction, meaning that they all have large values for that one variable. The last group of icons looks more mixed. This dataset is the same as that displayed in the scatter plot matrix in Fig. 17.1, where it is also possible to see three fairly distinct clusters. Arguably, icon plots are not sufficiently valuable for multivariate analysis to warrant space in the curriculum. For most real datasets, because one icon is used for each case, the sheer number of icons to be viewed is impractical. However, icon plots are still very popular—as a Google™ search of *Chernoff faces* will attest—but unfortunately more for novelty purposes than for serious data analysis.

17.2 New Material

17.2.1 Parallel Coordinate Plots

Parallel coordinate plots are drawn to examine overall trends in data, such as clustering and outliers. They are constructed by laying axes in parallel instead of the more common orthogonal axes. Values corresponding to a row in the data matrix are connected by line segments. Here is an example of how they are constructed. Take a row of the data matrix (e.g., 191 131 53 150 15 104 from Table 17.2) and, using the minimum/maximum data values of all the data (Table 17.3), convert the numbers in the row to a value between 0 and 1 (e.g., (191 – 122)/(242 – 121) = 0.57, (131 – 107)/(146 – 107) = 0.62, (53 – 43)/(58 – 43) = 0.67, (150 – 116)/(157 – 116) = 0.83, (15 – 8)/(16 – 8) = 0.88, (104 – 55)/(123 – 55) = 0.72)). Plot these values on the corresponding axis, and connect the values from one axis to the next using a line segment. The result is a line trace for each row of the data matrix. Figure 17.5 displays a parallel coordinate plot of six variables measured on flea beetles. We can roughly see two different patterns in the traces that would correspond to clusters, and we can see one outlier that has low values on variables tars2, head, and aede1.

Table 17.2 Sample of five rows of a data matrix (not all entries for datasets shown)

	Variables					
Samples	tars1	tars2	head	aede1	aede2	aede3
1	191	131	53	150	15	104
2	185	134	50	147	13	105
3	200	137	52	144	14	102
4	173	127	50	144	16	97
5	171	118	49	153	13	106

Table 17.3 Minimum and maximum for each column of the full dataset

	Variables					
Ranges	tars1	tars2	head	aede1	aede2	aede3
Minimum	122	107	43	116	8	55
Maximum	242	146	58	157	16	123

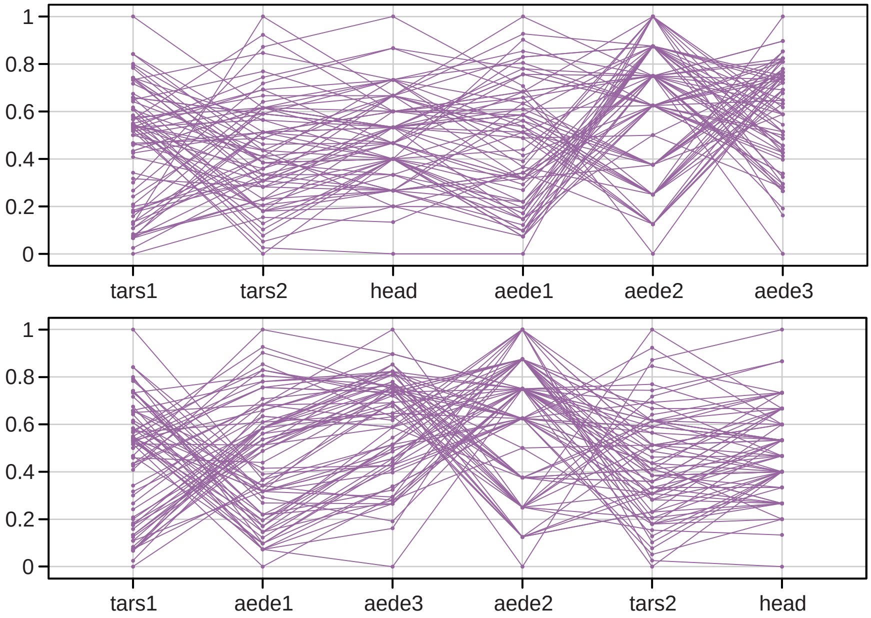


Fig. 17.5 Parallel coordinate plots of the same data, ordered two different ways

Slightly different structure can be read from the different ordering of the axes. The order of the axes is changed to produce the bottom plot. In the new ordering, the two different patterns in the traces corresponding to clusters are a little easier to see; a keen-eyed reader might even notice that there is a third pattern in the traces. The three clusters are characterized by the patterns (1) high, low, low/medium, high, …; (2) low, high, high, low, …; and (3) high, high, mixed, mixed, ….

Parallel coordinate plots are useful to see multivariate relationships. They can display a lot of variables; but too many data points produce too many lines, resulting in overplotting that makes the plots difficult to read. Interaction on the plot, discussed in Sect. 17.2.3, can help when there are a lot of data. It should be noted that sometimes it is a good idea to keep the original scale on each variable or use a global minimum and maximum instead of converting each variable separately to 0–1.

17.2.2 Trellis, Lattice, and Facetted Plots

Trellis plots (Becker, Cleveland, & Shyu, 1996), also known as lattice and facetted plots, are convenient ways to lay out *subsets* of data. These displays allow the exploration of conditional distributions of multivariate data. The plots in Fig. 17.6 are examples. On the left, a simple example is shown: box plots of a single variable for three groups in a dataset are plotted. The species Heptapot has low values on the variable, and the other two groups have relatively high values.

In the plot on the right, scatter plots of log (wages) versus experience for three race groups are drawn, with all the data shown on the far right facet. A smoother is overlaid on each plot, and the three smoothers are drawn on the overall plot. [*Smoothers* refers to regression methods that allow the data to influence the shape of a prediction line instead of forcing the decision to use a specific model in advance. These methods are usually called nonparametric regression or smoothing.] The grey regions around the smoothers indicate the confidence in the line estimate. We can see that the black subjects appear to have a different wage pattern than the other two races: it flattens out at the middle-experience levels. Also note that there are fewer data points for black subjects at high levels of experience, resulting in a broader confidence band.

17.2.3 Multiple Linked Plots

For multivariate data, not everything can be plotted in one graph; therefore, several plots of different aspects of the data are typically made. We can learn a lot about the joint distribution if these multiple plots are linked; highlighting a set of points in one plot should identify the corresponding points in the other plots. This is called simple linking, where points in plots are linked by their row identity.

Figure 17.7 illustrates simple linking between multiple plots. The row of the data matrix is linked between plots of different columns. A subgroup of species = 2, sex = 2 is brushed. [*Brushing* refers to the ability to change the color/size/glyph of points in the graphics window. It often refers to bolding or coloring points in a plot to draw attention to specific properties.] This corresponds to a group of points that has relatively small RW to FL values, as seen in the scatter plot of two other variables in the data. The parallel coordinates plot for these data says that the variables are all strongly linearly associated (profiles are very flat) and that the brushed group has relatively low RW relative to FL values (dip at RW).

Linking between plots can be more complex. Figure 17.8 illustrates an example where points and lines are linked by a categorical variable, identifying all values corresponding to a subject. An individual with a decline in wages after 8 years of experience is identified. This person has a steady increase in wages up to about 6 years, then suffers a decline, has another increase, followed by a decline. This plot

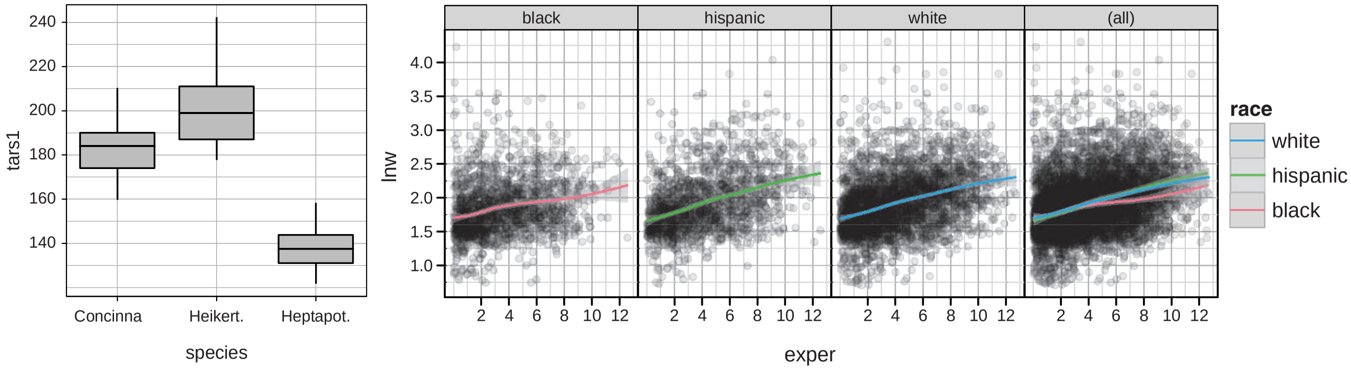


Fig. 17.6 Trellis, lattice, and facetted plots: (*left*) box plots for examining distributions for subsets of the data, and (*right*) scatter plots, with overlaid smooth curve, examining the relationship between log (wages) and experience for three subgroups, in comparison with the full data

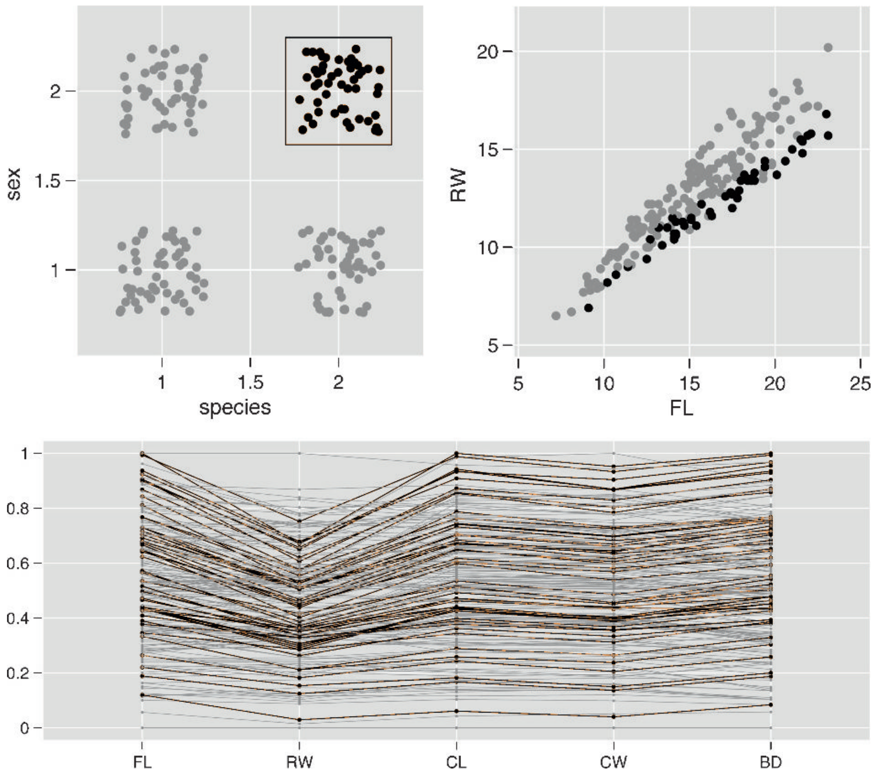


Fig. 17.7 Simple linking between plots can reveal a lot of multivariate structure. (*top left*) A subgroup of species = 2, sex = 2 is brushed. (*top right*) This corresponds to a group of points that has relatively small RW to FL values, as seen in the scatter plot of two other variables in the data. (*bottom*) The parallel coordinates plot for these data says that the variables are all strongly linearly associated (profiles are very flat) and that the brushed group has relatively low RW relative to FL values (dip at RW)

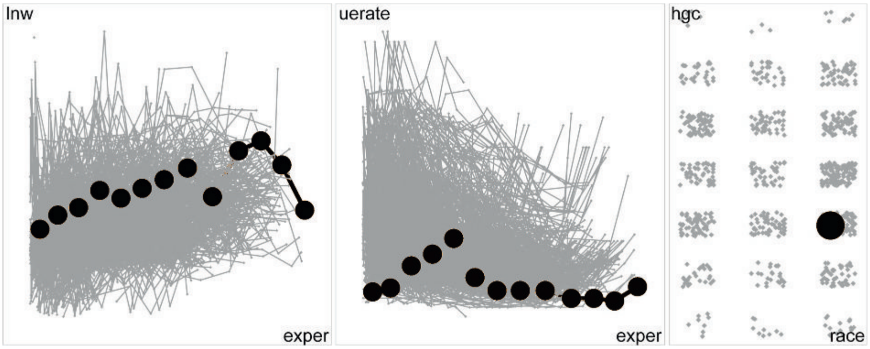


Fig. 17.8 Linking by a categorical variable enables us to explore longitudinal data. Here the profile for one subject's wage experience is linked to the unemployment in the local region and the subject's demographic information

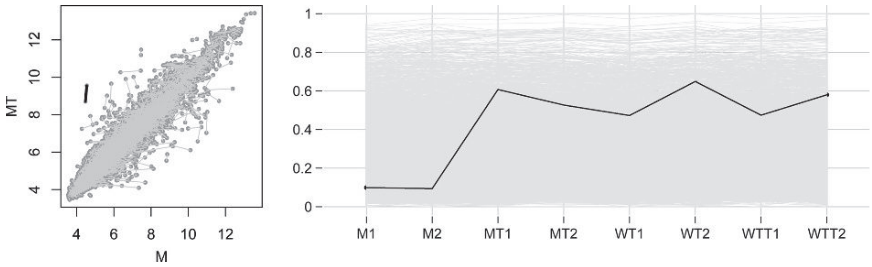


Fig. 17.9 Linking between different graphical elements. The line segment on the left corresponds to the profile on the right

is linked to another showing unemployment over the same period in this person's neighborhood (middle plot). The unemployment rate in that area varies differently from the person's wage pattern; it peaks around 5 years and then drops. We can also link to information about the person's demographic characteristics. This particular individual is in the white racial group (3) and attained a Grade 8 education.

The left and right plots in Fig. 17.9 link different tables of data. These data are from a designed experiment with two replicates for each treatment and multiple items. We are interested in the difference in treatment relative to difference in replicate. The most interesting samples have a big treatment difference relative to replicate difference, which in this plot corresponds to line segments that are short and far from the $x = y$ diagonal. A full profile of all the measurements for this case is shown in the parallel coordinate plot. The pattern for the highlighted sample is consistently low on first treatment (M) and high on remaining treatments. It is interesting because it has a big difference between treatment M and all others and relatively small variability in the replicates.

Linked brushing, as in the above examples, examines conditional relationships: If $10 < X_1 < 20$, what is X_2? [*Linked brushing* means that brushing conducted in any of the linked applications is immediately displayed in all others.] Learning about conditional distributions is a step toward understanding the joint distribution.

17.2.4 Tours

A tour is a continuous sequence of low-dimensional projections of high-dimensional data, shown as a movie, with several interaction modes. A projection is like the shadow of an object; if you see enough shadows, you might infer or recognize the shape of the object. Tours can be used for real-valued variables. They can help answer questions about the data related to the overall shape, or joint distribution, which includes clustering, linear or nonlinear association, or outliers. Generally, more structure is visible from a tour, showing combinations of the variables than in pairwise plots.

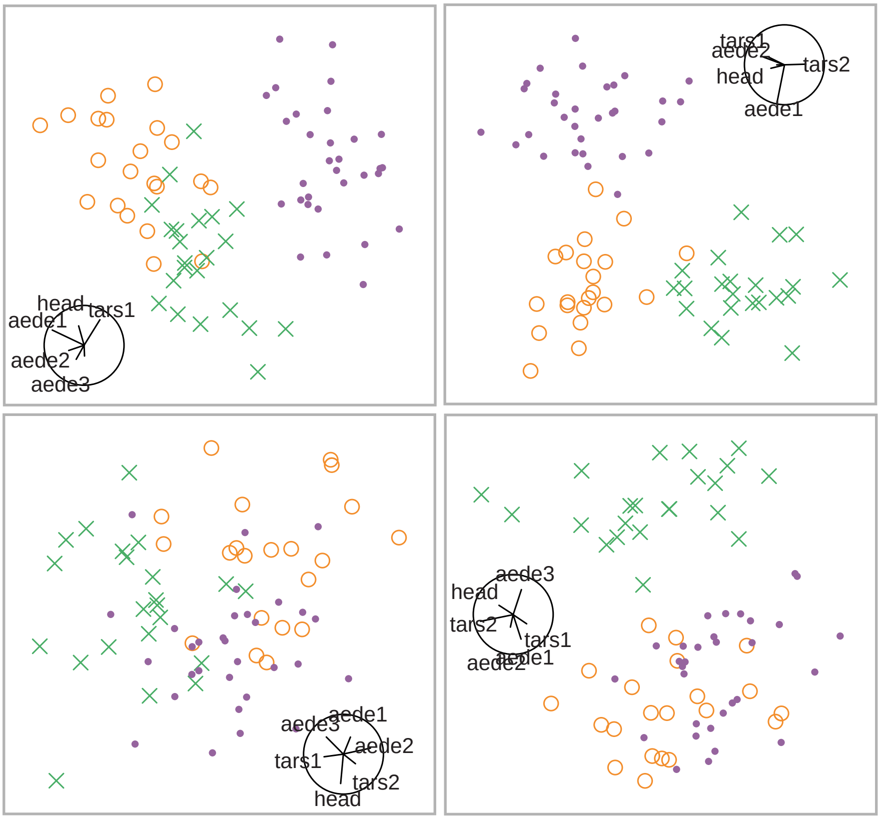


Fig. 17.10 Sample of projections seen in a tour of a dataset with six variables

Figure 17.10 shows four random, two-dimensional (2D) projections taken from a tour of the flea beetles data used in previous examples. The different symbols represent species. In some views, the three groups are more distinct and in others less so. The circle with lines is a representation of the projection of the orthogonal axes, which can help when interpreting structure seen in a single projection. Figure 17.11 compares the best projection seen in the tour (left) with the best of the pairwise plots (right). Although there is some distinction between the three groups in the pairwise plots, there is a bigger separation between the three when a combination of more than two variables is used, as was done in the tour projection.

There are several types of tours; the ones that we use are described in Buja, Cook, Asimov, and Hurley (2005) and Cook, Buja, Lee and Wickham (2008):

- **Random:** Choose new projections randomly; eventually see almost all possible projections.

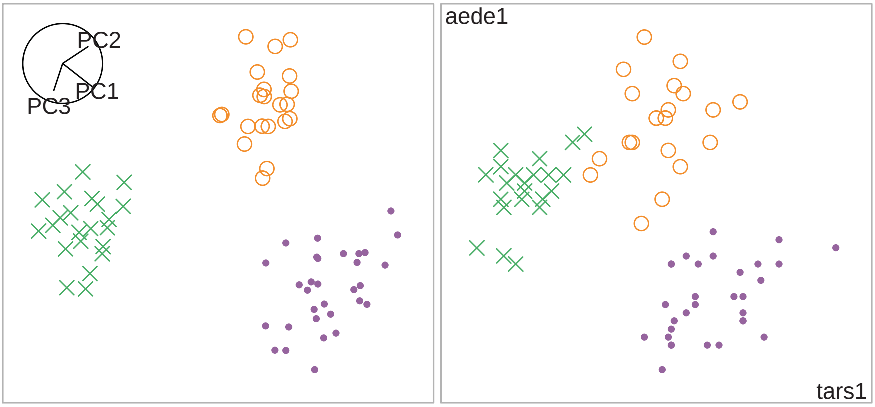


Fig. 17.11 Projection from a tour (*left*) shows a much bigger separation between the three groups than the best pairwise scatter plot (*right*)

- **Projection pursuit-guided:** Choose new projections according to a particular type of structure by optimizing a function describing the interesting structure.
- **Manual:** Pick one variable and change its projection coefficient; projection coefficients of all other variables are constrained on the values of the manipulated variable.

17.3 Applications

17.3.1 Data Analysis

For supervised classification, these new methods help to reveal how different methods operate and provide insights into the class structure in the data. An example of an interesting problem is to classify olive oil samples to geographic regions in Italy. This dataset has eight variables. Numerical methods give varying results. For one classification task, classification trees return a rule using just two variables: linoleic and eicosenoic (the first two variables in the scatter plot matrix in Fig. 17.12). The rule from linear discriminant analysis is similar, although a tad more complicated in that it has a combination of many variables, but the separation between groups is almost identical to the trees result. One group (1—open circle) is well separated from the other two groups (2—cross, 3—solid circle), which differ from each other but not by much. The tree classifier is a lovely, simple (parsimonious) solution; the separation is perfect. Undetected by the classifier, a third variable, arachidic, makes an interesting contribution to separating groups 2 and 3. On its own, it is not at all useful; however, in combination with linoleic, it produces a big gap. This is shown using a manual tour where arachidic acid is rotated into the horizontal axis containing linoleic acid (right plot), revealing the large divide between groups 2 and 3.

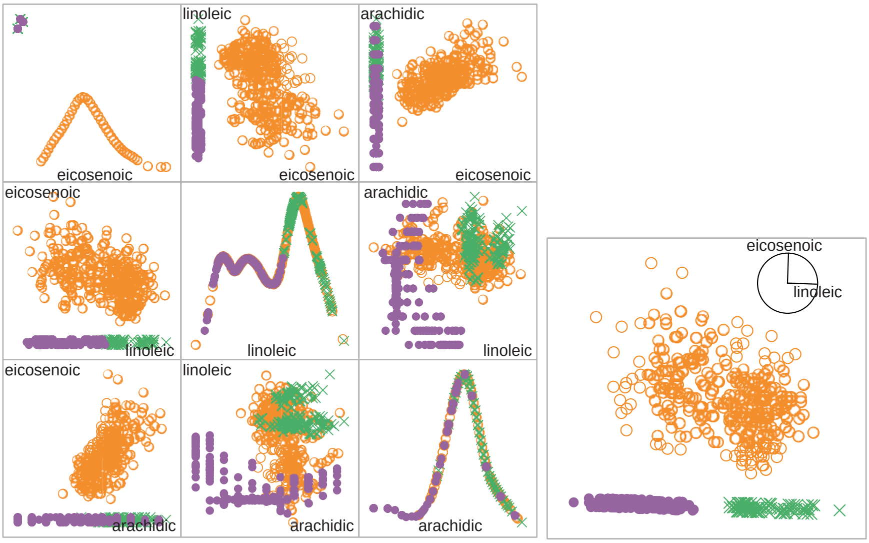


Fig. 17.12 Three variables important for separating the three groups in these data. Scatter plot matrix (*left*) shows that eicosenoic acid separates one group, but no single variable separates the other two groups. Using a manual tour (*right*) reveals a linear separation between all three groups, when linoleic and arachidic acids are combined

17.3.2 Multivariate Distributions

Density functions can be explored for more than 2D; two examples are provided in Fig. 17.13. The top one contains two linked windows showing a bivariate normal density and corresponding sample. A slice of density values is highlighted in one plot, revealing the elliptical contours in the other plot. The two bottom plots show the same process with a trivariate normal density and sample. The contours here are 3D ellipses. This same process can be used to look at density functions for any dimension distribution.

In a similar way, we can examine confidence regions. Figure 17.14 shows a 90% confidence region (ellipse) on three variables displayed along with the data (crosses), a hypothesized mean value (large solid circle), and the bounds of simultaneous confidence intervals for each variable generated by Bonferroni's method (cube). Several projections from a tour are shown. The hypothesized mean is just outside the confidence region, mostly due to the variables SweatRate and Na. A corresponding hypothesis test on whether this is a plausible value for the true mean will result in rejecting the null hypothesis at the 10% level, using Hotellings T^2 test statistic. If, instead of doing a formal hypothesis test, we simply computed Bonferroni simultaneous confidence intervals for each variable, we would find that the values of the hypothesized mean on each of these variables is inside the interval.

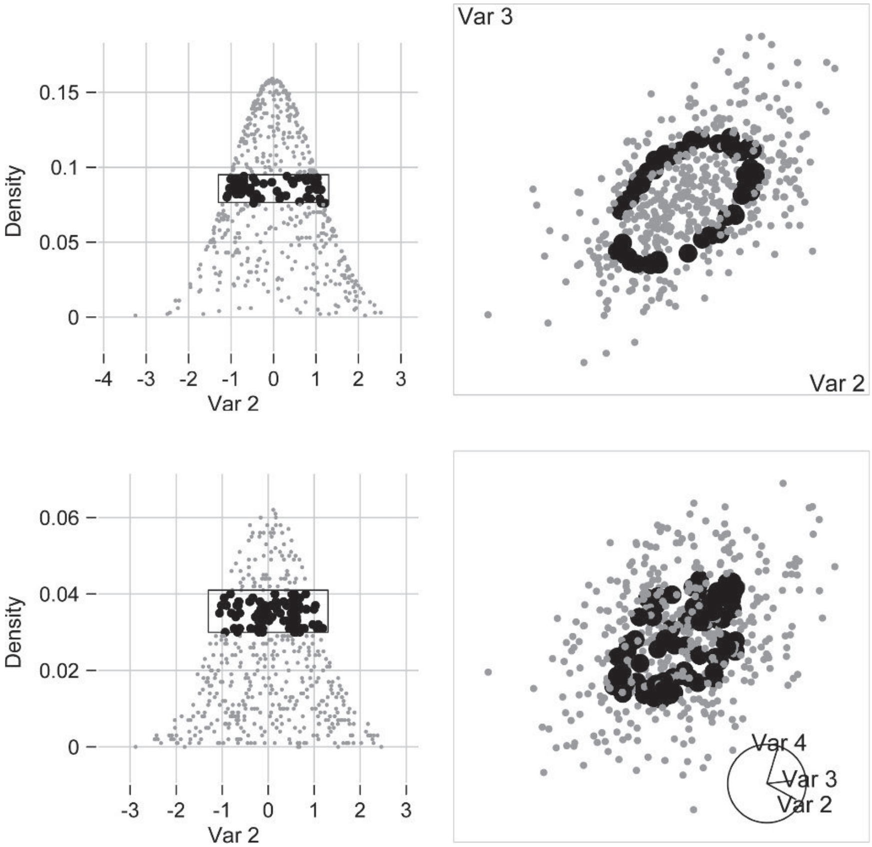


Fig. 17.13 Examining the multivariate normal distribution: (*top*) bivariate, (*bottom*) trivariate. Linked brushing between the density function and a sample from the multivariate normal reveals the elliptical variance–covariance structure

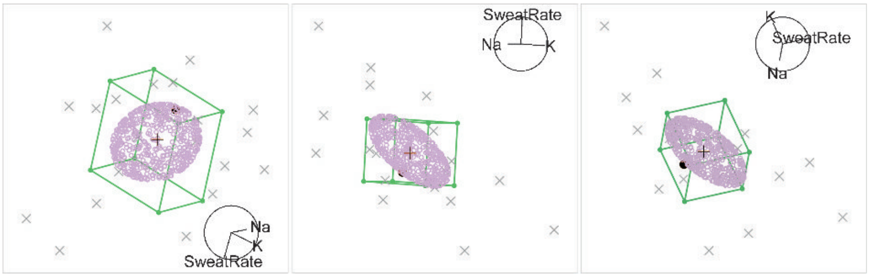


Fig. 17.14 Comparison of confidence regions, one being the Hotellings T^2 90% region (ellipse), and the other generated using Bonferroni's correction (cube). The hypothesized mean falls just outside the ellipse but inside the cube

We would not suspect anything amiss about the proposed values. Bonferroni's correction does not consider the variance–covariance of the data, basically requiring independence between the construction of the multiple intervals. This approach can be used to examine confidence regions for any number of variables.

17.3.3 Notes

Here are several comments about multivariate data visualization in general, some of which apply more broadly to plots of data.

- **Aspect ratio:** For multivariate data, each variable should be treated the same, which means the plot aspect ratio should be 1:1 (square) instead of rectangular, as shown in Fig. 17.15.
- **Overlaying additional information:** Information—such as mean, median, variance, smoothed curve, and regression curve—is good to overlay on plots of data.
- **Color and symbol:** High-contrast symbols (e.g., a closed circle, open rectangle, cross) are better to discern differences between categories. Contrasting colors, such as red/blue and light/dark shades, work well to discern different categories. It is also possible to check plots for color-blind robustness using tools available on the Web. As tempting as it might be to load further dimensions into color and symbol on a plot, it usually leads to confusion rather than clarity. It is really not helpful to have many different colors or symbols on each plot, as pretty as it might seem.

- **Working up into multivariate space:**

  1. Examine univariate marginal distributions (box plots, histograms, density curves) and lay out multiple histograms or box plots for each variable.
  2. Examine bivariate marginals: scatter plots, scatter plot matrix.

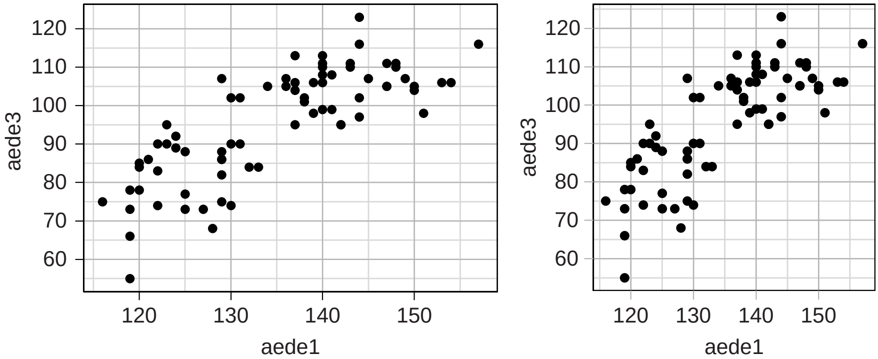


Fig. 17.15 A square aspect (*right*) ratio is correct for multivariate data; however, it is not the default for most plotting software

3. Examine conditional distributions: trellis, lattice, and facetted plots.
4. Examine the joint distribution: parallel coordinate plot, tours.

- **Resources:** The author employs the software GGobi (GGobi Data Visualization System, n.d.) to explore multivariate data using interactive and dynamic graphics. It has an associated R (R Project for Statistical Computing, n.d.) package, rggobi, which allows the graphics to be scripted somewhat from R. In the past year, several R application packages—classifly, clusterfly, meifly—have been developed for coordinating multivariate analysis procedures with visual representations using GGobi.

17.4 Summary

In summary, graphics treatment can be greatly enhanced in multivariate courses and should be improved in the new editions of the popular textbooks. A lot can be learned about multivariate structures in data and about theoretical quantities of multivariate distributions by using more graphics. Researchers in education generally—and perhaps especially in science and literacy education—can benefit greatly from applying lessons learned from statistical graphics. The impact suggested by the old adage *a picture is worth a thousand words* can be multiplied several fold by proper utilization of modern graphical methods. Therefore, for experts in education research, it is all the more important to achieve fluency in the modalities of visual communication and representation. Resources to help instructors incorporate material on interactive and dynamic graphics methods into their multivariate analysis classes are available in Cook and Swayne (2007) and the GGobi Web site (see http://www.ggobi.org/book/index.html). Another good source for current information on graphics for multivariate data is Wickham (2008).

Acknowledgments The resources for some material in this chapter were supported by NSF grant DMS0706949.

References

Becker, R. A., Cleveland, W. S., & Shyu, M.-J. (1996). The visual design and control of trellis fisplay. *Journal of Computational and Graphical Statistics, 5*(2), 123–155.

Buja, A., Cook, D., Asimov, D., & Hurley, C. (2005). Computational methods for high-dimensional rotations in data visualization. In C. R. Rao, E. J. Wegman, & J. L. Solka (Eds.), *Handbook of statistics: Data mining and data visualization* (Vol. 24, pp. 391–413). Amsterdam: North-Holland.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association, 68*(342), 361–368.

Cook, D., Buja, A., Lee, E. -K., & Wickham, H. (2008). Grand tours, projection pursuit guided tours, and manual controls. In C.-H. Chen, W. Härdle, & A. Unwin (Eds.), *Handbook of data visualization* (pp. 295–314). Dordrecht, The Netherlands: Springer.

Cook, D., & Swayne, D. F. (2007). *Interactive and dynamic graphics for data analysis: With R and GGobi*. Dordrecht, The Netherlands: Springer.

Friendly, M., & Denis, D. J. (2008). *Milestones in the history of thematic cartography, statistical graphics, and data visualization*. Retrieved July 13, 2008, from http://www.math.yorku.ca/SCS/Gallery/milestone/

GGobi Data Visualization System. (n.d.). *Homepage*. Retrieved July 13, 2008, from http://www.ggobi.org/

Johnson, D. E. (1998). *Applied multivariate methods for data analysis*. Pacific Grove, CA: Duxbury.

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th edn.). Upper Saddle River, NJ: Prentice-Hall.

Mayr, G., von. (1877). *Die Gesetzmäßigkeit im Gesellschaftsleben* [The regularity of social life]. Munich: Oldenbourg.

R Project for Statistical Computing. (n.d.). *Homepage*. Retrieved July 11, 2008, from http://www.r-project.org/

Tukey, J. W., & Tukey, P. (1981). Summarization, smoothing, supplemented views. In V. D. Barnett (Ed.), *Interpreting multivariate data* (pp. 245–275). New York: Wiley.

Wickham, H. (2008). *Practical tools for exploring data and models*. Unpublished doctoral dissertation, Iowa State University, Ames. See also http://had.co.nz/thesis.

Chapter 18
Approaches to Broadening the Statistics Curricula

Deborah Nolan and Duncan Temple Lang

Recently, there has been a lot of discussion about what a statistics curriculum should contain, and which elements are important for different types of students. For the most part, attention has been understandably focused on the introductory statistics course. This course services thousands of students who take only one statistics course. In the United States, the course typically fulfills a general education requirement of the university or a degree program. There has also been considerable activity regarding the use of computers to present statistical concepts and to leverage the Web and course management software to interact with students. Recently, there has been debate as to whether statisticians should make ambitious changes using resampling, the bootstrap, and simulation in place of the more traditional mathematical topics that are seen as the fundamentals or origins of the field (Cobb, 2007). It is unclear that we are achieving the goals of basic statistical literacy by focusing on formulae or even by concentrating almost exclusively on methodology. Instead, we believe the field and students would be significantly better served by showing the challenges and applicability of statistics to everyday life, policy, and scientific decision making in many contexts, and by teaching students how to think statistically and creatively.

In contrast to the activity at the introductory level, there has been much less attention paid to updating the statistics curricula for other categories of students. While smaller in number, these students—undergraduate majors and minors, masters, and doctoral students—are very important, as they are the ones who will use statistics to further the field and improve the quality of research. Other disciplines (e.g., biology, geography, and political and social sciences) are increasingly appreciating the importance of statistics and including statistical material in their curricula. Further, statistics has become a broader subject and field. However, the statistics curricula at these levels have not changed much past the introductory courses. Students taking courses for just

D. Nolan
University of California, Berkeley

D. Temple Lang
University of California, Davis

M.C. Shelley II et al. (eds.) *Quality Research in Literacy and Science Education,*

2 years may not see any modern statistical methods, leading them to a view that the important statistical ideas have all been developed. More importantly, few students will see how these methods are really used, and even fewer will know at the end of their studies what a statistician actually does. This is because statisticians very rarely attempt to teach this; instead, they labor over the details of various methodologies. The statistics curricula are based on presenting an intellectual infrastructure in order to understand the statistical method. This has significant consequences for improved quantitative literacy. As the practice of science and statistics research continues to change, its perspective and attitudes must also change so as to realize the field's potential and maximize the important influence that statistical thinking has on scientific endeavors. To a large extent, this means learning from the past and challenging the status quo. Instead of teaching the same concepts with varying degrees of mathematical rigor, statisticians need to address what is missing from the curricula. In our work, we look at what statistics students might do and how statistics programs could change to allow graduates to attain their potential.

18.1 What is Missing from the Statistics Curriculum?

Efron (as cited in Rossman & Chance, 2003) noted that theoretical statistics courses are caught in a time warp that bores course instructors and subsequently their students. Cobb (2007) supported this position: "We may be living in the early twenty-first century, but our curriculum is still preparing students for applied work typical of the first half of the twentieth century" (p. 7). For example, hypothesis testing takes up a lot of space in the current undergraduate curricula because statistics textbooks place so much emphasis on sets of rules developed in the 1950s for various test statistics (e.g., *z*-test, one-sample *t*-test, two-sample *t*-test, paired *t*-test, *t*-test with nonhomogeneous variances). As a result of the large number of formulae, too little attention goes to the main notions behind testing. Even worse, these same sets of rules for testing are taught over and over again in introductory, advanced undergraduate, and graduate courses. This approach fails to teach modern developments in statistics and fails to convey the excitement of statistical practice.

Cobb (2007) posited the reason for this focus on particular tests stems from the way the curricula have developed: "What we teach was developed a little at a time, for reasons that had a lot to do with the need to use available theory to handle problems that were essentially computational" (p. 6). However, modern computing offers alternatives to these approximations; and it offers the opportunity to break from the constraints of current curricula and to design syllabi from scratch, using the large collection of computational tools and statistical experiences currently available. If students are facile with computing, they will be able to actively explore methods and their characteristics and limitations in contrast to merely accepting mathematical statements about them. Also, computationally capable students will be able to work on interesting, topical, and scientific problems and to apply statistical ideas.

Computing is one dimension of the statistics curricula that can attract bright, talented students and educate them in a way that will broaden the focus and impact of statistics. We want graduates who not only *know* statistics but can *do* statistics; increasingly, that means nontrivial computation on rich, varied, and large datasets from many sources in collaboration with scientists from other fields. We must modernize the curricula to include computation, like mathematics, as an important medium for expressing statistical ideas—not only for being able to apply statistical concepts but also to be able to develop the computational tools needed for research.

In addition to addressing its computational inadequacies, we advocate teaching from the vantage point of statistical concepts flowing from contextual problem-solving with data. Traditional courses do “not attempt to teach what we do, and certainly not why we do it” (Efron, as cited by Rossman & Chance, 2003, p. 3), yet intuition and experience of methodology in the scientific context are essential to learning how to think statistically (Wild & Pfannkuch, 1999). Statistical thinking and practice involves so many more aspects than selecting and fitting statistical methods to data. Yet most courses focus on statistical methodology—either the theory or the application—and very few discuss in any detail the skills needed to approach a scientific problem from a statistical perspective. What is missing is the experience of connecting these methods to actual applications and rounding the student into a scientific collaborator. An application is too often just an example of how to apply a particular statistical method to some manageable data—preselected by the instructor to illustrate the strength of the method—rather than a scientific application that students identify and evaluate based on relevant statistical methods. For those learning statistics, the intuition and experience that are necessary for good data analysis are the hardest things to learn; they involve a very different dimension of both learning and thinking than are used in mathematical thinking developed when teaching statistical methodology. Undoubtedly, students need to understand both methodology and statistical experience. At present, the focus is primarily, if not exclusively, on the former (Bryce, Gould, Notz, & Peck, 2001).

In this chapter, we describe activities that we believe will reduce some of the hurdles in achieving these changes and promote quantitative literacy. The curricula changes we believe are needed come from two aspects: (a) embracing computing as an essential building block of statistical creativity and practice; and (b) focusing on statistical experience, reasoning, and applications. We are not suggesting that the mathematical approaches be discarded. Instead, we propose creating more balanced, relevant, and modern curricula for various levels of students that are determined by what they need for the future, rather than what is known from the past.

One project that we have embarked on seeks to first collect and then disseminate materials to help faculty members argue for, introduce, and teach computing within the statistics curricula. These materials include model or template syllabi—which are intended as discussion documents describing the different elements of statistical computing and how each is important for different types of students and to get courses adopted by departments—and lecture notes, exercises, projects, tutorials, textbook chapters, a textbook in data technologies, and workshops to assist faculty members teach statistical computing. By leveraging the existing, small community of those involved in statistical computing research, we are essentially trying to seed the statistical community with resources for teaching computing as a part of the undergraduate and graduate curricula.

A second project addresses the issue of statistical practice. The authors and a colleague (Hansen, Nolan, & Temple Lang, 2006) developed a model for a summer program in statistics where undergraduates were exposed to important, topical, and scientific research problems presented by statisticians working on a team with scientists who were addressing a problem. The statisticians brought data for the students to creatively explore. The students gained a sense of the data and how the data might be used to address the problem. Based on evaluations from students and faculty participants, the approach developed in the program was successful in exciting the students about the possibilities that statistics holds.

Finally, relating to statistical thinking in the curriculum, our third project (Nolan & Temple Lang, 2007) offers another approach for providing students with statistical experience. This project aims to provide a flow of materials from statistics researchers involved in scientific research to the pedagogical community using reproducible, dynamic, and interactive documents. The premise is to enable researchers to document their computations and analyses so that they can be reproduced for both themselves and others (e.g., peers, reviewers, editors, managers). Researchers would work in an environment that captures their writings, computations, and thought process in an electronic notebook. In essence, the notebook is a database of all the activities within the data analysis or simulation study; and it can be projected into different views to make the information it contains available for different audiences. These documents would provide resources to instructors to assist them in teaching in new ways because they would open up the thought process and experience behind a data analysis both to the instructor and the students. This technological approach would support a model for cooperation between statisticians active in research and consulting and the community of statistics educators. Instructors would then have libraries of real case studies that include data analysis projects and current research methodologies that show how statisticians think and work.

18.2 Computing in the Curricula

The current approach to teaching statistics focuses almost exclusively on mathematics. Mathematics is not always the best medium in which to teach statistical concepts; unfortunately, a heavy reliance on mathematics restricts an instructor’s ability to convey statistical concepts in a fuller light. Although many tasks are easier to convey through mathematics, others are more appropriately conveyed through a plot, a simulation study, or experience with data. That is, these computational approaches offer complementary means for presenting and understanding statistical concepts (Moore, 1997). In addition, not all students appreciate the insight mathematics offers, and not all uses of mathematics offer the best insight. With the computer, students can explore data to formulate scientific questions. They can explore statistical models to understand assumptions, operating characteristics, etc., and how they behave—and they can explore both together to answer scientific questions. Moreover, computing has revolutionized statistics; many modern statistical methods are feasible only because of today's computational opportunities. It is unimaginable that statisticians today would not be facile with the computer, for they are expected to be able to access data from various sources, apply the latest statistical methodologies, and communicate their findings to others. They should be encouraged to (a) create interesting presentations of statistical findings with important consequences (e.g., as exemplified by Gapminder, n.d.), (b) influence developments in the digital world (e.g., the semantic Web), and (c) increase the impact of good decision making with statistics.

18.2.1 What do Statisticians Need to Know?

Clearly, computing is not a fad, but something vital to the field of statistics. "Computation is now regarded as an equal and indispensable partner, along with theory and experiment, in the advance of scientific knowledge" (Society for Industrial and Applied Mathematics [SIAM] Working Group on CSE Education, 2001, p. 163). Although many agree that there should be more computing in the statistics curriculum and that statistics students need to be more computationally capable and literate, it can be difficult to determine how it should change because computing has many components or dimensions. These components need to be carefully considered and prioritized in order to understand where they might fit and which groups of students would benefit from a particular emphasis.

While statistics students must learn practical details of computing (e.g., programming language syntax), we must strive to teach higher-level concepts including a vocabulary and computational thinking that will enable them to discuss computational problems precisely and clearly. Vocabulary is necessary to be able to communicate—understand, express, and reason—about computational issues. As computing and data technologies continue to evolve rapidly, especially as statistics enters the era of mainstream parallel and distributed computing for scientific computing, it is essential that students are provided a good foundation rather than a thin memorization of specifics so that they are able to continue to learn new aspects of computation. Statisticians must not mistakenly think all that is needed to introduce computing into the curriculum is to teach students a particular programming language. We must aim higher and more generally, just as statistics is not taught as a collection of formulae and ad hoc tricks. It is helpful to look at three high-level components of statistical computing: programming languages, environments, and paradigms; algorithms and data structures; and data technologies.

18.2.1.1 Programming Languages, Environments, and Paradigms

The vast majority of, if not all, statisticians would agree that students need to learn a programming language (American Statistical Association [ASA], n.d.; Bryce et al., 2001). There will be different opinions about what language(s) should be taught. Some will want to cover the practical aspects of using common types of statistical software or packages, such as SAS or SPSS. However, we believe that computing should be viewed as a supporting skill for statistical practice and research and that courses should cover the concepts of computing as well as the specifics; that is, the teaching of computing needs to be approached in the same way as the teaching of mathematics or statistics. Students need to be able to transfer the concepts to the specifics of other languages and environments as they change from, for example, R to MATLAB or from MATLAB to Perl. When taught in isolation, programming languages are idiosyncratic and arcane. When taught more generally, the commonality and patterns emerge and provide a significantly simpler viewpoint and much more useful, general skills that will serve them well in the future. Students should be made aware that (a) different languages serve different roles and (b) learning just one language is likely to be quite limiting in the future.

For these reasons, we advocate that students learn a general-purpose programming language with which they can create new algorithms and functionality and express statistical ideas and computations at a relatively high level. Some students will need to learn lower-level languages (e.g., C or FORTRAN), but most will be well equipped with languages such as MATLAB or R. Our experiences with code written for student research (graduate and undergraduate) have included both the need for more sophisticated algorithms and better understanding of fundamental programming concepts. Teaching algorithms that are subtly different or whose applicability is somewhat subtle, before improving the basic programming skills of students, would seem to be misplaced. A course in computational statistics—essentially how to do statistical computations properly—is more appropriate as a follow-up to an introductory course in programming languages and environments.

In addition to programming languages, graduate students will need to learn new paradigms, such as parallel and distributed computing. These are no longer exotic, specialized topics but commonly used techniques for implementing real scientific computations. Similarly, as statisticians increasingly publish software implementing their methodological research, graduate students need to understand some essential principles of software engineering. Issues of portability, memory management, object-oriented programming, compilation and efficiency, version control, unit testing, and so on are very important in developing software for others to use. What might have been considered advanced computing a decade ago is becoming more important for doctoral students so that they can successfully function in the scientific community.

18.2.1.2 Algorithms and Data Structures

If one were to ask academic statisticians what computing should be taught to statistics graduate students, many would list linear algebra decompositions, numerical optimization, and aspects of approximation and numerical analysis (Gentle, 2004; Lange, 2004; Monahan, 2004). Interestingly, none of those topics requires a computer; however, they are methods for obtaining solutions efficiently or approximately, or both, and often become necessary in advanced research. These are undoubtedly good things for students to know. However, when prioritizing the importance of this material relative to other statistical and computational topics, their importance is less clear. For the most part, students will not implement the general algorithms as the implementations available in well-tested, widely available software are efficient and robust. Teaching the circumstances under which each algorithm might be best utilized is undoubtedly useful, but it is not necessarily limited to a computational course. Rather, it should be part of a theory or methodology course in which the need for optimization is raised (e.g., maximum likelihood estimation, robust regression).

If such topics are to be taught in a computational course, it is imperative that the students have the skills to be able to express computations so that they can quickly perform experiments to explore and understand the characteristics of these algorithms. For the most part, these are topics more appropriate for graduate students than undergraduates—and not all graduate students will need such skills early in their research—as these topics only make sense when the student has studied statistical methods that require such computational approaches. On the other hand, all students will need to be able to program, perform simulations, and access and manipulate data. While the choice of algorithm is often critical for developing efficient code, poor understanding of programming concepts is often the primary cause of inefficiency. Furthermore, human time is expensive relative to computer cycles; therefore, optimizing performance may be a waste of precious resources. These considerations imply an order and a priority for the different computational topics.

Classical computational statistics topics are undoubtedly of importance, and, all else being equal, students should master them. Statisticians need to question this legacy and consider new topics and their importance. We wish to provoke thought about their importance relative to other potential topics for a computational statistics course. Simulation, computer experiments, Markov chain Monte Carlo (MCMC), the Expectation-Maximization (EM) algorithm, and resampling methods (bootstrap, cross-validation) are of greater importance from a pedagogical perspective than matrix calculations and optimization algorithms—because they are less amenable to general-purpose implementations and so do not exist as well-tested implementations in common software environments. Furthermore, since matrix calculations and optimization algorithms are extremely well implemented in widely available libraries and environments, students should not write their own versions of these highly tuned implementations.

18.2.1.3 Data Technologies

It is much easier to teach more algorithmic, mathematical material, such as the topics found in many computational statistics courses, than it is to teach topics in data technologies. For many, merely defining data technologies may prove difficult. Instead, think of these topics as new computational tools, techniques, and standards that allow access to data and functionality from varied sources and the presentation of information, results, and conclusions in rich, dynamic ways. These technologies include regular expressions for text manipulation, relational database management systems for storing and accessing data, the eXtensible Markup Language (XML) used for many purposes, Web services for distributed functionality and methods, Web publication and asynchronous documents, interactive and dynamic graphics, etc. In fields such as bioinformatics, finance, and astronomy, these are essential tools by which researchers access and share data. They are becoming important for statisticians who work in these fields, and a handful of statistics departments around the world are beginning to teach these topics. However, they are much less amenable to the definition–theorem–corollary–application style of teaching. They require instructors to think about teaching in a different manner; thus, it is necessary to rebuild much of the usual infrastructure for teaching.

We argue that, while difficult to teach, the topic of data technologies is growing in importance in the field. As statisticians deal with larger amounts of data from many and varied sources, often the challenges to data analysis start well before the computational steps involved in model fitting. Rather, simply accessing the large volume of data and getting it into the programming environment in a manageable way (e.g., from a relational database) can pose a problem. The choice of data structure and understanding when and when not to copy data are examples of issues that may be far more important and immediate hurdles. Further, rather than being concerned with potential inefficiencies in an algorithm, it is often more productive to use profiling tools to determine where lie the bottlenecks in the code. These profiling tools, data structures, and management of large datasets may well be more important than learning about efficient algorithms that are needed primarily for worst-case situations.

In addition to these three core topics in the statistics curriculum, one other topic deserves mention and consideration: visualization. Visualization clearly offers an invaluable tool, and computing plays an important role in modern visualization techniques for data analysis. Skills in visualization may well be the most valuable of all computing skills when considering the ubiquity of visual presentations of data and the great potential for communicating complex data structures simply with appropriate images. However, when promoting one set of computational topics over another, we must be quite specific about the goals, the audience, and their interests.

18.2.2 What are the Challenges to Making This Change?

There is little doubt that statistical ideas and concepts are important topics for students to learn. Mathematics and computing are supporting tools that aid learning these concepts and provide complementary approaches to this end. Ideally, both approaches are mastered. However, unlike mathematics, if one does not have computational skills, one simply cannot engage in the application and practice of statistics regardless of one's knowledge of the concepts. Computation is the currency of statistical action while mathematics is typically the currency of statistical description. Since most statistics students go on to apply statistics rather than study it academically, computational skills are vital.

At the graduate level, most statistics students have studied mathematics for at least 6 years, have taken at most one course in computing, and have no experience in statistical or scientific computing. While their mathematical skills may not be as strong as instructors would like, most students do not have a vocabulary for computation and often arrive with bad habits from point-and-click applications. Statistics departments have historically admitted doctoral students solely on the basis of their mathematical background. More balance in graduate curricula is needed so that students can leverage both mathematics and computation to understand and practice statistics and play a more active role in current and future developments.

We have heard many reasons or explanations why computing is not a larger part of the curricula. In our opinion, the primary reasons computing is omitted are: (a) it is difficult and time consuming to teach or retool to teach, and (b) the discipline is conservative and clings to its mathematical past. It is useful to consider these explanations because some are legitimate points of view and obstacles to change.

18.2.2.1 "We don't know that material."

One explanation sometimes offered for not having computing in the statistics curricula is that statistics faculty were never taught computing, they have not had the opportunity to learn it, they cannot teach it effectively, and so do not. This is very unfortunate as it means that new students do not have the opportunity to learn it either. At some point, statisticians need to break the cycle and learn this material. The situation is improving as some students are learning this material on their own, albeit in an ad hoc fashion and often incorrectly. With a willingness to change, the cycle can be broken.

18.2.2.2 "We send our students to Computer Science to learn computing."

This approach seems reasonable until one tries to determine more precisely what statistics students should learn and then map these needs to the courses available in computer science (CS). For the most part, CS courses are justifiably concerned more with abstract, theoretical aspects of computer science and technology than statistics students need to learn. That is, computing is an important means to an end for statisticians, not a study in its own right. For example, statisticians generally need not worry about the optimizations performed in the execution of a relational database query; instead, they should understand the general principles of the relational database model and the common elements of the Structured Query Language (SQL) used to extract data from a database. This material is easily covered in several classes; statistics students do not need to spend an entire course on other, less important topics. As for programming concepts, we would argue that the emphasis and tone of general programming courses are not appropriate. Rather, high-level, interpreted languages that use vectorized operations and provide garbage collection (e.g., R, S-PLUS, MATLAB) are more appropriate. This is very different from the more traditional, object-oriented languages (e.g., C, C++, Java) used in introductory CS courses.

This is not to suggest that there are no CS courses that are relevant to statistics students. However, they must acquire fundamental scientific computing knowledge and skills that are the prerequisites to the more advanced topics. For database design, a CS course in more detailed database topics would be valuable. Understanding algorithms is a very important skill; therefore, a data structures and algorithms course would be useful. For disseminating statistical methods as software, a software engineering course is important. Scientific visualization is another course that is highly recommended.

18.2.2.3 "We let the students learn it on their own."

Many of our colleagues advocate—or at least practice—the approach in which students are told to learn about computing by themselves, from each other, or from their teaching assistant. This sends a strong signal that the material is not of intellectual importance relative to the material covered in lectures. With this approach, students pick up bad habits, misconceptions, and, more importantly, the wrong concepts. They learn just enough to get what they need done, but they do not learn the simple ways to do things or take the time to abstract what they have learned and assimilate these generalities. For many, they are unaware of the possibilities that surround them and so continue to do everything in the same, limited way. They cannot learn about new topics as they lack a basic vocabulary. Their initial knowledge shapes the way they think in the future and typically severely limits them, making some tasks impossible. They lack the ability to deal with new problems, and they typically lack the necessary confidence to approach new tasks. Their lack of computational skills makes it difficult for them to work in a team where others are computationally capable, independent, and autonomous. The curricula must provide computing fundamentals; we believe that adding a small amount of structure and guidance would yield large professional gains for students, research assistants, and professionals.

One of the rather ironic aspects of this approach is the relative paucity of material with which the students can learn. There are very few textbooks on general aspects of statistical computing or programming for statistics. In contrast, there are hundreds of textbooks on each statistical topic that instructors present in lectures. There seems to be an inversion in teaching, where students are left with few aids in computing while valuable contact time is spent repeating what they can read for themselves.

18.2.2.4 "Students only need to learn basic programming."

Some statisticians think that graduate students only need to learn MATLAB or R, others think just SAS is needed as it is widely used, while others think a language such C or Java is the right choice because it is the common language of scientific computing and is easily transferred. The fact that there is a difference of opinion and various options illustrates that there are differing goals and needs for statistics students. However, it is a big leap from a single course in basic computing to embracing problem-solving methodologies and general computing principles; the latter should be taught and then fostered by the culture of a modern, vibrant department that contributes to advances in statistical methodology and application. Additionally, to omit data technologies (e.g., relational databases, XML, Web services, distributed and parallel computing, etc.) is a disservice to those students having only basic computational literacy skills when they work with others from different fields who are vastly more skilled in the practicalities and advanced skills of working with data.

18.2.2.5 "Computing is not as important as our core statistical concepts."

While this may have once been true and may still be relevant to those with a very narrow view of statistics, the growth in data analysis in all sciences and the relative intractability of complex models and methods makes computational skills of immense importance in a modern view of statistics. The goal of statistics education is statistical concepts and thinking. Mathematics and computing are supporting tools that aid learning these statistical concepts; both must be mastered, not just the former. To function in the practice of statistics, one must be capable of increasingly complex computation. Since most students go on to practice statistics rather than study it academically, computational skills are vital.

18.2.3 How do we Make it Feasible to Teach Computing?

It is an immense amount of work for an individual instructor to integrate computing into an existing statistics curriculum. First, the instructor must socialize the idea with colleagues, foster their support to add or change a course, and argue about which course or topics can be discarded to make way for this new material. This change can be met with resistance or apathy, which can herald the end of the process for all but those who feel sufficiently strongly about the new direction to persevere. Then there is the need to create a syllabus, debate what should and should not be included, and outline the topics on a week-by-week or lecture-by-lecture basis for the course. The process will typically involve multiple iterations. With all this done, the course may be submitted to a campus-wide committee for approval. After this, one must still convince other relevant individuals that the course should be scheduled and taught and not simply listed in the course catalog.

Having cleared the typically lengthy administrative hurdles, the instructor now has to teach the material, which can involve the following steps:

- Decide what the basic programming language will be, for example, R, S-PLUS, MATLAB, SAS, Perl or Python, C/C++.
- Be familiar with the topics at a level that goes at least slightly beyond what is being taught.
- Prepare exercises and longer projects, which involves identifying and evaluating the main topics to be covered and deciding how problems can be combined to reach the overall goal.
- For projects, find interesting datasets with an associated scientific or social problem of interest and then create a sequence of doable tasks that lead to the pedagogical goal, which typically means trying three or four datasets to find one that fits all the necessary criteria.

This collection of hurdles makes it apparent why statistical computing is not taught more. However, computing is too important to merely accept the difficulties and continue along the current, traditional path. We advocate pooling resources so that the materials needed to clear the administrative hurdles and teach the topics are available as templates that can be quickly adapted and customized for different situations.

To this end, we are creating, gathering, and disseminating materials to help faculty members initiate new courses or modules on computing within the statistics curricula (Hansen, Nolan, & Temple Lang, n.d.). To promote discussions among faculty members and assist the decision-making process, these materials include model or template syllabi—discussion documents that describe the different elements of statistical computing, why each element is important for different programs of study, and why the topic was selected. To aid in teaching, the materials include lecture notes, exercises, case studies, projects, tutorials, textbook chapters, and a textbook on data technologies. In today's Web-based world of information exchange, we no longer need to think in units of textbooks but smaller units that can be combined creatively for different courses. In addition to these materials, we have organized 1-week workshops for faculty on how to teach statistical computing and established an electronic forum for discussing aspects of teaching computing in statistics. Essentially, we hope to seed the statistical community with resources for introducing and teaching computing by leveraging the existing, small community of those involved in statistical computing research.

This work is part of a 3-year grant funded by the US National Science Foundation from its Division of Undergraduate Education. We began with a workshop that brought together experts in statistical computing to discuss different topics and evaluate the areas to be taught. The wiki (Hansen et al., n.d.) includes findings from the workshop, syllabi from computing courses taught by workshop participants and others, and an annotated bibliography on statistics curriculum reform. In July 2008, we held a workshop for faculty members from around the USA who teach or plan to teach statistical computing. The participants are from departments with a commitment to introducing or continuing to develop their computing courses. The workshop covered the basic material and discussed different teaching approaches. Participants worked through case studies and projects, thinking about how to get students involved and be creative.

18.2.3.1 Building for the Future

More recently, the statistical community has seen the advent of several systems developed by and for the community. XLisp-Stat and Quail are systems that explore different paradigms for statistical computing and have significant results and merits. The S language and its two implementations—S-PLUS® (Insightful Corporation, n.d.) and R (R Development Core Team, 2006)—have been very important for the practice of statistics and also for the development of over a thousand add-on packages providing cutting-edge methodologies, often before they are published in journals. This new form of disseminating work is a terrific, modern change. As excellent as R and other systems are, they are aging, relative to the dramatic changes and innovations of both hardware and software from the engineering and information technology communities and relative to the ever-increasing size of datasets of interest. It should be clear that statistics as a field must continually innovate and build new systems and infrastructure to handle the challenges and new directions for statistics so as to remain relevant. The popularity and impact that R has had on the field should encourage us to put more resources into development, rather than hold the misguided belief that we have all we will need. The infrastructure must adapt to the changing needs of statisticians by importing innovations in general computing and technology and helping transform and shape statistics.

Statisticians cannot depend on commercial entities to develop the tools they need. Nor can they rely on research laboratories (such as Bell Labs, where S was created and developed) to produce the next generation of computational innovations. Traditional university statistics departments do not necessarily provide the stimulating, supportive environment that encourages faculty members to conduct research in computational infrastructure. (The University of Auckland, New Zealand, which is the home of R, is a rare exception.) This must change. While the field needs only a few bright students to focus on computational infrastructure, it definitely needs them to be educated with the fundamentals so that they can easily specialize in this work.

More than just developing the infrastructure needed for statistics itself, statisticians must aim to influence and guide some of the technological innovations that are underway. For example, access to self-describing data with rich metadata describing its content and origins are at the heart of the semantic web. Statisticians should help to incorporate ideas that could revolutionize access to data for statistical decision making in this effort.

18.3 Statistical Experience

Most statisticians would agree that students need to know how to apply statistical methods, which is in many regards the goal of statistics education (ASA, n.d.; Bryce et al., 2001; Cobb & Moore, 1997). However, while many courses teach methodology—either the mathematics or the applied heuristics—very few focus primarily on teaching the skills of approaching a scientific problem from a statistical perspective. Instead, courses often focus on understanding methods and their characteristics in the belief that providing students with a set of tools and an understanding of those tools is the necessary background for using statistics in an applied setting. This is a rather big leap as there are so many other skills needed. For as Wild (2007) noted, “the biggest holes in our educational fabric, limiting the ability of graduates to apply statistics, occur where methodology meets context (i.e., the real world)” (p. 325).

Typically, the applications in statistics courses are merely examples where students have to identify the inputs and plug them into the method. Examples do not have the same rich, complicated context, extraneous information, and decision-making issues as real applications. Examples focus on methodology and ignore the many other dimensions needed to apply statistical ideas to a real problem. At other times, the applications are derived from data collected from students and, again, typically lack a real question and context. Applications should involve uncovering the relevant information, understanding the needs of the problem, drawing conclusions and understanding their limitations, incorporating less quantitative considerations, refining the goal, and communicating the essential findings. These steps are far broader than estimating parameters or performing an F-test. On a more specific level, one needs to break the task into steps, figure out how to combine the steps, and then perform each step; this involves very different skills than using a particular statistical method or tool. The methodology is one important detail in the bigger picture, but it is just one, and too often students miss this important experience.

While statistics students learn the mathematical fundamentals of the field, students in other fields are often learning to apply more modern statistical methods than we even teach our doctoral students (e.g., Gelman & Hill, 2007; Sekhon, n.d.). To ensure that statistics students remain relevant, they must have skills that cannot be replaced by reading a textbook of modern methods. Statistical experience is such a skill; it gives students the skills needed to become collaborators in scientific endeavors. With these skills they can work as part of a team and bring a particular way of thinking about a problem, along with a different set of tools that they have experienced in real situations.

For students learning data analysis and statistics, the intuition and experience that are used in data analysis are the hardest things to learn and the elements that are least often taught. Of course, it is also the hardest thing to teach, being somewhat subjective, context-specific, and an art. But we cannot shy away from this difficulty, and we can attempt to distill the more objective aspects. At the very least, students need to be exposed to the paths that are followed during a real analysis and understand the statistician's thinking and decision-making process.

18.3.1 When and Where do Students Encounter the Experience Component?

Intuition and experience of methodology in the scientific context are essential to this thought pattern. Ironically, these are rarely presented in books—in sharp contrast to the large collection of textbooks that offer similar, formal descriptions of common methods. Many statisticians would do well to adopt the pedagogical technique where students read material outside of class and the professor spends time on material not in the book. Wild and Pfannkuch (1999) noted that to teach statistical thinking we often simply "let them do projects" (p. 224). Although a valuable exercise, as a single, unguided encounter with statistical thinking in a real setting, it is far from adequate. Another approach is to leave the statistical experience until after they have learned the basic, traditional tools of statistics, such as probability, hypothesis testing, estimation, and inference. This might be in a capstone course for undergraduate majors. Again, this single exposure pales in comparison to opportunities appearing in multiple courses earlier in their studies. For graduate students, consulting courses in which clients bring statistical problems to a class have potential. However, the problems are of varied quality and interest and somewhat random in the lessons taught.

Another statistical experience teaching venue is its integration into an existing methodology or applied course via case studies (Brady & Allen, 2002). Case studies of data analyses often hide much of the thought process that is required. In a case study, an analysis is typically presented as a completed work, but the thought process that led to the conclusions and choice of approaches and methods is usually omitted. There is rarely mention of alternative approaches that were not fruitful or that were not pursued and the reasons why. Also not typically identified are the alternative approaches that led, or would lead, to almost equivalent solutions to the one presented in the final analysis.

Another option is for the practical experience to be inserted as tangents in the flow of a course that show how a particular method came to be used. Again, it is difficult in this scenario to get away from the use of a particular method (i.e., the one just learned) to the data at hand. However, it can be effective if it truly emerges from a scientific problem and includes contrasts with other approaches and methods that may be applicable (Nolan & Speed, 1999). It takes time to prepare and to teach, so one must decide if it is worth it. Unfortunately, we may be allowing our decision making to be clouded by limited time rather than the good of the students.

Some graduate programs require students to take a minor subject for a year in which they apply statistics to problems in that field. This seems to be successful, but these schools are usually quite forward looking and already have a broader view of statistics. Many students, both graduates and undergraduates, take a minor in statistics while majoring in another subject. Why is statistics typically the minor? One reason is that students believe they can make more of a difference in their work in other fields; this partly comes from being exposed to actual applications of statistics and appreciating its challenges and impact.

Yet another opportunity for displaying statistical thinking and imparting experience is in the introductory statistics course. Instead of assuming that this course is the only chance to teach students statistics and so must cover a long list of fundamentals; a more novel and potentially more effective approach would be to teach backwards; that is, rather than students learning methods as formulae and applying them to draw conclusions, compelling scientific and social problems are presented for students to grapple with, debate, and make decisions based on data exploration. With a well-guided discussion that pushes students to justify opinions and conclusions, they can discover and deduce commonsense statistical concepts and methods. We might do well to recognize that (a) there are students who may be interested in studying statistics but who do not know much about it, and (b) these students can be attracted to the field by showing them the bigger picture of how statistics is used and in what ways it is important. When students grapple with intellectually demanding questions and discover personal expression and creativity in the statistical experience, rather than the drier material of a traditional introductory statistics course, they may be attracted to the field or at least gain an appreciation for it.

The occasional course that presents different aspects of a broader view of statistics might get students thinking in new and interesting ways, and foster activity and innovation. Courses entitled *Weird Science* and *Disruptive Technologies* at the University of Texas, Austin, are both thought provoking and engaging. Similar experiments could be tried in statistics.

18.3.2 What Should Be Taught?

While there is much variability in how statisticians operate, a statistician often approaches a consulting problem or scientific collaboration in ways that can be abstracted. From interviewing practitioners and researchers, Wild and Pfannkuch (1999) identified four dimensions of statistical thinking: the investigative cycle, types of thinking, the interrogative cycle, and dispositions. Their framework would complement the mathematical models used in analysis and address areas of the process of statistical investigation that the mathematical models do not, particularly areas requiring the synthesis of problem-contextual and statistical understanding.

We offer a concrete list of these aspects of statistical thinking that captures the elements of a typical data-analysis process: decompose the problem and identify key components, abstract and formulate different strategies, connect the original problem to the statistical framework, choose and apply methods, reconcile the limitations of the solution, and communicate findings. There are many nuances that we have omitted; and it is a subjective, informal process. Yet, the overlap between our list and the US National Research Council (US NRC, 1996) science education standards is notable:

> Inquiry is a multifaceted activity that involves making observations; posing questions; examining books and other sources of information to see what is already known; planning investigations; reviewing what is already known in light of experimental evidence; using tools to gather, analyze, and interpret data; proposing answers, explanations, and predictions; and communicating the results. (p. 23)

We consider each of these aspects in more detail.

18.3.2.1 Decompose the Problem and Identify Key Components

The statistician asks the scientist questions about the general subject matter to get the context of the problem and to understand the goals. As they interact, there is a discussion that iterates between what the scientist wants and what the statistician needs to know more generally about the problem. The conversation often revisits earlier topics to get more information that earlier did not seem necessary or occur to the discussants. Students need to learn how to identify this information. It is hard to mimic an interaction with a scientist, but making the information available from this process in the document-database and summer statistics program (described earlier and also later) is important.

18.3.2.2 Abstract and Formulate Different Strategies

As more details are uncovered, the statistician is collecting potential approaches, identifying potential problems or additional information needed, and formulating strategies by mapping the scientific problem to statistical approaches. This high-level work involves classification, prediction, or parameter estimation. As more information becomes available, particular techniques come into the picture (e.g., CART or *k*-nearest neighbors, generalized linear model or GLM). There are many possible methods for each high-level statistical goal, and there may be several statistical goals that lead to an approach to the more general scientific problem. Exposing the dynamic picture the statistician builds as the investigation proceeds would be very valuable to students.

18.3.2.3 Connect the Original Problem to the Statistical Framework

A key component at this stage is understanding what is really of interest. If the goal is to make a decision about a particular social phenomenon, the statistician must understand the decision space, that is, for what outcomes is there the same decision and where is the boundary between the different decisions. These will ultimately be determined by the available data and the statistical methodology that is used. However, the boundary may be quite invariant with respect to the different statistical techniques being used; differentiating between various methods may not be of importance. In other cases, the sensitivity of the decision process may be quite extreme and requires very careful understanding of the choices. The decision-making process is not determined entirely by the statistical approach but by the context in which the decision is being made. The particular method or the actual parameter estimates may not be important, but rather their impact further up the decision-making chain is. Further, before pursuing a statistical approach, it is important to understand the accuracy that is needed. For example, there can be a large difference between estimating a value to within 2 m or to within 2 cm. Quite different data or a different approach may be required for the latter whereas any, even an ad hoc, method may suffice for the former.

18.3.2.4 Choose and Apply Methods

With the array of possible directions, the statistician will prioritize these, realizing that some are easy, some are more complex, and some are long shots that may not lead anywhere. When mapping the problem to a statistical formulation, it is useful to consider if the data will be able to answer the problem in this setting. The statistician considers the limitations of the data and whether other data are needed and accessible. After choosing a statistical approach, the statistician performs the necessary computations to obtain the results. In so many ways, this is really the easy part of the process. Unfortunately, picking and fitting the model is the common focus of most courses.

18.3.2.5 Reconcile the Limitations of the Solution

With the statistical results in hand, the statistician puts them back into the decision-making context to evaluate whether a conclusion can be made and to determine the possible limitations of that conclusion. This information leads to an updated formulation of the problem so that the statistician can iterate through the entire process to get better results leading to a final conclusion.

18.3.2.6 Communicate Findings

Finally, the statistician communicates the results in a meaningful, interesting way. While the effectiveness might be related to the subject matter, the presentation needs to be engaging, clear, and context-specific in telling an important story.

18.3.3 What are the Challenges?

Some reasons for the lack of change in this direction in the statistical curricula are: (a) the difficulty in finding good problems that are both compelling and at the right level for students, (b) conservatism in the field coupled with a deep-rooted belief that mathematics is the most important topic for students in spite of dramatic technological changes, (c) the effort in preparing course materials (i.e., providing all the details of a compelling analysis) and concern that they will be out of date quickly, and (d) unfamiliarity with teaching from this approach. There is no question that it takes energy, time, and creativity to successfully convey the statistical experience. Models of how to begin to think about making the change are needed. We present two different models here. We have experimented with a summer program where undergraduates engage in research problems and statistical thinking through data visualization. The second is an approach where research papers are augmented by instructors to create interactive, dynamic documents that guide students in data analysis and statistical thinking.

18.3.4 The Summer Statistics Program

In June of both 2005 and 2006, the authors, along with Mark Hansen of the University of California, Los Angeles (UCLA), organized and conducted a 1-week workshop at UCLA. Each year, about 25 undergraduate students, 5 graduate students, 3 researchers, and 3 additional statisticians participated. The undergraduates ranged in background from statistics majors to promising students who had taken only one statistics course but who were keen to see a broader side of statistics. On the first day, we taught essential computing skills, thus enabling students to manipulate data, extract subsets, and create numerical and graphical displays.

The workshop format involved different research statisticians introducing a scientific problem in which they were active contributors. This typically involved motivating the problem and its importance, discussing the challenges (e.g., what data were needed and available), what outcomes were feasible, and how to start thinking about the problem from a statistical perspective. Each high-level problem was typically broken into about six steps over 2 days. The students first familiarized themselves with the data and explored it. Typically, there were some suggestions of aspects to explore; as the researchers moved between groups and focused discussions, new ideas emerged. After about an hour, each group presented something of interest, which the class discussed and critiqued. After this more free-form exploration, the researcher guided the students through a particular statistical formulation of the problem and explained how this would help lead to a result that could be used to address the problem. The remainder of the session oscillated between explaining some statistical techniques and sending the students off in groups or individually to use these techniques on the data to solve the problem. Again, there were several different approaches to explore; students either selected those that interested them or groups agreed to try different approaches for comparison purposes. Having a discussion at the end of each breakout session ensured the students were engaged and acted as participants in the problem-solving activity. They exhibited a sense of involvement and creativity to find an unusual perspective or hidden feature—in contrast to the more typical, computational, data-analysis course laboratories where students complete tasks or exercises in a prescribed manner. (We were coincidentally fortunate to have the opportunity to observe this more traditional approach within one morning of the same workshop, where the students' reactions were markedly less enthused.)

There have been several outcomes from this summer school. Firstly, it exposed students to a very different aspect of statistics than they had experienced in their courses. We found that the focus on real problems with statistical thinking, as opposed to learning about methods, interested the students and motivated them to learn more about the methods. Secondly, the experience confirmed for us that one can teach statistics in this manner to good effect. The students were able to quickly master sufficient computing skills so that they could then work relatively independently and be in charge of exploring their own creative ideas. We found that, while statisticians often teach as if the mathematics provides intuition for students, in these workshops students were able to rapidly grasp statistical concepts and intelligently apply methodology when described more heuristically. For some, a subsequent mathematical description helped to clarify the idea. Furthermore, they were able to suggest adaptations of the methods and were unencumbered by mathematical formalities and the sense of there being a unique, correct answer. A sense of creativity and a can-do attitude, even if erroneous, are desirable attributes of statistics students.

One very important outcome of the summer school is the case studies of the presentations in the form of an extended laboratory. We often use one or two as exercises or entire projects in our regular courses. This flow from researcher to instructor to student, where there are some mutually beneficial gains for the researchers and instructors, leads to very rich and somewhat unique teaching materials. However, there is no doubt that it is time consuming to gather this material. Two of the three organizers worked with each presenter before the workshop so as to understand the scientific problem, reconstruct the analysis and the computations in R, discuss how to decompose the topics for the students, and often provide higher-level functions and preprocessed data to expedite analysis within the short period available. This process consumed several person-weeks; it would be greatly beneficial to reduce this time and to access more potential applications without burdening the researcher directly. This is one motivation for the database document concept described in Sect. 18.3.5.

18.3.5 Reproducible, Dynamic, Interactive Documents

A second, novel avenue that we are pursuing in the area of teaching statistical experience is to provide infrastructure that induces a passive flow of research and case-study documents from researchers to educators. The vehicle is a *database document* or a reproducible, dynamic, interactive document. The essential idea is to enable authors to document their actual computations and analyses, along with notes and explanations of their thoughts and decisions, so that the analytical process is reproducible. Authors use the document as an electronic notebook that captures their writings, computations, thought process, and notes or comments. This document is not what is intended for the readers, but a database or repository of all the activities within the analysis or task. The document can be projected into different views to make articles, papers, stories, code, and software available for distinct audiences. Readers can switch between the projected view and access the details at various resolutions, from seeing data to the general computational flow to specific lines of code. One can replay all the computations up to a specific point or change inputs and recalculate the tables and figures.

This style of documenting one's work aids researchers by allowing them to archive material in a structured manner rather than the more personal style currently in use. Critically from a pedagogical perspective, it makes real analyses and applications of statistics available to educators in a manner that can be easily used for teaching students about this subtle, elusive process. The collection of such documents has vague similarities to open-source software, which has served the statistical community very well with R and S-PLUS, and software repositories at Comprehensive R Archive Network (CRAN, n.d.) and StatLib (n.d.). Here, the idea is to share details of analyses across discrete communities, allow analyses to be used for different purposes, encourage greater verification and understanding of results, facilitate further extensions of approaches, and enable students to observe and participate in the statistical experience.

Instructors can take such a document and know that it has all the details involved in a real analysis. They can annotate the material with links to explanations of the science and the statistical terms. They can annotate the computations (either programmatically or manually) to identify the inputs and outputs of the different subtasks. Such annotations can be used to display interactive controls for students who can then control various aspects of the computations—set nuisance parameters to different values, remove subsets of the data, introduce alternative datasets, create new plots, or introduce entirely different ways of analyzing data (e.g., using a different classification method in one step of the overall analysis). This is the interactive aspect of the document, which allows for student control via graphical user interface (GUI) elements rendered when displaying a projection of the original document. It provides a semiguided exploration of the details that can go on to delve deeper and eventually go to free-form analysis.

Our goal is for students to experience the thought process of the *masters* in context, seeing their choices, approaches, and explorations. We want to avoid simplifying the scientific–data problems; instead, we want to simplify how students see these details initially while allowing them to gradually see them to their full extent to experience the reality of statistical practice. As Wild (2007) noted, these documents give instructors a mechanism to (a) control complexity and other environmental settings, (b) provide multiple pathways to explore, (c) focus attention on what is new and accelerate through what has already been mastered, (d) allow students to efficiently unlock the stories in data, and (e) encourage students to just try things out.

This system (see Nolan & Temple Lang, 2007) is based on widely used and standardized technologies and frameworks; it readily supports multiple and different programming languages. Because it is highly extensible, it allows adaptation and will accommodate future developments (e.g., different aspects of the analysis process). The approach is to create a programming and authoring environment designed for professional statisticians that supports communication of statistical results and the data analysis process. The document created by the statistician would be both interactive and dynamic—dynamic in this case meaning that the code for the analysis and plots is contained in the document and this code is run to create the view of the document. The document is interactive in that the reader can control the embedded computations by, for example, dragging a slider that leads to code reevaluation and subsequent update and redisplay of the output.

Our prototype is based upon the R computational environment. The document is a collection of text, code, output, and other metadata and is written in XML—XML syntax is similar to HTML, having elements or nodes of the hierarchical form. The XML document-database can be converted to a variety of formats, such as PDF, HTML, and what we call interactive HTML. R packages provide the functionality to transform and display these XML documents. The interactive controls are provided by a general purpose GUI toolkit called wxWidgets, which is also available from within R via a plug-in package. Information can be programmatically queried and extracted from the document database via R commands that identify the XML nodes of interest. While XML underlies the representation of the document, these documents can be authored without any knowledge of XML using tools such as Microsoft Word. However, the richness, flexibility, extensibility, and generality emerge from the XML infrastructure.

One might think that this is yet an additional burden on the author and so is unlikely to be adopted. We are more optimistic because essentially this archiving of the actual computations and noting of ideas, decisions, and thoughts is what is done more informally in every analysis. Statisticians store code used in the computations in separate directories and files, adding comments to LaTeX or Word documents as notes to themselves. At the very simplest, we are describing a system that facilitates such archiving and provides ways to retrieve and manage the elements, allows extraction of notes and code for other uses, and simplifies the creation of documents from the centralized master document—an important feature as XML and related technologies continue to dominate in publishing. The adoption of Sweave (Leisch, 2002) within R for dynamic documents illustrates that people are willing to use such tools. Our approach is a more general notion of a document with Sweave essentially as a special case. We provide a much richer concept of a document acting as a database rather than merely as a dynamic, presentation-based, document mechanism. The concept of having many other dimensions within the document makes it much richer. However, not all documents are required to have these extra dimensions; they can be added after the document is first authored, which allows authors to gradually move from Sweave-like use to leveraging these extended facilities as appropriate. The use of standard, ubiquitous technologies makes it more broadly applicable across different communities and more amenable to interesting extensions.

18.4 Summary

Computing and statistical thinking and experience are very important elements of a statistics education. To bring these elements into statistics curricula, statisticians must think boldly, unconstrained by legacy, starting from a blank slate and bringing back the best of the existing curricula along with new important topics. For example, instead of reteaching the same concepts at progressively higher levels of mathematical abstraction, the time gained could be used to teach other topics, including computing, statistical experience, and modern statistical methods. Perhaps more importantly and ambitiously, once the entire curricula is evaluated from the viewpoint of what is no longer needed because the available computational power is so much greater, many topics can be streamlined or eliminated entirely. As Cobb (2007) stated:

> [A] huge chunk of statistical theory was developed in order to compute things, or approximate things, that were otherwise out of reach. … The computer has offered to free us and our students from that, but our curriculum is at best in the early stages of accepting the offer. (p. 7)

Teaching computing and statistical thinking is very hard. We have outlined various approaches that attempt to make it easier for individual instructors to introduce and teach this material within the statistics curriculum. Common to all of them is the notion of pooling resources across one or more communities. For computing courses, we are working to create model syllabi and documents to discuss the importance of different topics for different types of students. Also, we are working to create an archive for tutorials, chapters, case studies, course notes, videos for use within courses; and we are holding workshops for faculty on how to teach this material. The intent is to enlarge the community of instructors capable of, and willing to, teach statistical computing by leveraging the existing small community of those who already do.

To aid the teaching of statistical reasoning and experience, we aim to unite the research community and instructors by providing a flow of real-world data analyses from the former to the latter. This is done by providing an infrastructure for reproducible results for the researcher that allows the capture of computational details and thought process in an electronic notebook that acts as a project database. While this is beneficial to individual researchers and their community of fellow researchers and reviewers, it is also useful to course instructors. These documents allow students to enter the world of the researcher and to engage in the research process. Much remains to be done before this approach is complete and effective; software must be written, and communities must be engaged. However, the infrastructure is in place to achieve these ends.

The suggestions in this chapter represent more than incremental changes motivated by constrained resources and conservatism. Computing, the Web, the digital world, and interdisciplinary science present a changepoint for the field of statistics and require statisticians to think about what a modern statistics curriculum would look like if they had both the freedom to change and resources to implement. For too long, the field of statistics has acted more passively to such changepoints and responded by merely adding topics to courses—and not seeking, considering, or embracing new paradigms. For statistics to flourish in this new era of science and technology and to have the impact that it could and should, educators must seize the opportunity to move the field of studies towards the modern needs of scientific research with data.

References

American Statistical Association. (n.d.). *Curriculum guidelines for undergraduate programs in statistical science*. Retrieved May 14, 2008, from http://www.amstat.org/education/index.cfm?fuseaction = Curriculum_Guidelines

Brady, J. E., & Allen, T. T. (2002). Case study based instruction of DOE and SPC. *The American Statistician, 56*(4), 312–315.

Bryce, G. R., Gould, R., Notz, W. I., & Peck, R. L. (2001). Curriculum guidelines for bachelor of science degrees in statistical science. *The American Statistician, 55*(1), 7–13.

Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education, 1*(1). Retrieved from http://repositories.cdlib.org/uclastat/cts/tise/vol1/iss1/art1/

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly, 104*(9), 801–823.

Comprehensive R Archive Network (CRAN). (n.d.). *Homepage*. Retrieved July 1, 2008, from http://cran.r-project.org/

Gapminder. (n.d.). *Homepage*. Retrieved July 1, 2008, from http://www.gapminder.org/

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Gentle, J. E. (2004). Courses in statistical computing and computational statistics. *The American Statistician, 58*(1), 2–5.

Hansen, M., Nolan, D., & Temple Lang, D. (2006, June 17–24). *Data visualization and its role in the practice of statistics* [second annual Undergraduate Summer Program in Statistics]. Retrieved May 14, 2008, from http://summer.stat.ucla.edu/

Hansen, M., Nolan, D., & Temple Lang, D. (n.d.). *Model courses and curricula*. Retrieved May 14, 2008, from Computing in Statistics Wiki: http://www.stat.berkeley.edu/twiki/Workshop/CompCurric

Insightful Corporation. (n.d.). S-PLUS® 8 Enterprise Developer [Computer software]. Seattle, WA: Author. Available from http://www.insightful.com/

Lange, K. (2004). Computational statistics and optimization theory at UCLA. *The American Statistician, 58*(1), 9–11.

Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. In W. Härdle & B. Rönz (Eds.), *Compstat 2002: Proceedings in computational statistics* (pp. 575–580). Heidelberg, Germany: Physika Verlag.

Monahan, J. (2004). Teaching statistical computing at North Carolina State University. *The American Statistician, 58*(1), 6–8.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review/Revue Internationale de Statistique, 65*(2), 123–137.

Nolan, D., & Speed, T. P. (1999). Teaching statistics theory through applications. *The American Statistician, 53*(4), 370–375.

Nolan, D., & Temple Lang, D. (2007). Dynamic, interactive documents for teaching statistical practice. *International Statistical Review, 75*(3), 295–321.

R Development Core Team. (2006). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rossman, A., & Chance, B. (2003). *Notes from the JSM 2003 session "Is the Math Stat course obsolete?"*. Available from American Statistical Association Section on Statistical Education Website: http://www.amstat.org/sections/educ/SFPanelFinal.doc

Sekhon, J. S. (n.d.). *Quantitative methodology in the social sciences seminar* (Political Science 239 course syllabus). Available from http://sekhon.berkeley.edu/seminar/syllabus_ps239.pdf

Society for Industrial and Applied Mathematics (SIAM) Working Group on CSE Education. (2001). Graduate Education in Computational Science and Engineering. *SIAM Review, 43*(1), 163–177.

StatLib. (n.d.). *Data, software and news from the statistics community*. Retrieved July 1, 2008, from http://lib.stat.cmu.edu/

United States National Research Council. (1996). *The national science education standards*. Washington, DC: The National Academies Press. Available from http://www.nap.edu/catalog.php?record_id = 4962

Wild, C. J. (2007). Virtual environments and the acceleration of experiential learning. *International Statistical Review, 75*(3), 322–335.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review/Revue Internationale de Statistique, 67*(3), 223–248.

Chapter 19
Dr. Fox Rocks: Using Data-mining Techniques to Examine Student Ratings of Instruction

Morgan C. Wang, Charles D. Dziuban, Ida J. Cook, and Patsy D. Moskal

Few traditions in higher education evoke more controversy, ambivalence, criticism, and, at the same time, support than student evaluation of instruction (SEI). Ostensibly, results from these end-of-course survey instruments serve two main functions: they provide instructors with formative input for improving their teaching, and they serve as the basis for summative profiles of professors' effectiveness through the eyes of their students. In the academy, instructor evaluations also can play out in the high-stakes environments of tenure, promotion, and merit salary increases, making this information particularly important to the professional lives of faculty members. At the research level, the volume of the literature for student ratings impresses even the most casual observer with well over 2,000 studies referenced in the Education Resources Information Center (ERIC) alone (Centra, 2003) and an untold number of additional studies published in educational, psychological, psychometric, and discipline-related journals.

There have been numerous attempts at summarizing this work (Algozzine et al., 2004; Gump, 2007; Marsh & Roche, 1997; Pounder, 2007; Wachtel, 1998). Student ratings gained such notoriety that in November 1997 the *American Psychologist* devoted an entire issue to the topic (Greenwald, 1997). The issue included student ratings articles focusing on stability and reliability, validity, dimensionality, usefulness for improving teaching and learning, and sensitivity to biasing factors, such as the *Dr. Fox* phenomenon that describes eliciting high student ratings with strategies that reflect little or no relationship to effective teaching practice (Ware & Williams, 1975; Williams & Ware, 1976, 1977).

Because of the persisting interest in student ratings, a comprehensive assortment of measurement and psychometric techniques serve as analysis models for assessing these data. Latent-trait approaches incorporated factor and component analysis in an attempt to resolve the dimensionality issues associated with these responses (Bangert, 2006; Clayson, 1999; Cohen, 2005; Feldman, 1976; Lannutti & Strauman, 2006; Marsh & Roche, 1997; Smith & Anderson, 2005). Some investigators developed hypothesis-based dimensionality studies using

M.C. Wang, C.D. Dziuban, I.J. Cook, and P.D. Moskal
University of Central Florida

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

confirmatory and hierarchical factor models; others used methods such as cluster (Ginns & Ellis, 2007) and smallest space analysis (Cohen) to define teaching profiles for effective instructors (Abrami & D'Apollonia, 1991; Apodaca & Grad, 2005; Ginns, Prosser, & Barrie, 2007). Elegant reliability studies (Chang & Hocevar, 2000) using generalizability theory resolved ratings into variance components for students, instructors, course level, items, and actions trying to account for the fact that students are nested within instructors. Other investigators incorporated classical test theory (Cook, Gelula, Dupras, & Schwartz, 2007; Lannutti & Strauman; Ustünlüoglu, 2007; Wilson, 2006).

Causal and predictive approaches applied methods such as path analysis and structural equation modeling (Chang, 2000; Ginns et al., 2007; Greenwald & Gilmore, 1997; Renaud & Murray, 2005; Rinderman & Schofield, 2001; Shevlin, Banyard, Davies, & Griffiths, 2000) that augmented more traditional regression and correlational analysis (Cohen, 2005; Davidovitch & Soen, 2006; Eiszler, 2002; Nasser & Fresko, 2006; Read, Rama, & Raghunandan, 2001; Renaud & Murray; Sheehan & DuPrey, 1999; Stapleton & Murkison, 2001). A large body of research featured hypothesis-testing models such as analysis of variance (Crumbley, Henry, & Kratchman, 2001; Maurer, 2006; Renaud & Murray; Riniolo, Johnson, Sherman, & Misso, 2006; Smith & Anderson, 2005) and chi-square contingency analysis (Howell & Symbaluk, 2001). In addition, an important approach to SEI involves deductive analysis typified by studies that incorporate criticism techniques to clarify the role of student ratings in the instructional process (Gump, 2007; Kolitch & Dean, 1999; Oliver & Sautter, 2005; Pounder, 2007). Any attempt to summarize this body of research converges on defining robust elements that underlie students' conceptions of instruction in higher education.

19.1 Student Ratings in the World of Web 2.0

During the recent decade, the emerging Internet—and in particular the concept of Web 2.0 (see http://www.oreilly.com)—impacted students' evaluations of their instructors. This phenomenon is interacting with a generation of young people on campus who have been alternatively termed millennials, the net generation, the digital generation, and generation Y, among others. Their learning and technology characteristics are described as operating at twitch speed (miniscule response time), using parallel processing for information intake, preferring information in graphic rather than textual form, using their digital, personal, and mobile technologies to stay continually connected, preferring active rather than passive learning scenarios, incorporating play into their working lifestyles, embracing learning through virtual environments, and seeing technology as fun rather than a challenge (interested readers may see http://www.marcprensky.com).

For them, the Web 2.0—with its sharing, communicating, blogging, text messaging, social networking, group writing though wikis, and interactive social opportunities—is a seamless and continuous communication medium. These developments present a learning model far different from one-directional, teacher-to-student techniques that served as the prototype for most SEI research of the past decades. Today's students experience education though online and blended courses (partly face-to-face and partly online) and extending devices, such as podcasts, chat rooms, and worldwide virtual collaborative groups.

These trends have implications for students and their instructors. One example of emerging issues is the Web site http://www.ratemyprofessors.com where students formed a worldwide community to share their perceptions about their instructors' teaching abilities. Further, they share their impressions on social networking tools, such as Facebook (http://www.facebook.com) and MySpace™ (http://www.myspace.com), or post videos of their instructors in the act of teaching on YouTube (http://www.youtube.com). On many campuses students rate their professors online rather than using the paper-and-pencil scansheets of old. Students respond, not only to their face-to-face courses, but evaluate any number of technology-mediated classes in which they might be involved.

These emerging trends make it even more important to explore elements that underpin effective teaching in the eyes of students. In order to do this, the authors explored the use of data-mining techniques to develop rule-based models that best predict what students consider excellent and poor teaching in the academy.

19.2 Data for the Present Study

The University of Central Florida (UCF) administers an end-of-course student evaluation instrument. The *Student Perception of Instruction* (SPI) form is a 16-item, Likert-type device that students use to rate their instructors (e.g., excellent, very good, good, fair, poor). Respondents have the opportunity to provide written comments about the instructor, and considerable demographic information (course level, college, department, and instructor) can be obtained from the instrument because the class and date are recorded on the form. After classes end, instructors receive the original forms with student comments and a summary of course-rating responses. Presently, many students have an online response option as well.

The instrument comprises two separately designed item sets. A university-wide committee developed the first group of eight questions, and the Florida Board of Regents provided the second set of items that were common to Florida State University System institutions. However, this distinction of item sets is not evident on the instrument. Instructors may customize the form by adding items to the preset 16-item form. No other student demographic information is collected (e.g., anticipated grade). Table 19.1 provides the items of UCF's student rating instrument.

The UCF Faculty Senate authorized a study of the results from the instrument to explore its validity for assessing alternative instructional modes. Operationally, this study sought to determine which of a number of independent variables (demographic and rating response) would predict student response to an overall rating item for their instructor.

Table 19.1 Student perception of instruction items for the University of Central Florida

Source	Questions
Administration	1. Feedback concerning your performance in this course was:
	2. The instructor's interest in your learning was:
	3. Use of class time was:
	4. The instructor's overall organization of the course was:
	5. Continuity from one class meeting to the next was:
	6. The pace of the course was:
	7. The instructor's assessment of your progress in the course was:
	8. The texts and supplemental learning materials used in the course were:
Board of regents	9. Description of course objectives and assignments:
	10. Communication of ideas and information:
	11. Expression of expectations for performance:
	12. Availability to assist students in or outside of class:
	13. Respect and concern for students:
	14. Stimulation of interest in the course:
	15. Facilitation of learning:
	16. Overall assessment of instructor:

19.3 Data Collection and Ethics Protocol

The investigators assembled a dataset containing all student ratings of instructors for the 5 academic years beginning 1996/97 through 2000/01. The file contained 588,575 student records with responses to the 16 items and corresponding demographic information. The investigators reformatted the file so that it comprised only the responses to the 16 items (five levels) and indicators of course level (lower undergraduate, upper undergraduate, graduate), college (Arts and Sciences, Business Administration, Education, Engineering and Computer Science, Health and Public Affairs), and the academic year. No further identifying information was available in the analysis file. Throughout the study, the investigators preserved department and instructor anonymity. Therefore, this study investigated the independent measures, college, course levels, academic year, and items 1 through 15 (Table 19.1) on the SPI instrument for their ability to predict overall rating of the instructor (item 16).

19.3.1 About the Analysis

In order to explore these data, the authors incorporated decision trees (Breiman, Friedman, Olshen, & Stone, 1984), a data-mining technique that identified classification rules for an instructor receiving an excellent, very good, good, fair, or poor overall rating. Justification for the authors' approach is presented below.

First, decision trees are readily applicable to large datasets such as this. To deal with missing values, the user does not have to impute values because decision trees have built-in mechanisms, such as floating-category approaches implemented by Enterprise Miner™ (SAS Institute, 2008) and the surrogate method in classification and regression trees (CART, Breiman et al., 1984). For datasets such as this one, there are many missing values; and imputation is a very difficult, time-consuming task. Second, decision trees are among the most efficient methods for studying problems of this nature. For example, a logistic regression method cannot efficiently handle all variables under consideration. There are 18 independent variables involved here; 1 variable has three levels, the other 17 have five levels. This means the logistic regression model must incorporate 68 dummy variables and 2,278 two-way interactions. Even with today's computers, this is very difficult. On the other hand, the decision-tree approach can perform this analysis very efficiently since it needs fewer computer resources (e.g., computing time and memory) even if the investigator considers higher-order interactions. Third, decision trees constitute an appropriate method for studying this problem because many of the variables are ordinal in their scaling. Although we can assign numerical values to each category, assignment of values to each category is not unique. However, decision trees use the ordinal component of the variables to derive a solution analysis. Fourth, the rules found in decision trees have an *if–then* structure that is readily comprehendible. For example, one rule derived in the analysis found that students who selected the excellent category in both *Facilitation of learning* and *Communication of ideas* had a 96% chance of selecting the excellent category for the *Overall satisfaction* item as well (Table 19.2). Fifth, the quality of these rules can be assessed with percentages of accurate classification or odds ratios that can be easily understood. The final analysis procedure produces tree-like rule structures that predict outcomes. Customarily, researchers test the quality of the rules on a dataset independent of the one on which they were developed.

Table 19.2 Decision rules that lead to an overall instructor rating of "excellent"

	Rating					
Question	E	VG	G	F	P	Excellent (*p*)
Rule 1 (*n* = 46,805)						
Facilitation of learning	•					0.96
Communication of ideas and information	•					
Rule 2 (*n* = 3,462)						
Facilitation of learning	•					
Communication of ideas and information		•				0.85
Organization of the course	•					
Assessment of student progress	•	•				
Rule 3 (*n* = 6,215)						
Facilitation of learning				•		0.78
Communication of ideas and learning	•	•				
Organization of the course	•					
Instructor interest in your learning	•					

19.3.2 The Model-building Procedure for Predicting Overall Instructor Rating

For this study, the investigators used the CART method (Breiman et al., 1984), executed with SAS Enterprise Miner (SAS Institute, 2008). Because of its strong variance-sharing tendencies with the other variables, the dependent measure for the analysis was the rating on the item *Overall rating of the instructor*, with the previously mentioned indicator variables (college, course level, academic year, and the remaining 15 questions) on the instrument. Tree-based methods are recursive, bisecting data into disjoint subgroups called terminate nodes or leaves. CART analysis incorporates three stages: data splitting, pruning methods, and homogeneous assessment.

Data splitting into two (binary) subsets at each stage is the first feature of the model. For example, all students who selected the excellent category for *Facilitation of learning* were classified into a single category; all other students were classified into another subset. After splitting, the data in the subsets become more and more homogeneous. The tree continues to split the data until the numbers in each subset are either very small (i.e., say the number of observations is less than 100) or all observations in a subset belong to one category (e.g., all observations in a subset have the same rating). Typically, this *growing-the-tree* stage results in far too many terminate nodes for the model to be useful. The extreme case occurs when the number of terminate nodes equals the number of observations. Such models are uninformative because they produce very few rules with explanatory power.

The CART procedure solves this problem by using pruning methods that reduce the dimensionality of the system. In practice, CART splits the data into two pieces: the first dataset grows the tree, and the second prunes the tree, thereby validating the model. In practice, CART methods reduce the original tree into a nested set of subtrees. Although homogeneousness based on the training dataset can always be improved, it is not necessarily true in the validation set. Typically, because the validation data are not used in the growing process, they give an honest estimate of the best tree size.

The final stage of the analysis involves assessing homogeneousness in growing and pruning the tree. One way to accomplish this is to compute the misclassification rates. For example, a rule that produces a 0.95 probability that an instructor will receive an excellent rating has an associated error of 5.0%.

An important feature of this approach involves a performance assessment of the finally developed model—accomplished by the application of rules that have been developed and validated initially to an independently collected dataset. In this case, the data from the 1996 through 1998 academic years were developmental while additional data for the 1999/2000 and 2000/01 years provided the basis for model performance assessment. Accordingly, the model development used 424,498 observations, and the performance assessment used 164,077 independent records.

19.3.3 Consequences of Using Decision Trees

Although decision-tree techniques are effective for analyzing datasets such as this, the reader should be warned of consequences of the procedure. First, decision trees only use ranks to handle both ordinal and interval variables. At times, this might lead to lost distribution information about some variables—although the use of ranks does not create any information loss in this analysis. Second, decision-tree algorithms will combine categories if a given category variable has excessive partitions. For example, most decision-tree algorithms will combine mailing code into several groups before applying a split search. This feature, however, was not problematic in this study because no categorical variable used in this study had more than ten categories. Third, the most serious weakness of decision trees is that the results can be unstable because the technique is data-driven and small variations can lead to substantially different, final solutions. Techniques such as boosting (Schapire, 1990) and bagging (Breiman, 1996) provide some remedy to the instability of tree methodology. However, these treatments make interpretation of the rules much less intuitive, countermanding the fact that ease of interpretation is one of the most important advantages of decision-tree modeling. Therefore, we did not incorporate these techniques; instead, we used a logistic regression to confirm that the resulting rules exhibit strong validity.

19.3.4 The Rules for an "Excellent" Instructor Rating

The CART method developed three rules that predicted a high probability that an instructor would receive an overall rating of excellent while three other rules led to a poor rating. All six rules only used other questions on the SPI instrument and eliminated college membership, course level, and academic year. The final solution incorporated some combination of *Facilitation of learning*, *Communication of ideas and information*, *Overall organization of the course*, *Assessment of student progress*, *Instructor was interested in your learning*, and *Instructor showed respect and concern for students*. Table 19.2 displays the three rules that led to an overall excellent instructor rating.

Rule 1 indicates that if an instructor received an excellent rating on *Facilitation of learning* and *Communication of ideas and information* then the probability of receiving an excellent overall rating is 0.96, irrespective of college, course level, academic year, or responses to any remaining questions on the rating form. Since 41.8% of the instructors in the dataset received an excellent overall rating, the odds ratio for this rule is 2.29, indicating that instructors that conform to Rule 1 are 2.29 times as likely to get an excellent overall rating than a randomly chosen instructor.

The pattern for Rule 2 also signals instructors that are good candidates for an excellent overall rating (0.85). These individuals receive excellent for *Facilitation of learning*, very good for *Communication of ideas and information*, excellent for

Organization of the course, and excellent or very good for *Assessment of student progress*. The odds ratio associated with this pattern of responses is 2.03, indicating that these instructors are slightly over twice as likely to receive an excellent overall rating as one drawn at random.

The third rule also leads to a high probability (0.78) of an instructor being viewed excellent overall. This rule blends *Facilitation of learning* (excellent), *Communication of ideas and information* (excellent or very good), and *Organization of the course* (excellent) with an additional question: *Instructor was interested in your learning* (excellent). The odds ratio for this rule was 1.87. Of the 56,482 students whose ratings conformed to either Rules 1, 2, or 3, the largest percentage (82.9%) represented Rule 1, followed by Rule 3 (11.0%) and Rule 2 (6.1%).

19.3.5 The Rules for a "Poor" Instructor Rating

Three informative nodes produced substantially high probabilities that an instructor would receive an overall poor rating. There was a high correspondence among the questions of the SPI form that predicted an overall rating of excellent and an overall rating of poor. Once again, the poor rules used only other questions on the rating instrument and eliminated college, course level, and academic year. The poor rules replaced the question *The instructor was interested in your learning* that appeared in the excellent rules with *The instructor showed respect and concern for students*. Table 19.3 depicts the outcomes associated with the three rules.

Rule 4 illustrates if an instructor receives a fair or poor on the question *Facilitation of learning* and a poor on both *Communication of ideas and information* and *Instructor is interested in your learning* then the probability of an overall

Table 19.3 Decision rules that lead to an overall instructor rating of "poor"

	Rating					
Question	E	VG	G	F	P	Poor (p)
Rule 4 (n = 1,821)						
Facilitation of learning				•	•	
Communication of ideas and information					•	0.83
Instructor interested in your learning					•	
Rule 5 (n = 1,135)						
Facilitation of learning				•	•	
Communication of ideas and information					•	0.58
Organization of the course					•	
Rule 6 (n = 532)						
Facilitation of learning				•	•	
Communication of ideas and learning			•	•		0.54
Assessment of student progress					•	
Respect and concern for students					•	

rating of poor is 0.83. Because the percentage of instructors receiving an overall rating of poor in the dataset is 1.9%, the odds ratio for this rule is extremely high (43.6). This means that students classified in this category are significantly more likely to designate poor as the instructor's overall rating. However, the odds ratio of 43.6 might overestimate the magnitude of this likelihood.

Rule 5 states if an instructor receives a fair or poor on *Facilitation of learning* and a poor on both *Communication of ideas and information* and *Overall organization of the course*, then the probability of an overall rating of poor is 0.58. Although the probability of a poor rating with this combination of responses seems somewhat lower than the previous rule, one should note that the odds ratio associated with this rule is 30.3. This means this rule still has a significantly higher likelihood of giving the instructor an overall poor rating than a student randomly selected from the university.

Rule 6 indicates if an instructor's rating for *Facilitation of learning* is fair or poor, *Communication of ideas and information* is good or fair, and *Assessment of student progress* and *Instructor shows respect and concern for students* are poor, then the probability of an overall rating is 0.54 with an associated odds ratio of 32.3. The probability of an instructor receiving an overall rating of fair or poor for Rule 4 = 0.99, for Rule 5 = 0.97, and for Rule 6 = 0.96.

19.3.6 Model Validity

The investigators used three approaches to validating the decision-tree model—two logical and one statistical. The logical approaches involved harvesting all instructors across the university that conformed to the excellent and poor decision rules and examining the degree to which the rules leveled college differences. Table 19.4 presents the results of that procedure for excellent rules (academic years 1999/2000/2001). The unadjusted column depicts the percentages of overall excellent instructor ratings by college in the absence of the rules.

Ratings ranged from a high of 53.79% for Education to a low of 36.33% for Business Administration. The columns under *Adjusted for rule* portray the results when instructors across colleges are selected according to their compliance with the rules. In this case, the differences virtually disappear. Rule 1 produces overall excellent instructor ratings in the colleges, ranging from a high of 97.12% (Education) to a low of 95.03% (Business Administration). Rule 2 adjusts the excellent ratings from a high of 86.23% in Education to a low of 83.07% in Business Administration. Rule 3 produces a high of 80.05% in Arts and Sciences to a low of 74.00% in Health and Public Affairs. Table 19.4 demonstrates that college differences equalize around the proportions specified by each rule when instructor ratings conform to the rules that lead to a high probability of excellent.

Table 19.5 shows the impact of the poor rules on instructor ratings across the colleges. Again, the unadjusted column indicates the percentages of instructors that received an overall rating of poor, not taking into account the rules. Those

Table 19.4 Percentage of instructors receiving “excellent” overall ratings by college unadjusted and adjusted by rules 1–3

			Adjusted for rule					
College	Unadjusted	*n*	1	*n*	2	*n*	3	*n*
Arts and Sciences	41.83	31,914	95.30	19,699	85.41	1,358	80.05	2,419
Business Administration	36.33	12,463	95.03	7,950	83.07	628	77.12	974
Education	53.79	8,819	97.12	6,634	86.23	313	75.21	458
Engineering and Computer Science	32.19	4,434	95.52	2,604	83.71	185	78.16	365
Health and Public Affairs	47.80	11,138	96.13	7,894	85.53	455	74.00	632

Table 19.5 Percentage of instructors receiving “poor” overall ratings by college unadjusted and adjusted by rules 4–6

			Adjusted for rule					
College	Unadjusted	*n*	1	*n*	2	*n*	3	*n*
Arts and Sciences	2.31	1,761	78.95	630	57.14	264	54.44	135
Business Administration	3.01	1,033	85.87	383	63.36	166	50.37	68
Education	1.68	276	89.77	79	52.27	46	50.00	21
Engineering and Computer Science	4.81	662	83.69	272	56.25	90	62.90	39
Health and Public Affairs	1.89	441	87.80	144	53.99	88	53.33	24

unadjusted percentages ranged from a high of 4.81% in Engineering and Computer Science to a low of 1.68% in Education. Viewing the ratings according to the poor rules produces dramatic changes. Those instructors who conformed to Rule 4 were overall rated poor, ranging from a high of 89.77% in Education to a low of 78.95% in Arts and Sciences. Percentages of poor ratings for instructors that conformed to the pattern of Rule 5 showed the highest percentage of poor ratings in Business Administration at 63.22% with a low value found for Education at 52.27%. Finally, Rule 6 defines instructors who were rated poor in Engineering and Computer Science at a rate of 62.90% with the lowest value found in Education at 50.00%.

The second logical validation approach involved comparing the results of the UCF study with two national initiatives on teaching excellence. A model that identified seven principles of effective undergraduate education has gained widespread acceptance as a national standard for higher education (Chickering & Gamson, 1987). These seven principles describe an instructor who encourages contacts between faculty and students, develops cooperation and reciprocity, uses active learning techniques, gives prompt feedback, respects diverse talents and ways of thinking, emphasizes time on task, and communicates high expectations. A parallel initiative, the National Study of Student Engagement (Kuh, 2001), described five benchmarks: student interaction with faculty, collaborative learning, active learning, supportive environments, and academic challenge. Table 19.6 presents the correspondences between these two initiatives and the UCF-CART study. A comparison of these initiatives shows a close correspondence with components in each system grounded in facilitation of learning, instruction interest in student learning, effective communication, a well-organized learning environment, respect for students, and effective assessment of student progress.

In order to further examine the model validity, the investigators completed two separate logistic regression analyses. Table 19.7 presents the results of the analyses for the items contributing to an overall excellent rating of the instructor. All items selected by the decision tree contributed to the equation with the Wald chi-square probabilities rounded to 0.00. The model predicted with 97.6% accuracy producing a Somer's *D* of 0.945.

The logistic regression results for those items leading to a poor overall rating selected by the decision tree are presented in Table 19.8. Once again, the analyses showed that all items in the rules produced Wald chi-square values with associated probabilities rounded to 0.00. This equation produced a predictive accuracy of 97.6% with a Somer's *D* of 0.963.

Table 19.6 A comparison of the seven principles of good practice, the National Study of Student Engagement, and UCF's rule-based items

Seven principles of good practice (Chickering & Gamson, 1987)	National Study of Student Engagement (Kuh, 2001)	UCF rule-based items (applicable rule)
Encourages contacts between faculty and students	Student interaction with faculty	Facilitation of learning (1,2,3)
Develops reciprocity and cooperation among students	Collaborative learning	Instructor interested in your learning (3)
Uses active learning techniques	Active learning	Communication of information and ideas (1,2,3)
Gives prompt feedback	Supportive environment	Well-organized course (2,3)
Respects diverse talents and ways of thinking		Respect and concern for students (6)[a]
Emphasizes time on task	Academic challenge	Assessment of student progress
Communicates high expectations		

[a] Poor rating on this item correlates with an overall rating of Poor.

Table 19.7 Logistic regression for "excellent" rule items[a]

Items	*df*	Coefficient	Wald χ^2	*p*
Intercept	1	0.895	23,385.1	0.0001
Interest	1	0.796	23,157.2	0.0001
Organization	1	0.798	23,903.1	0.0001
Assessment	1	0.847	12,454.4	0.0001
Communication	1	0.847	25,162.3	0.0001
Facilitation	1	1.092	33,328.6	0.0001

[a] Percent correctly predicted = 97.6, Somer's *D* = .963.

Table 19.8 Logistic regression for "poor" rule items[a]

Items	*df*	Coefficient	Wald χ^2	*p*
Intercept	1	0.171	349.9	0.0001
Interest	1	−0.691	6,472.9	0.0001
Organization	1	−0.919	12,650.7	0.0001
Assessment	1	−0.667	7,062.0	0.0001
Communication	1	0.924	13,598.8	0.0001
Facilitation	1	−0.961	14,867.4	0.0001

[a] Percent correctly predicted = 97.6, Somer's *D* = .963.

19.3.7 A Discussion of the Dimensions

In their book on facilitative teaching, Wittmer and Myrick (1974) provided characteristics for what students considered poor teaching. Those instructors were insensitive, cold, disinterested, authoritarian, ridiculing, arbitrary, sarcastic, demanding, punitive, and disciplinarians. The students described excellent teachers as good listeners, empathetic, caring, concerned, genuine, warm, interested, knowledgeable, trusting, friendly with a sense of humor, dynamic, and able to communicate effectively. This second list resonates with all three of the excellent rules.

Rogers (1993) described a facilitative teacher as one who creates a learning environment rather than simply transmitting knowledge. The key element in Rogers' theory of teaching emphasized the facilitator's empathetic understanding when he or she comprehended and valued a student's perceptions. Straus (1988) examined facilitation from a leadership perspective and built facets of the process that might be construed as a teaching model. His theory demonstrated seven components: sharing an inspiring vision, focusing on results process and relationship, seeking maximum possible involvement, designing pathways to action, bringing out the best in others, celebrating accomplishment, and modeling behaviors that facilitate collaboration. Not only do students respond positively to a facilitative class environment, several theories support facilitation as an effective teaching model.

The ability to communicate effectively has long been accepted as a standard for effective teaching. Our findings suggest that this ability is fundamental to an instructor being viewed positively by students. In fact, one of the terminate nodes we obtained involved only two items that led to a high probability of a poor overall rating. This happened when students rated an instructor high on interest in student learning but low on communication ability—most likely a frustrating and ambivalent situation for students.

The CART analysis suggests that students reward instructors who develop effective course organization and evaluation techniques. These components may be viewed as skills obtainable through professional development. Because of recent emerging modalities for classes (e.g., fully online, blended, Web-enhanced), course organization has gained prominent attention. In addition, instructors are under increasing pressure to make assessment of student learning an organic component in their courses. One should note, however, that organization and assessment impact an instructor's rating in the presence of facilitation and communication. By themselves, they are not strong enough to carry the instructor's rating.

Respect for students and interest in their learning weighted differently in the student evaluation process. Instructor interest in student learning contributed to an excellent rating while student perception of low instructor respect for them resulted in poor overall ratings. These results led us to conclude that a supportive class climate created by instructors is a strong motivating factor for students to view their class experience positively.

19.4 Conclusion

Classification and regression tree analysis of student rating of instruction appears to have lived up to the expectations we placed upon it. By efficiently handling missing data and multiple interactions, the procedure produced reasonably robust decision rules that identify qualities by which students characterize excellent and poor university instructors. Another advantage of the rule-based solution comes from the ability of multiple constituencies (students, faculty, administrators) to integrate results such as these into their decision-making processes. Decision-tree methodology provides compelling outcomes through probability statements, odds ratios, and misclassification assessment, thereby allowing users to judge the quality, usefulness, and opportunity costs found in rule-based outcomes. In addition to operational and specific if–then rules, another advantage derives from the tree-like structures that provide comprehensive and systematic solutions to examining student evaluation in complex systems.

This approach produces an interactive and recursive model whereby an individual (e.g., a dean of a college) might review the results of the analyses and through more extensive investigation examine the effects of additional independent variables (e.g., class size, laboratory sections, online classes) on student ratings. All these added values indicate that the decision-tree analysis, above all, is responsive to a number of elements in the emerging information society. Today's students and faculty live in a world of ambient fundability (Morville, 2005) that comprises a fast-moving society where anybody can find anyone or anything, anywhere, anytime. These findings put new pressure on faculty members to respond to students' information and communication needs while at the same time maintaining the rigor required by their disciplines. However, when decision rules portray excellent teachers as facilitative, communicative, organized, interested, and equitable, they configure a prototype learning situation far different from the traditional *paper chase* of a few years ago. One of the most efficient methods for building these profiles comes from data-mining techniques.

Seldom do new technologies replace old ones immediately but, rather, begin a complex pattern of interactions with them over a period of time. For instance, the workplace watercooler has gone digital although not completely. From the decision rules in this study, one might infer that digital networks are creating a collective intelligence in which problem solving becomes an activity of the commons where students expect a participatory learning environment. In the digital world, knowledge comes from real-life experiences (or their simulations) rather than from formal education. For many years, it was precisely that formal education to which SEI research directed its attention. Decision-tree analysis appears to work well as a flexible format for examining student responses as they evaluate a much more recursive learning environment.

Current higher education environments feature community, collaboration, and self-organization, which create learning climates that are cognitively complex, reliant on technology, and much less dependent on physical geography. Peer production becomes an important part of this new learning space, displacing many features of the academy as we have known it. The decision rules suggest that students wish to lessen the ambiguity they experience in their classes with the concomitant reduction of their ambivalent feelings toward higher education. They prefer active involvement because they participate in a highly interactive world that employs multiple learning facets.

The method of categorization and regression tress for analyzing student satisfaction with instruction is well suited to the evolving nature of higher education. By providing grounded decision rules, it avoids the difficulties encountered in nonobservable, latent-trait approaches and the prohibitive assumptions underpinning many predictive and hypothesis-testing procedures. Certainly, one must be sensitive to the fact that this method does not produce a one-time solution and that selection of the variables for inclusion in the analysis has a major impact on the results, thereby underscoring the need for context planning in such studies. In addition, investigators should be cognizant of the large number of observations required for these methods. However, the fact that the final results produce clearly interpretable rubrics permitting one to take actions on such issues as instructional design, curriculum planning, course offerings, and administrative policy bring data-mining techniques into the mainstream of higher education as a decision tool for the information age.

References

Abrami, P. C., & D'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness: Generalizability of "N = 1" research: Comment of Marsh. *Journal of Educational Psychology, 83*(3), 411–415.

Algozzine, B., Gretes, J., Flowers, C., Howley, L., Beattie, J., Spooner, F., et al. (2004). Student evaluation of college teaching: A practice in search of principles. *College Teaching, 52*(4), 134–141.

Apodaca, P., & Grad, H. (2005). The dimensionality of student ratings of teaching: Integration of uni- and multidimensional models. *Studies in Higher Education, 30*(6), 723–748.

Bangert, A. W. (2006). Identifying factors underlying the quality of online teaching effectiveness: An exploratory study. *Journal of Computing in Higher Education, 17*(2), 79–99.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. New York: Chapman & Hall.

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44*(5), 495–518.

Chang, L., & Hocevar, D. (2000). Models of generalizability theory in analyzing existing faculty evaluation data. *Applied Measurement in Education, 13*(3), 255–275.

Chang, T.-S. (2000, April). *An application of regression models with student ratings in determining course effectiveness*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED455311)

Chickering, A. W., & Gamson, Z. F. (1987). Seven principles for good practice in undergraduate education. *AAHE Bulletin, 39*(7), 3–7.

Clayson, D. E. (1999). Students' evaluation of teaching effectiveness: Some implications of stability. *Journal of Marketing Education, 21*(1), 68–75.

Cohen, E. H. (2005). Student evaluations of course and teacher: Factor analysis and SSA approaches. *Assessment & Evaluation in Higher Education, 30*(2), 123–136.

Cook, D. A., Gelula, M. H., Dupras, D. M., & Schwartz, A. (2007). Instructional methods and cognitive and learning styles in web-based learning: Report of two randomised trials. *Medical Education, 41*(9), 897–905.

Crumbley, L., Henry, B. K., & Kratchman, S. H. (2001). Students' perceptions of the evaluation of college teaching. *Quality Assurance in Education, 9*(4), 197–207.

Davidovitch, N., & Soen, D. (2006). Using students' assessments to improve instructors' quality of teaching. *Journal of Further and Higher Education, 30*(4), 351–376.

Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education, 43*(4), 483–501.

Feldman, K. A. (1976). The superior college teacher from the students' view. *Research in Higher Education, 5*(3), 243–288.

Ginns, P., & Ellis, R. (2007). Quality in blended learning: Exploring the relationships between on-line and face-to-face teaching and learning. *The Internet and Higher Education, 10*(1), 53–64.

Ginns, P., Prosser, M., & Barrie, S. (2007). Students' perceptions of teaching quality in higher education: The perspective of currently enrolled students. *Studies in Higher Education, 32*(5), 603–615.

Greenwald, A. G. (Ed.). (1997). Student ratings of professors [Current Issues]. *American Psychologist, 52*(11), 1182–1225.

Greenwald, A. G., & Gilmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209–1217.

Gump, S. E. (2007). Student evaluations of teaching effectiveness and the leniency hypothesis: A literature review. *Educational Research Quarterly, 30*(3), 55–68.

Howell, A. J., & Symbaluk, D. G. (2001). Published student ratings of instruction: Revealing and reconciling the views of students and faculty. *Journal of Educational Psychology, 93*(4), 790–796.

Kolitch, E., & Dean, A. V. (1999). Student ratings of instruction in the USA: Hidden assumptions and missing conceptions about 'good' teaching. *Studies in Higher Education, 24*(1), 27–42.

Kuh, G. D. (2001). Assessing what really matters to student learning. *Change, 33*(3), 10–19.

Lannutti, P. J., & Strauman, E. C. (2006). Classroom communication: The influence of instructor self-disclosure on student evaluations. *Communication Quarterly, 54*(1), 89–99.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187–1197.

Maurer, T. W. (2006). Cognitive dissonance or revenge? Student grades and course evaluations. *Teaching of Psychology, 33*(3), 176–179.

Morville, P. (2005). *Ambient findability: What we find changes who we become*. Sebastopol, CA: O'Reilly Media.

Nasser, F., & Fresko, B. (2006). Predicting student ratings: The relationship between actual student ratings and instructors' predictions. *Assessment & Evaluation in Higher Education, 31*(1), 1–18.

Oliver, R. L., & Sautter, E. P. (2005). Using course management systems to enhance the value of student evaluations of teaching. *Journal of Education for Business, 80*(4), 231–234.

Pounder, J. S. (2007). Is student evaluation of teaching worthwhile? An analytical framework for answering the question. *Quality Assurance in Education, 15*(2), 178–191.

Read, W. J., Rama, D. V., & Raghunandan, K. (2001). The relationship between student evaluations of teaching and faculty evaluations. *Journal of Education for Business, 76*(4), 189–192.

Renaud, R. D., & Murray, H. G. (2005). Factorial validity of student ratings of instruction. *Research in Higher Education, 46*(8), 929–953.

Rinderman, H., & Schofield, N. (2001). Generalizability of multidimensional student ratings of university instruction across courses and teachers. *Research in Higher Education, 42*(4), 377–399.

Riniolo, T. C., Johnson, K. C., Sherman, T. R., & Misso, J. A. (2006). Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *Journal of General Psychology, 133*(1), 19–35.

Rogers, C. R. (1993). The interpersonal relationship in the facilitation of learning. In M. Thorpe, R. Edwards, & A. Hanson (Eds.), *Culture and processes of adult learning* (pp. 228–242). London: Routledge.

SAS Institute. (2008). SAS® Enterprise Miner™ (Version 5.3) [computer software]. Cary, NC: Author. Available from http://www.sas.com/technologies/analytics/datamining/miner/index.html

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning, 5*(2), 197–227.

Sheehan, E. P., & DuPrey, T. (1999). Student evaluations of university teaching. *Journal of Instructional Psychology, 26*(3), 188–194.

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: Love me, love my lectures? *Assessment & Evaluation in Higher Education, 25*(4), 397–405.

Smith, G., & Anderson, K. J. (2005). Students' ratings of professors: The teaching style contingency for Latino professors. *Journal of Latinos and Education, 4*(2), 115–136.

Stapleton, R. J., & Murkison, G. (2001). Optimizing the fairness of student evaluations: A study of correlations between instructor excellence, study production, learning production, and expected grades. *Journal of Management Education, 25*(3), 269–291.

Straus, D. A. (1988). *Facilitative leadership: Theoretical underpinnings*. Cambridge, MA: Interaction Associates.

Ustünlüoglu, E. (2007). University students' perceptions of native and non-native teachers. *Teachers and Teaching: Theory and Practice, 13*(1), 63–79.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief overview. *Assessment & Evaluation in Higher Education, 23*(2), 191–212.

Ware, J. E., Jr., & Williams, R. G. (1975). The Dr. Fox effect: A study of lecturer effectiveness and ratings of instruction. *Journal of Medical Education, 50*, 149–156.

Williams, R. G., & Ware, J. E., Jr. (1976). Validity of student ratings of instruction under different incentive conditions: A further study of the Dr. Fox effect. *Journal of Educational Psychology, 68*(1), 48–56.

Williams, R. G., & Ware, J. E., Jr. (1977). An extended visit with Dr. Fox: Validity of student satisfaction with instruction ratings after repeated exposures to a lecturer. *American Educational Research Journal, 14*(4), 449–457.

Wilson, J. H. (2006). Predicting student attitudes and grades from perceptions of instructors' attitudes. *Teaching of Psychology, 33*(2), 91–95.

Wittmer, R. D., & Myrick, R. D. (1974). *Facilitative teaching: Theory and practice*. Pacific Palisades, CA: Goodyear Publishing.

Chapter 20
Process Execution of Writing and Reading: Considering Text Quality, Learner and Task Characteristics

**Huub van den Bergh, Gert Rijlaarsdam, Tanja Janssen,
Martine Braaksma, Daphne van Weijen, and Marion Tillema**

20.1 Studying Processes: What Research Questions Do We Need to Ask?

We have conducted systematic reflections, data reanalyses, and incorporated results from several studies to promote discussion, enhance understanding, and build theory. Two models guide our research and analyses: The Descriptive Interactive Process (DIP) model (Fig. 20.1, left), and the Experimental Interactive Process (EIP) model (Fig. 20.1, right). In the DIP model, the main idea is to study processes: What happens during task execution, and how does the process change accordingly? The complexity can be illustrated by adding three components to the model: (a) quality of the output—what variation in processes is related to variation in output quality?; (b) task characteristics—what degree do processes vary with task features (e.g., computer versus pen-and-paper writing)?; and (c) learner characteristics—what degree does the way skilled versus unskilled writers adjust their process to tasks vary?

In the EIP model (Fig. 20.1, right), the general aim is to detect the effect of interventions on processes: Do different instructional variables affect the target process differently? This model can be extended by adding the product variable—Do instructional variables affect the target process differently, and does the product quality vary accordingly?—and learner characteristics: Does the way instructional variables affect the target process vary with regard to learner characteristics? Do good writers profit as much from the experimental instruction as poor writers? Does the experimental instruction change the processes carried out while writing in the same way for good and poor writers?

H. van den Bergh
Utrecht University and University of Amsterdam

G. Rijlaarsdam, T. Janssen, and M. Braaksma
University of Amsterdam

D. van Weijen and M. Tillema
Utrecht University

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

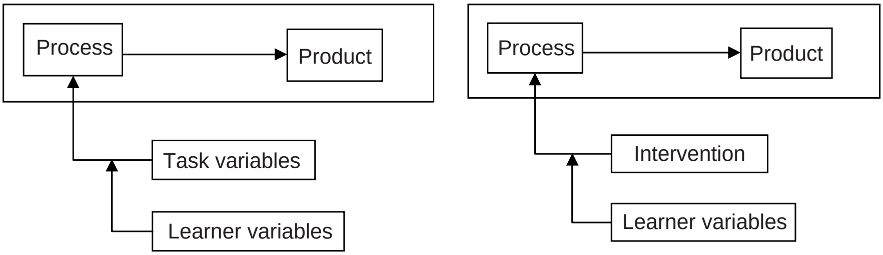


Fig. 20.1 The Descriptive Interactive Process model (*left*) and the Experimental Interactive Process model (*right*)

Both models are needed to understand the effects of educational interventions. We need to understand not only the relation between process and product characteristics, but also how we can influence the process characteristics that determine the quality of the resulting product through tasks and interventions for specific learners. The EIP model resembles in some ways the more traditional Aptitude Treatment Interaction (ATI) research (Cronbach, 1957, 1975). The main difference is that the interaction is more complex in many cases as it concerns an interaction between process characteristics (which change during task execution) and the result of these processes (quality of the product) and learner characteristics.

In this chapter, we present some cases and observations from our research on both models carried out in the last decade. We have chosen *observations* advisedly to indicate the tentative nature of our assertions. Process models for writing and literary reading subsume different distinguishable (sub)processes. Flower and Hayes (1980) suggested that writers have to plan what to write, put their ideas into words, and revise the product. The occurrence of such cognitive processes depends on the moment in the writing or reading process. We will show that it is worthwhile to distinguish each process (planning, formulating, revising, etc.) according to the moment it is carried out. It is, to give an example, rather unlikely that writers will start to revise before they have written a single word. Therefore, revising activities are less likely to occur in the beginning of the writing process. This kind of pattern appears to hold for all the cognitive activities we have studied thus far.

We model the occurrence of cognitive activities during reading and writing as a function of the moment at which they occur as ongoing, real-time processes. The occurrences of each cognitive activity need to be related to the quality with which the task is executed. Next to the occurrence of activities for individual readers and writers, the individual trajectories have to be related to outcome variables. These process–product relations form a fruitful basis for the interpretation of differences in the orchestration of cognitive activities; therefore, they form the foundation of theory development. We will show how flexible statistics address rather complex questions related to task and learning processes. Multilevel modeling (Goldstein, 2003; Snijders & Boskers, 1999) appears to be suited for answering these types of questions.

20.2 What We Want to Know about Processes

These models illustrate that the production and interpretation of print-based text involves a variety of processes dealing with making sense. Writing involves the global task of reporting ideas in a permanent form (print using established grammars) that may also shape the writer's conception and understanding as the task proceeds. Reading involves making sense of the text, not simply taking meaning from the text, and involves the orchestration of prior knowledge about the ideas and about discourse conventions, traditions, and strategies.

Many early studies of reading and writing emphasized a linear, lock-step model involving mechanical application of skills. The dominant views of writing and reading at this time were knowledge-telling and text-driven models in which either writers converted recollections, mental models, and conceptions of ideas into print representation unaltered or readers decoded the message and took meaning from the print. Frequently, the writing and reading processes were devoid of any socio-cultural interactions, de-emphasized self-regulation, and emphasized the mechanics of the language. Clearly, our theoretical foundation and underlying assumptions do not embrace such views.

20.3 Writing Processes and Resulting Product Quality

20.3.1 Observation 1: Activities and their distribution over time define a process

Writing is seen as a problem-solving task in which several processes or cognitive activities are distinguished (Flower & Hayes, 1980; Hayes, 2005; Hayes & Flower, 1980). It includes cognitive activities, such as planning (what you want to say and how you want to do it), generating information (coming up with ideas), formulating (putting these ideas into words), structuring, revising, and so on. The writing process is recursive and not strictly sequential; each cognitive activity can follow any other, and each activity can occur at any moment during the writing process (Flower & Hayes). However, during writing, the task situation continually changes. As the text grows, writers have to make rerepresentations of the changing task situation. Writers adapt to these changes by carrying out different processes or cognitive activities.

Figure 20.2 (top) illustrates the mean occurrences of reading the assignment and generating ideas plotted against the moment (time) in the writing process. As time progresses and the text grows, the probability of occurrence for reading the assignment decreases. Average writers will engage less often in this activity near the end of the writing process than at the beginning (see also Sect. 20.8; first part of Eq. 20.5).

Generating ideas, however, shows a completely different pattern over the writing process. The mean probability of occurrence for generating ideas to write about

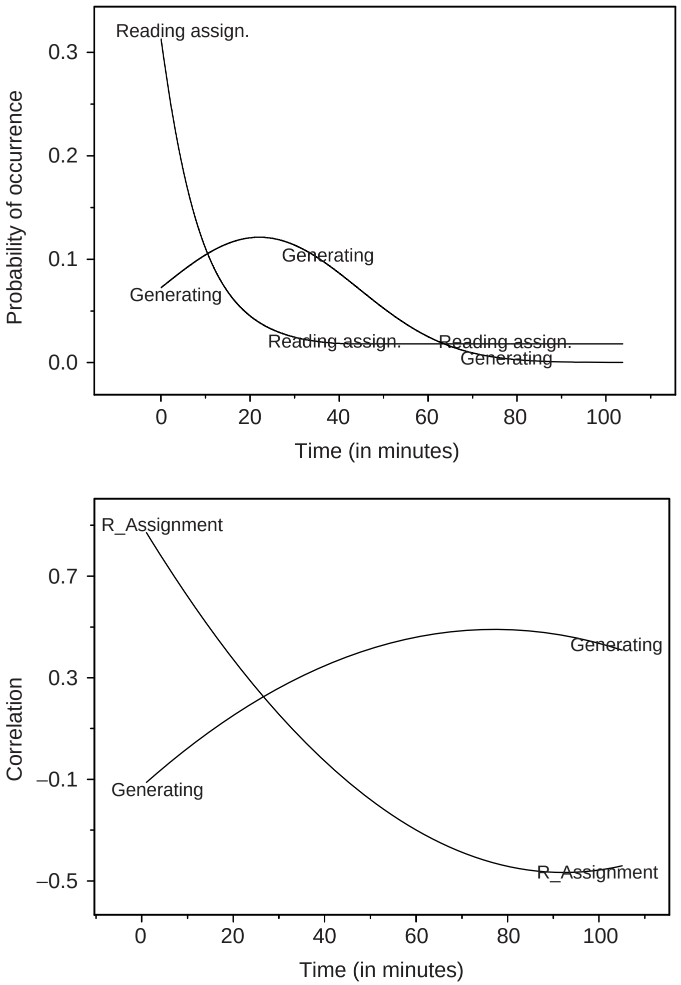


Fig. 20.2 Mean occurrence of reading the assignment and generating during the writing process (*top*) and correlations between the temporal distribution of reading the assignment and text quality (R_Assignment) and between generating and text quality (*bottom*) (Data from Breetvelt et al., 1996)

gradually increases from a low initial value to a peak (around minute 20), after which it decreases to a very low level. Hence, in the beginning of the writing process, writers refrain from generating ideas and engage in other activities (reading the assignment), while at later stages they focus on other activities (formulating, structuring, revising, etc.). These general patterns, observed with 15-year-olds, also appear with 11-year-olds (van der Hoeven, 1997) and first-year college students (van Weijen, van den Bergh, Rijlaarsdam, & Sanders, 2008c).

Most cognitive activities (planning, formulating, evaluating, revising, etc.) studied thus far show a distinct pattern of occurrence during the writing process. The point is that each cognitive activity has a higher or lower probability of occurrence depending on the moment. Different cognitive activities are dominant at different points; they are not absolutely sequential but rather recursive and situation-dependent.

20.3.2 Observation 2: Dynamically changing relations between processes and text quality

Figure 20.2 (bottom) illustrates the correlations between the occurrence of two cognitive activities (reading the assignment and generating) and text quality at various moments during the writing process. The most remarkable feature is that the correlation between process characteristics and text quality changes over time (see also Sect. 20.8: Eq. 20.9). These changes in the correlations show the importance of distinguishing the temporal occurrence of cognitive activities to the moment at which a writer carries them out and they also provide insights into the orchestration of cognitive activities during task execution. The orchestrations that are effective in producing good texts can thus be inferred. Reading the assignment is only positively related to text quality during the initial stages of the writing process, after which the correlation decreases and soon becomes negative. Conversely, writers who hardly consulted the assignment in the beginning wrote poorer texts. Writers who consulted the assignment frequently at the end wrote poor texts, whereas writers who refrained from reading the assignment at the end produced better texts.

The correlation between generating ideas and text quality also changes during the writing process. It increases during the initial phases, peaks in the middle, and subsequently decreases. Other things being equal, writers who refrained from generating in the beginning and the end, but did so in the middle, wrote better texts than writers who were preoccupied with generating information at the beginning or the end of the writing process.

We have seen that a description of each cognitive activity as a time-context process is worthwhile because the occurrence of these activities changes during task execution (Fig. 20.1, top). Furthermore, the relation between cognitive activities and text quality changes during the writing process. Although this may seem complicated already, it could still be an oversimplification of the writing process as cognitive activities can and do correlate with each other.

20.3.3 Observation 3: Individual differences in writing processes

We have presented general patterns of two cognitive activities. Figure 20.2 portrays the temporal distribution of two cognitive activities for the average writer; but to understand the writing process, we turn to individual writers, the cognitive activities they carry out, and the differences among writers and processes. The estimated individual patterns of occurrence for reading the assignment and generating activities are presented in Fig. 20.3 (see also Sect. 20.8; second part of Eq. 20.5).

For each writer (*N* = 36), the estimated probability of occurrence is plotted against time (the moment at which it occurs during the writing process). Each line (probability trace) represents one writer for either reading the assignment (top) or generating (bottom). The differences between writers are relatively large and

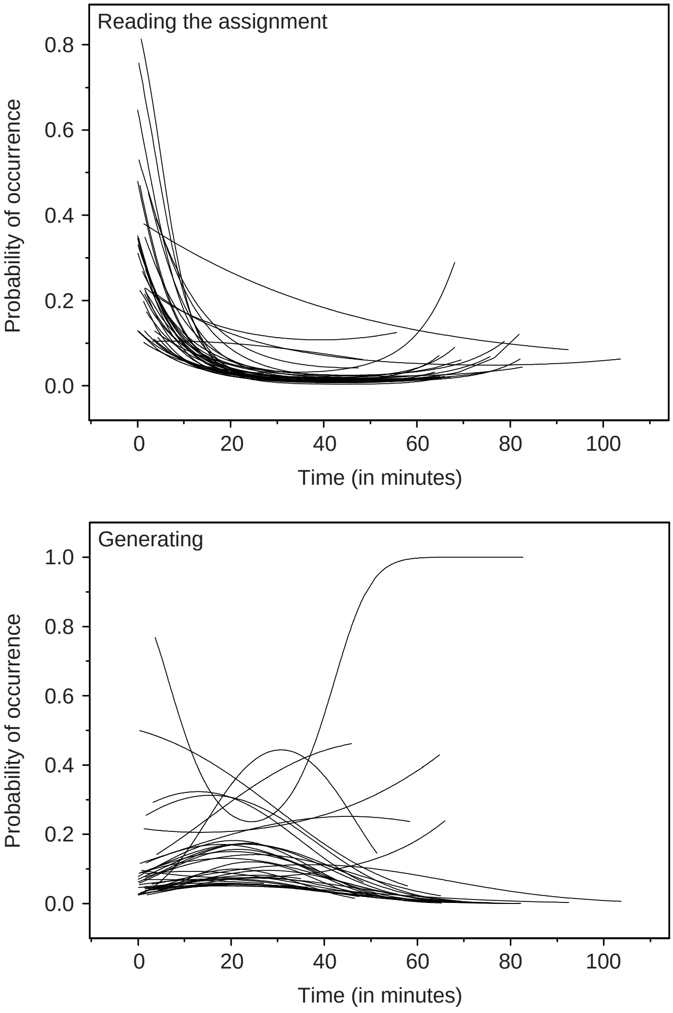


Fig. 20.3 Changes in the probability of occurrence during the writing process for individual writers: reading the assignment (*top*) and generating (*bottom*) (Data from Breetvelt et al., 1996)

depend on the moment in the writing process. In the beginning, the probability that a writer reads the assignment varies between 0.80 and 0.20. Some writers consult the assignment rather often during the beginning and considerably less during later phases. Other writers demonstrate an increase in the probability of reading the assignment at the end. A third group of writers demonstrates relatively constant probability for this cognitive activity throughout the writing process.

The temporal distribution for generating shows that most students increase their generating activities during the beginning of the writing process (first 20 min or so) followed by a decrease later (30–60 min). Most writers appear to follow the mean pattern although the differences in generating between writers are large. The clear exceptions are demonstrated by a student who generates often in the beginning and in the end but less in the middle stages and by a few students for whom the probability of generating continuously decreases.

The relation between both activities also differs among writers. For some writers, there is a positive relation between reading the assignment and generating, whereas for other writers the relation between these two activities is negative. In fact, the correlation between the temporal orders of both activities varies from −0.90 to 0.90 for individual students. Hence, some writers use the information in the assignment to generate new information to write about, whereas others do not need the information in the assignment to generate content information. Perhaps this difference in functional relations is mediated by prior topic knowledge; if one knows enough about a topic, one does not need to consult the assignment to come up with ideas. However, if one does not know what to write, it seems a plausible strategy to see if the assignment contains useful information.

We have observed and discussed two essential differences between writers in the temporal organization of cognitive activity. First, these differences are related to differences in text quality. Second, the differences in the temporal organization of one activity are related to the temporal organization of other activities. There appears to be a huge difference in the temporal organization of these two activities between writers. These differences might be related to their general procedural knowledge (Alamargot & Chanquoy, 2001; van der Hoeven, 1997) or be a consequence of the specific task execution, as both influence the ongoing writing process.

20.3.4 Observation 4: Functional compensatory relations between cognitive activities

It is assumed that cognitive activities can fulfill different functions depending on the context in the process. Hence, it is assumed that an activity occurring at moment t does not need to fulfill the same function in the writing process as that same activity occurring at moment $t + x$. Take, for instance, an activity like rereading already written text. This activity is part of the reviewing component in the Writing Process Model (Hayes & Flower, 1980). Hayes (1996) assigned reading a more central role in the writing process distinguishing several functions of reading during writing:

“In addition to reading to evaluate, two other kinds of reading play an important role in writing: reading source texts and reading to define tasks” (p. 18). However, it is tempting to attribute even more weight to reading during writing by hypothesizing that rereading already written text can, in some instances, fulfill a supportive function for generating ideas (cf. Galbraith, 1999). Take, for example, a sequence during which a writer writes something down, rereads what she wrote, and then generates another chunk of information. Such a sequence is not only logical but was also observed regularly (Breetvelt, van den Bergh, & Rijlaarsdam, 1996). In order to demonstrate the interdependency between these two cognitive activities, the correlations between rereading and generating and text quality were calculated (Fig. 20.2 bottom). The rereading-already-written-text correlation changes over time but is always positive. However, writers who generated ideas relatively often at the beginning wrote weak texts while writers who gradually increased the number of generating activities wrote the best texts. The point to be made here is that the correlation between either activity and text quality changes if the other activity is taken into account.

Figure 20.4 compares the general correlation between generating and text quality and the correlation between generating and text quality where rereading already written text is accounted for in the analysis. The figure clearly shows that the influence of generating on text quality is larger when we correct for rereading. That correction, however, has a different effect in both halves of the writing process—negative in the beginning and positive later. Generating activities in the beginning is ineffective; writers who generate much text in the beginning while relying on rereading already written text as input appear to have written a poor text. In the second half of the process, generating contributes strongly to text quality.

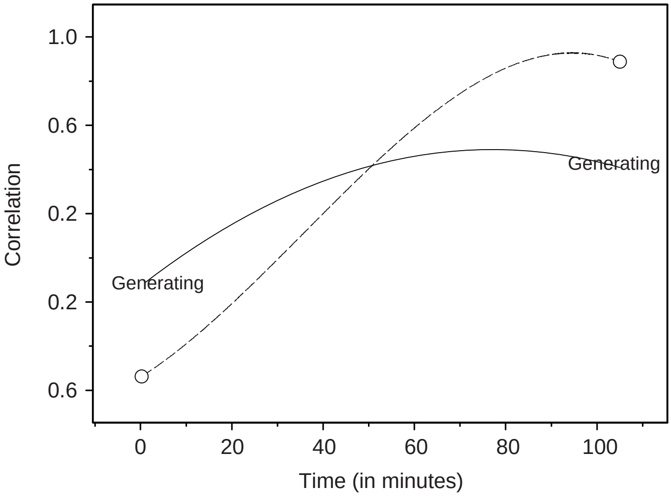


Fig. 20.4 Correlations between generating and text quality in two conditions: raw correlation (*solid line*) and corrected for rereading (*dashed line*) during the writing process (Reanalyzed data from Breetvelt et al., 1996)

However, writers who use rereading already written text in the second half as input profit more from this effect compared to writers who do not use rereading as input for writing. Therefore, the combination is inhibiting in the first half but facilitating in the second half; in this way, functional relations and their effectiveness appear to change over the duration of the writing process.

The correlation between rereading and text quality, not shown in Fig. 20.4, changes only marginally if generating is accounted for in the analysis (Breetvelt et al., 1996). This implies that generating does not have the same effect on rereading as rereading has on generating. Therefore, it seems plausible that rereading not only serves generating but that rereading has other functions as well. This does not only hold for the combination of rereading and generating but for many other relations between cognitive activities as well (cf. van den Bergh & Rijlaarsdam, 1999, 2001). It was shown, for instance, that the relation between translation-driven generation (i.e., generation preceded by an act of formulating) is only positively related to text quality in the beginning of the writing process. During later phases in the writing process, this correlation becomes negative.

This analysis shows that it is reasonable to map functional relations between cognitive activities. It also points to individual differences in the way activities are functionally combined. One possible interpretation of the presented relation between rereading and generating is that writers with weak generating skills need the input of the already written text as a knowledge resource for generating a new idea to write about. This interpretation could be tested if data about the writers' generation skills were available, which is not the case in the present studies (see van der Hoeven, 1997, for an analysis of the relation between revision activities and revision skill).

A second piece of evidence comes from another study (van Weijen, van den Bergh, Rijlaarsdam, & Sanders, 2005, 2008b) that focused on so-called writing process blocks: a protocol unit containing multiple occurrences of interrelated activities aimed at producing a certain amount of text, as

> the composing process has an episodic pattern of its own which is not dictated by the patterns of the text. Writers appear to work in composing 'episodes' or units of concentration which are organized around a goal or plan. Understanding the overall architecture of these episodes and the logic which begins and ends them will, we think, tell us a great deal about how writers combine planning and text production. (Flower & Hayes, 1981, p. 242)

Each individual writing process analyzed contained 2 to 14 blocks. We coded the presence or absence of planning behavior at the beginning of a block. Good writers appear to plan significantly more at the beginning. Planning within writing blocks does not differentiate between good and weak writers; nor is it related to the quality of the text produced. It was concluded that planning at the beginning appears to influence the processes within such a block.

We have tried to show that a univariate view on cognitive activities during the writing process limits the interpretation of the data and the building of a writing process theory. Such a view neglects the context in which cognitive activities occur or neglects the interdependency between activities, which can change during the process. The presented research results lead us to reconsider the unit of analysis for theory building. When combinations of cognitive activities behave as functional relations—implying that the function of each activity varies according to the context (i.e., the preceding activity or subsequent activities), then combinations rather than single activities might be considered a more appropriate unit of analysis.

20.4 Process and Task Characteristics

20.4.1 Observation 5: Differences in temporal distribution due to task

Thus far, we have not addressed differences in process characteristics across several tasks since the vast majority of studies involved writing only one or sometimes two texts. Within-writer differences in text quality are well documented (Coffman, 1966; Wesdorp, 1974), but within-writer variability in process characteristics is infrequently addressed (Rijlaarsdam & van den Bergh, 1996). In one study, writers (*N* = 20) were confronted with eight different writing tasks that all resembled each other, such as essays about use of cell phones in public transport, downloading of music from Internet, video surveillance, tolerance for soft drugs, etc. (van Weijen et al., 2008c). The results allowed us to infer differences in task execution within writers across several tasks. Figure 20.5 (top) illustrates the upper and lower boundaries of 80% confidence level for generating for these tasks. The pattern indicates that the occurrence of generating activities can vary due to the task with relatively small differences between tasks (within writers) in the beginning of the writing process, but the extent to which a writer varies how he carries out generating activities increases during task execution. It is tempting to assume from these probability limits that writers start more or less in the same way with all the writing assignments and that within-writer variance increases later. However, decreases and increases in within-writer-between-task variance cannot be interpreted at face value. They must be related to the other variance components (within-task variance and between-writer variance), which also depend on the moment. Therefore, the between-task variance has to be expressed as a proportion of the total variance and as a function of the moment in the writing process.

Figure 20.5 (bottom) illustrates the proportion between-task variance for generating expressed as a function of the moment in the writing process. The differences between tasks at the beginning and at the end are relatively large. The way writers start with a task varies greatly; for some tasks they start generating information much sooner than for other tasks. In the middle of the writing process, the approach to generating ideas in different tasks is much more alike; but in the end, the differences between tasks have increased enormously. Nevertheless, the correlation between generating and text quality remains remarkably stable (van Weijen et al., 2008c; Fig. 20.2, bottom). However, why the differences between tasks exist has yet to be determined. We do not know whether these differences in generating activities between tasks are related to differences in topic knowledge, a consequence of the occurrence of other cognitive activities, random differences, or other possibilities.

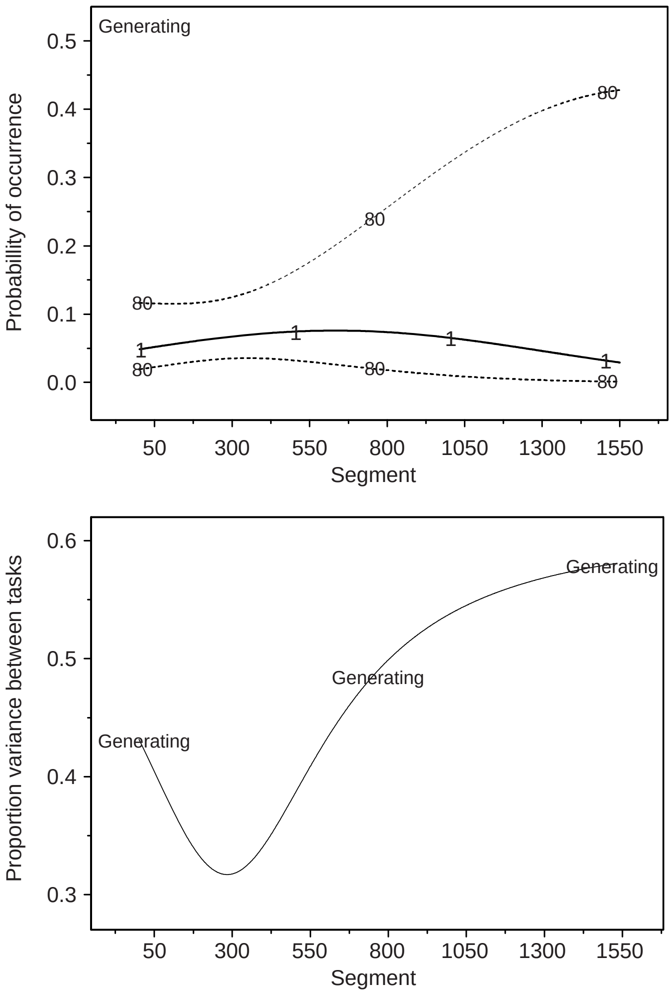


Fig. 20.5 Differences in temporal distribution of generating activities for different tasks: 80% confidence intervals (*top*) and proportion between-writer variance (*bottom*) (Data from van Weijen et al., 2008c)

20.4.2 Observation 6: Differences in activities due to the task situation (L1 versus L2)

An obvious distinction between writing tasks is the language in which the texts have to be written. In a series of studies, we focused on the influence of writing in a second language (L2) on the occurrence of cognitive activities during the writing process (Couzijn, van den Bergh, & Rijlaarsdam, 2002; Tillema-Kortman, van den Bergh, Rijlaarsdam, & Sanders, 2005, 2008; van Weijen et al., 2005, 2008a, 2008b, 2008c). All the studies had a comparable design but differed with respect to the number of tasks (four or eight) and the type of writers (Grade 9 or first-year university students). Writers wrote several argumentative essays in Dutch (L1) and English (L2). Assignments were counterbalanced (i.e., writer #1 wrote assignment A and B in L1 and C and D in L2, whereas writer #2 wrote assignment A and B in L2 and C and D in L1) and administered under think-aloud conditions.

The existence of a language effect can only be claimed if it adds to any differences found amongst the four tasks. Thus, it must be determined whether the temporal distribution of activities differs between tasks (Observation 5) *and* whether the distribution of activities differs between languages as well. Results indicate that the temporal distribution of each activity (i.e., the moment at which it occurs during the writing process) varies in both languages although differences between languages appear to be smaller than differences due to task. However, between-task variation plays a larger role in L1 than in L2; writers' behavior seems to be somewhat more stable between tasks when writing in L2. Finally, the correlation between each cognitive activity and text quality varies over time in both languages, depending on the moment at which each activity occurs and on the specific activity being carried out. For some activities (reading assignment), the correlation with text quality appears to be stable over languages. For other activities (generating, formulating, planning), the correlation with text quality clearly depends on the language.

van Weijen et al. (2005, 2008b) examined the effect of writing in different languages on planning behavior. In these studies, the unit of analysis was what we called writing process blocks (see Observation 4). Results show that good L1 writers do not only show adequate planning behavior at the beginning of blocks in L1 but are also likely to do the same in L2. Students who refrain from planning at the beginning of blocks in L1 are also likely to refrain from doing so in L2. Therefore, a clear influence of mother-tongue writing processes on foreign-language writing processes was claimed. Writers who have an effective strategy in L1 are more likely to also have a relatively effective strategy in L2. Likewise, writers who show an ineffective strategy in L1 are also likely to show the same ineffective writing strategy in L2. However, it has to be noted that the orchestration of cognitive activities while writing in a foreign language differs from writing in the mother tongue.

20.5 Process and Learner Characteristics

20.5.1 Observation 7: Differences between strong and weak readers

Our insights into literary reading interweave cognitive models (van Dijk & Kintsch, 1983) and results from empirical studies. A body of research provides evidence of differences between expert and novice readers of literature. Andringa (1995) distinguished between types of processes: identification (determining the literary genre), selection (focusing on information), (re)construction (filling in gaps), elaboration (making personal associations), evaluation (judging, criticizing), emotional processes (experiencing suspense, pity, etc.), and metacognition (reflecting on your reading activities). Rosenblatt (1938/1995) argued that “flexibility of mind” (p. 99) is part of the essence of reading and a fundamental goal of teaching reading. More mature readers are flexible in their use of reading strategies to achieve comprehension; that is, they are able to adapt their reading strategies to their reading purposes, to the nature of the reading material, and to the context (Pressley & Afflerbach, 1995; Pressley & Gaskins, 2006; Rayner & Pollatsek, 1994).

One interpretation of flexibility is that proficient readers adapt their reading process to the current task demands. In a study of 19 Grade 10 students, the orchestration of activities differed between good and poor readers (based on teacher assessments) who read four to five short stories under think-aloud conditions (Janssen, Braaksma, Rijlaarsdam, & van den Bergh, 2005). Figure 20.6 presents the differences in retelling (top) and emotional responding (bottom) over time measured as story fragments (each story was divided into 10 to 15 fragments following the original structure or paragraphs). The mean probability of occurrence for retelling and emotional responding are presented for the two ability groups (W for the 9 weak readers, S for the 10 strong readers).

The average strong and weak readers display different patterns of retelling and emotional responding. For strong readers, the probability of occurrence of retelling and emotional responding fluctuates during the reading process over the five stories. At the beginning of the reading process, the probability for these activities by strong readers is rather small (<5%). They refrain from retelling story content and from responding emotionally when reading the first few story fragments. The probability of occurrence gradually increases during reading but diminishes toward the end of the reading process, resulting in a curvilinear pattern. Weak readers demonstrate no such changes; their pattern of responses during reading for these activities and other reading activities, such as inferring (not addressed in Fig. 20.6), remains rather constant.

We hypothesize that strong readers are more flexible than weak readers; that is, they adjust their activities to the particular phase in the reading process, their understanding of the story, and the particular part of the story they are reading. This might not only hold for the reading of narrative literature but also for reading expository genres. Indeed, it has been shown that especially good readers profit from, and adapt to, specific characteristics of text structure, like logical connectives, signal words (Land, Sanders, & van den Bergh, 2008; Land, Sanders, Lentz, & van den Bergh, 2002; Mulder, 2007), and sentences (Kozijn, 2006), which also appear to depend on topic knowledge (Kamalski, 2007).

20.5.2 Observation 8: Individual differences within ability groups

Besides the average differences between ability groups just established (Fig. 20.6), there are large individual differences in process activities within ability groups (weak and strong). The variance within ability groups is related both to story and

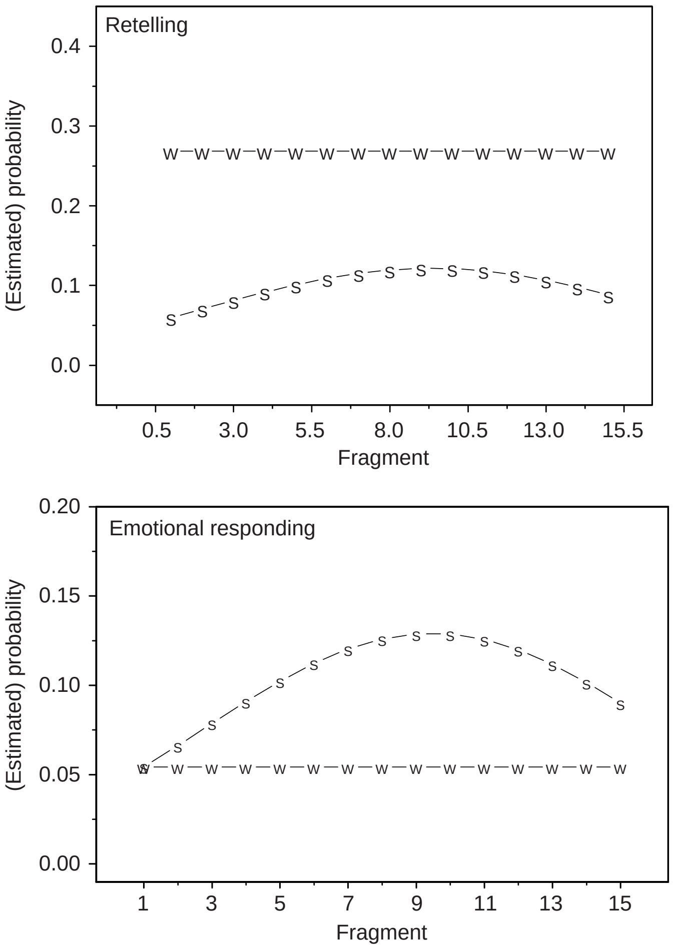


Fig. 20.6 Mean probability of occurrence of retelling and emotional responding for two ability groups (W: weak readers, *n* = 9; S: strong readers, *n* = 10 (Data from Janssen et al., 2005)

to reader characteristics. In general, the differences between readers appear larger for weak readers than for strong readers (Janssen et al., 2005). Averaged over different stories, good readers all show (more or less) the same patterns while the patterns for weak readers differ greatly. Nevertheless, all weak readers showed the same pattern during reading; they only differed in the amount a given cognitive activity is carried out.

20.6 Processes, Task, and Learner Characteristics

Insights into critical processes, strategic placement, adapting to task demands, and flexibility are central to better understanding writing in their first or second language and reading and interpreting literary texts. Whether good readers or writers differ from weak readers or writers in the ways they process tasks has become part of our research agenda in recent years.

van Weijen et al. (2005) found differences in the way that good and weak writers dealt with changing task demands (writing in L1 and in L2). The distinction between good and weak writers was made on the basis of text quality. Subjects' writing processes were analyzed to determine their planning behavior. It was expected that good writers would plan relatively more at the beginning of blocks than weak writers, and that good writers would be able to adapt to the demands of the task. Therefore, good writers were expected to show changes in planning behavior at block boundaries across tasks. However, this would only be true if it concerned intentional behavior.

Weak writers who wrote poor texts planned relatively little at the beginning of blocks (van Weijen et al., 2005, 2008b). Writers who did plan at the beginning wrote relatively better texts than weak writers. The best writers, however, were those with relatively high levels of planning at block boundaries but whose resulting approach varied between tasks (van Weijen et al., 2005). This could be a sign of adaptive behavior; because even though their approach to the task appeared to change to some extent, their text quality was consistently high. It has to be mentioned that we did not determine in what way the activities within writing blocks were affected by the presence or absence of planning at the block boundaries. Therefore, more research must be conducted in order to broaden and validate this observation to within text blocks and to other processes.

The results from this study were corroborated by a small-scale study of ten writers writing four essays (Tillema-Kortman et al., 2005, 2008). Writers who are not flexible in L1 are not flexible in L2; good writers, however, who showed signs of possible adaptive behavior in L1 also showed (relative) stable behavior in L2. Only the best writers in L2 showed some signs of adaptive behavior in L2. Hence, flexibility in one language does not guarantee flexibility in another language. Adaptation to the task in terms of processes seems to pay off in terms of text quality. However, this observation is based on a small-scale study with statistically significant differences; therefore, we plan a follow-up study to replicate the inquiry and verify these findings.

20.6.1 Observation 9: Strong readers respond differently to different stories, whereas weak readers tend to maintain the same pattern of response across different stories

Adolescents' responses to five different stories revealed that the between-story variance in reading activities was generally larger for strong readers than for weak readers (Janssen et al., 2005). Figure 20.7 illustrates the estimated probability of occurrence of retelling. Each line represents a specific reader reading a single story. In total, the figure contains 92 story traces (19 participants reading one of four to five stories). The story traces of the strong readers are relatively far apart with the probability of occurrence for retelling varying between 3% and 22%, depending on the story and the particular segment within a specific story. This indicates story and story-segment effects for strong readers. The story traces of the weak readers, on the other hand, are closer together and horizontal with the estimated probability of occurrence for retelling varying between 21% and 32%. There were no story or within-story effects found for the weak readers.

Similar results were found for other reading activities (inferring, problem detecting, associating); strong readers appeared to monitor and adjust their response to the story they were reading, whereas weak readers did not change their flatline pattern of processing much in response to different stories or segments within stories. This finding supports the hypothesis that good readers are more flexible and sensitive to text features they are reading than are weaker readers. The effects of story on students' reading processes underline the necessity of using several stories in think-aloud research instead of just one or two as in previous studies. Different stories elicit different responses, especially in strong readers. The question remains whether these differences are related to genre (narrative, expository, etc.) or to types of stories within a genre (fantasy, realistic stories, etc.).

All in all, we think we have provided some arguments in support of differences in process execution of strong and weak writers or readers. The former appear to adapt their processes much more to the task and the circumstances encountered than the latter.

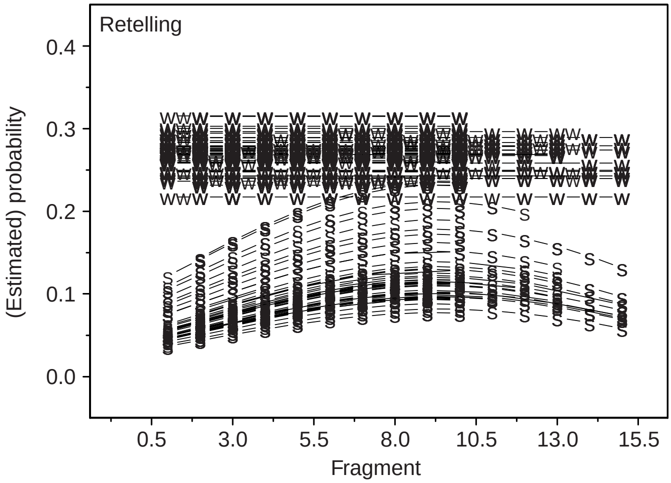


Fig. 20.7 Mean probability of occurrence of retelling for individual readers belonging to one of two ability groups (W: weak readers, $n = 9$; S: strong readers, $n = 10$) per story (Data from Janssen et al., 2005)

20.7 What We Want to Achieve in Educational Experiments: Processes as Output

When studying the effects of interventions, one must focus on the effects of interventions on differences in process execution for different learners. We have concentrated on the effects of observational learning on the orchestration of processes with the writing task. Braaksma, Rijlaarsdam, van den Bergh, and van Hout-Wolters (2004) examined the effects of observational learning on writing processes and the subsequent influence on writing products. An experiment was conducted in which participants (*N* = 52, Grade 8 students) learned to perform new writing tasks by observing peer-models' writing (experimental instruction) or by doing it themselves (control instruction). Two versions of the observational learning condition (focus on good models or focus on weak models) were implemented to increase our understanding of the generalizability of observational learning. The participants' orchestration of writing processes was measured by posttest writing tasks under think-aloud conditions.

20.7.1 Observation 10: Learning conditions influence the orchestration of processes

The study by Braaksma et al. (2004) showed that observational learning influenced the writing processes differently than learning-by-doing. Writers who learned by observing performed relatively more metacognitive activities (goal-orientation, analysis) at the start and relatively more execution activities (writing, rereading) in the second part than the writers who learned by doing. Over the whole writing process, writers who learned by observing showed more planning activities than writers who learned by doing (Fig. 20.8). In the middle and last part of the writing process, writers who learned by observing performed increasingly more meta-analyzing activities, indicating monitoring and regulating processes, than writers who learned by doing. Furthermore, writers who learned by observation showed a changing execution over time for some activities, whereas writers who learned by doing performed these activities at a constant rate during the writing process. In addition, variations in the observational learning conditions were larger than in the control condition, indicating more heterogeneous processes. Finally, it was found that students' orchestration of processes was related to text quality. Students who performed more goal-orientation and analyzing activities at the start of their writing process wrote better-quality texts.

20.7.2 Observation 11: Different interventions result in different reading processes

Janssen, Braaksma, and Couzijn (in press) studied the effects of self-questioning approaches to literature on Grade 10 students' (*N* = 67) processing activities, story appreciation, and quality of their postreading interpretations. The experimental

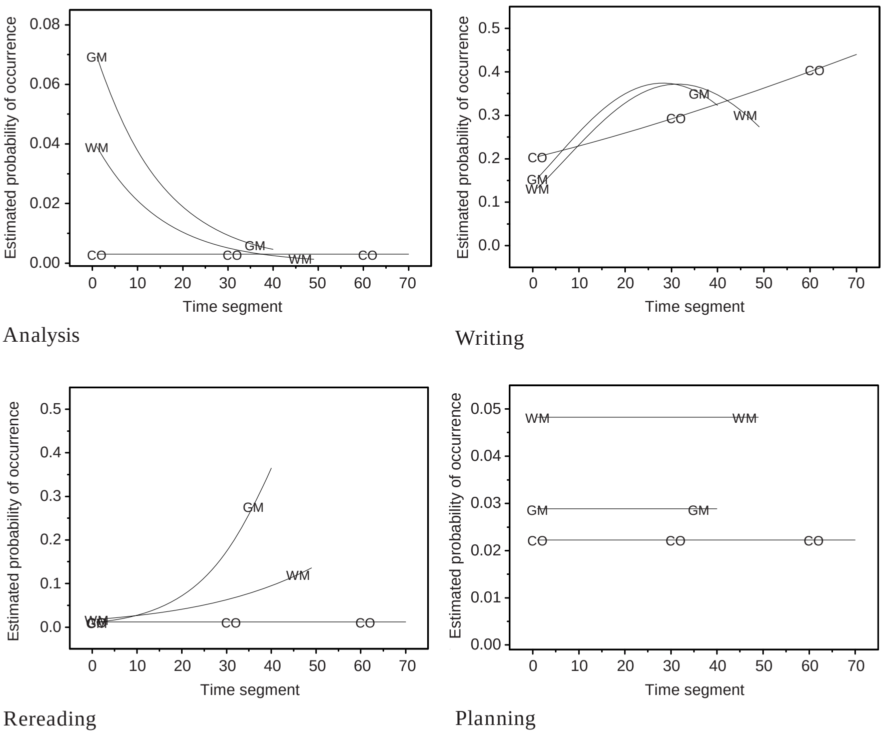


Fig. 20.8 Orchestration of analysis, writing, rereading, and planning per experimental and control conditions (CO: control condition, WM: weak model condition, GM: good model condition) (Data from Braaksma et al., 2004)

group learned to generate authentic, reader-based questions while reading short stories. The students exchanged and discussed their questions in small groups. The control condition resembled the experimental condition except students received questions generated by the instructor (based on questions posed by students in a pilot study) and discussed these questions in small groups. Pretreatment and posttreatment think-aloud protocols of students' *real-time* processing of short stories were collected. The protocols were analyzed to determine the occurrence of several reading activities.

The results indicated that the number of questions asked during reading increased significantly in both conditions. However, the growth in the experimental condition was larger than in the control condition. Furthermore, the experimental group pondered more often on their own questions during reading; students returned to their questions later in the reading process, formulated hypotheses, and searched for answers in the story or in their established knowledge resources. The control students made fewer attempts to find answers to their questions. This indicates that the experimental instruction influenced students' self-questioning and search processes during reading.

Other reading activities (retelling, associating, generalizing, evaluating) did not change significantly between the pretest and posttest. Nor did we find a significant change in the quality of students' postreading responses as measured by ratings of students' written reviews. However, the self-questioning approach did have a significant effect on students' appreciation of stories. Students who had learned to generate reader-based questions responded more positively toward the stories they read than the students in the control group.

20.7.3 Observation 12: Different interventions result in different reading processes for some learner characteristics

In a second quasi-experiment, the effects of self-questioning instruction were studied in 10 Grade 10 classrooms (*N* = 245 students, 9 experienced teachers, language of instruction was Dutch) in The Netherlands and Belgium (Janssen & Braaksma, 2007a, 2007b). Students learned to interpret complex literary stories by generating and discussing authentic, reader-based questions in response to the stories during six lessons as part of the regular literature curriculum. Students responded to two different stories during pretests and posttests. Again, a positive effect was found on students' self-questioning behavior during reading and on their story appreciation after reading, supporting the results of our previous experiment. This time, we also found an effect on the quality of students' interpretations using open-ended questions. However, only students who did not read fiction or literature in their spare time (nonreaders) appeared to profit by learning to generate reader-based questions in the literature classroom. Their story interpretations received significantly higher ratings on the posttest than on the pretest, whereas no significant changes in the quality of interpretations were found for students who were moderate or frequent readers of fiction and literature. Thus, students' extracurricular reading experience or proficiency appears to be an important factor.

20.8 How Can Statistics Help Us Answer Questions?

We have made several claims concerning changes in the occurrence of cognitive activities during writing and reading as well as their relation with the quality of the task executed. In order to substantiate these claims empirically, we need a statistical model that describes the observations and takes into account that these observations are nested within respondents; observations are not randomly distributed across respondents but define specific respondents' behavior observations. (The term *observation* is used here in the traditional scientific sense to denote the occurrence of cognitive activities, not as earlier to denote a tentative assertion.) Therefore, observations are not interchangeable between respondents. Multilevel models meet these requirements (Bryk & Raudenbush, 1992; Goldstein, 2003; Snijders & Boskers, 1999). In this section, we give a description of the models used and the data analysis.

First, each occurrence of a cognitive activity observed is treated as a single element in the analysis. Therefore, there are as many repeated observations of each individual as there are activities performed during task execution. These observations can be coded as 0 or 1; 1 for the target activity and 0 for all other activities. Note that the number of observations is allowed to vary between respondents; for instance, one respondent may need 10 cognitive activities to perform a certain task whereas another respondent needs 100 activities to perform the same task. Furthermore, we need to relate the occurrence of the target variable to the moment it occurred during task execution. We want to relate the 0s and 1s to the time elapsed or to another time-sequence factor (story segments, sentences produced, etc.) since the start of an assignment that allows us to map the time-related function of the target activities (cognitive activity as a function of time).

Let Y_{ij} represent the occurrence of the target variable of individual *j* at moment *i*, and t_{ij} represent the passing of time. Now the model to be analyzed can in principle be written as:

$$Logit(Y_{ij}) = \mathrm{f}(t_{ij}) \tag{20.1}$$

Equation 20.1 indicates that there is some relation between both variables (occurrence of a cognitive activity and time). As the response variable is dichotomous, a logit transformation is appropriate. Please note that, according to the equation, the same pattern is assumed for all individuals; the function is not allowed to differ between individuals. However, different individuals can carry out the same task in different ways; one writer may take only a short look at an assignment in the beginning but fall back on information in the assignment during later phases of the writing process; another writer might start by reading the assignment carefully but subsequently never look at it again. So, the model has to be extended to allow for individual differences in the distribution of an activity during task execution. We can easily allow the general function to differ between individuals (*j*):

$$Logit(Y_{ij}) = \mathrm{f}_j(t_{ij}) \tag{20.2}$$

Equation 20.2 specifies a function that describes the occurrence of a cognitive activity during task execution for different individuals. These two models are at the heart of our approach to the time-series analysis of process data.

In the past, many types of function have been proposed for models such as those presented by Eqs. 20.1 and 20.2 (van den Bergh & Rijlaarsdam, 1996; Breetvelt, van den Bergh, & Rijlaarsdam, 1994; Chatfield, 2004; Hoeksma & Koomen, 1991). However, we prefer polynomials because of their flexibility and ease of interpretation. That is, the occurrence of the target activity is modeled as a function of powers of time elapsed since the start of the writing or reading process. Depending on the number of coefficients (powers of time), polynomials can take almost any shape. Now we can rewrite Eq. 20.1 as a polynomial expanded for different powers of time, that is:

$$Logit(Y_{ij}) = \beta_0 * t_{ij}^{\ 0} + \beta_1 * t_{ij}^{\ 1} + \beta_2 * t_{ij}^{\ 2} + \ldots + \beta_k * t_{ij}^{\ k} \tag{20.3}$$

Equation 20.3 represents the mean change in the occurrence of an activity over time. The number of coefficients (*β*) required for an adequate representation of the observations can be seen as an empirical question; only significant coefficients are kept in the model, and higher-order coefficients are only included if all lower-order coefficients are significant (Goldstein, 1979). The parsimony restriction governs the inclusion and exclusion of terms in the polynomial in such a way that if none of the coefficients reach significance, we can restrict ourselves to an analysis of proportions.

We can view the temporal distribution of each individual as a deviation from the mean distribution or the general pattern during task execution. Therefore, we can write the coefficients of individual *j* as deviations from the mean coefficients:

$$\begin{aligned} \beta_{0j} &= \beta_0 + \mu_{0j} \\ \beta_{1j} &= \beta_1 + \mu_{1j} \\ &\ldots \\ \beta_{kj} &= \beta_k + \mu_{kj} \end{aligned} \tag{20.4}$$

Equation 20.4 has coefficients for each individual *j* noted as deviations from the mean. The residuals (μ_{0j}, …, μ_{kj}) indicate whether the regression coefficient for individual *j* deviates from the mean regression weight. It is assumed that these differences are normally distributed with an expected value of 0 and a variance of $S^2\mu_{0j}$, $S^2\mu_{1j}$, …, $S^2\mu_{kj}$.

If Eq. 20.4 is substituted in Eq. 20.3, the model to be analyzed can be written as follows:

$$\begin{aligned} &Logit(Y_{ij}) = \beta_0 * t_{ij}^{\ 0} + \beta_1 * t_{ij}^{\ 1} + \beta_2 * t_{ij}^{\ 2} + \ldots + \beta_k * t_{ij}^{\ k} + \\ &[\mu_{0j} * t_{ij}^{\ 0} + \mu_{1j} * t_{ij}^{\ 1} + \mu_{2j} * t_{ij}^{\ 2} + \ldots + \mu_{kj} * t_{ij}^{\ k}] \end{aligned} \tag{20.5}$$

The resulting multilevel model (Eq. 20.5) contains two parts: a fixed part and a random part (Bryk & Raudenbush, 1992; Goldstein, 2003; Quené & van den Bergh, 2004, in press; Snijders & Boskers, 1999). The fixed part of the model describes (the logit of) the average occurrence of a cognitive activity during task execution (Fig. 20.2, top). Please note that the fixed part of the model can be extended in order to incorporate individual characteristics (such as indicators for weak and strong readers, see Fig. 20.6) and interactions with the time variables. The random part (between the square brackets) describes the differences between individuals (as a deviation from the mean pattern, e.g., Figs. 20.3 and 20.7). Only significant coefficients are taken into account, and higher-order coefficients are only included in the model if lower-order coefficients are significant for both fixed and random parts of the model. Please note that the individual residuals are not estimated but their variances (e.g., $S^2\mu_{0j}$, $S^2\mu_{1j}$) and covariances (e.g., $S\mu_{0j,}\ \mu_{1j}$) are. However, these estimates can be used to approximate the residuals for all individuals. That is, for individual 1, the regression line would look like:

$$Logit(Y_{i1}) = [\beta_0 + \mu_{01}] * t_{i1}{}^0 + [\beta_1 + \mu_{11}] * t_{i1}{}^1 + [\beta_2 + \mu_{21}] * t_{i1}{}^2 + \\ \ldots + [\beta_k + \mu_{k1}] * t_{i1}{}^k \tag{20.6}$$

An individual's terms in the polynomial may be small or large depending on the size of the residuals (μ_{kj}) for that specific individual. The shape of the regression line for each individual depends on the residuals for that individual and hence on the variance between individuals.

One of the characteristics of the model in Eq. 20.5 is that differences between individuals are modeled in terms of variance. The estimated variance is time-dependent:

$$VAR(Individuals \mid T = t) = S^2{}_{\mu 0 j} + 2 * T * \text{cov}(\mu_{0j}, \mu_{1j}) + T^2 * S^2{}_{\mu 1 j} \\ + \ldots + 2 * T^k * \text{cov}(\mu_{0j}, \mu_{kj}) + \ldots + T^{2k} * S^2{}_{\mu kj} \tag{20.7}$$

It is important to note that the differences between individuals are allowed to change during task execution; the general assumption of homoscedasticity of variance, so explicit in many unilevel models, is not made in any way here (van den Bergh & Rijlaarsdam, 1996). The residuals or interindividual deviations from the average pattern are estimated in the random part of the model. Therefore, the random part of the model allows, as we have seen (Fig. 20.3), for the identification of individual regression lines.

Learner characteristics can easily be added to the model. A learner characteristic, say C_j, can be related to cognitive activities during task execution in different ways. For instance, if C_j would be added to the first line in Eq. 20.4—which in that case would read as: $\beta_{0j} = \beta_0 + \gamma_0 {}^* C_j + \mu_{0j}$—a main effect of this characteristic would be assumed. In essence, this means that the distribution for individuals with more or less of this characteristic has the same shape but only differ in their onset (intercept). The characteristic, however, can be related to linear, quadratic, or cubic changes during task execution as well. This implies that the distribution of an activity during task execution may vary with levels of the characteristic, and in fact a specific type of ATI model is specified.

In order to model relations between processes and product (i.e., between writing and reading processes and the quality of the final text or postreading responses), we need to expand the model to form a multivariate model. If we assume that the process characteristics are adequately described by a model such as Eq. 20.5, then we need a separate part of the model to describe differences in the quality of the product.

Let Y_{2j} be the quality of the product of individual *j*, then we can write the quality of the product of individual *j* as a deviation from the mean quality:

$$Y_{2i} = \gamma_0 + v_{0j} \tag{20.8}$$

As the process characteristics (see Eq. 20.5) and the product characteristics are estimated simultaneously, the residual scores (μs and v) are allowed to covary. In the case where only the linear random component of the process part of the

Table 20.1 Relations between process and product when only the linear component is significant

	Process		Text quality
Process	VAR (μ_{0j} * t_{ij}^{0})		
	COV ($\mu_0 j$, μ_{1j} * t_{ij}^{1})	VAR (μ_{1j} * t_{ij}^{1})	
Text quality	COV (μ_{0j}, v_{0j})	COV (μ_{1j} * t_{ij}^{1}, v_{0j})	VAR (v_{0j})

model is significant, the random part of the model would resemble the parameters presented in Table 20.1.

There are several points worth noting in Table 20.1. The first three components of the table describe the differences in process characteristics: the intercept variance (VAR (μ_{0j} * t_{ij}^{0})), the variance in the linear component (VAR (μ_{1j} * t_{ij}^{1})), and the covariance between the intercept and the linear component (COV (μ_0, μ_{1j} * t_{ij}^{1})). VAR (v_{0j}) indicates differences in product characteristics. The two remaining components describe the relation between process and product. One of these two coefficients relates to time (COV (μ_{1j} * t_{ij}^{1}, v_{0j}); therefore, the covariance between process and product may vary during process execution. Finally, as we have a model that describes both changes over time and between individuals, and we have estimated the relation between these differences and product characteristics, we can approximate the correlation at every moment in time:

$$(r \mid T = t) = \frac{\mathrm{cov}(\mu_{0j}, v_{0j}) + \mathrm{cov}(\mu_{1j}, v_{0j}) * T}{\sqrt{[\mathrm{var}(\mu_{0j}) + 2 * \mathrm{cov}(\mu_{0j}, \mu_{1j}) * T^1 + \mathrm{var}(\mu_{1j}) * T^2] * + \mathrm{var}(v_{0j})}} \qquad (20.9)$$

Equation 20.9 illustrates that the correlation between process characteristics and text quality are time (*T*)-dependent; the correlation fluctuates during the writing or reading process (see Fig. 20.2, bottom).

We have shown that there is a rather strict correspondence between the questions behind the observations and the models. Only two or three statistical models need to be understood in order to answer a wide range of questions related to differences in task execution, the relation with individual variables, and the quality of the resulting product. It is very tempting to interpret the results in a causal manner. However, statistics cannot answer the causality question. In order to answer questions related to the causality of process–product relations, more experimental studies must to be undertaken leading to better theories, cause–effect mechanisms, and insights.

20.9 Closing Remarks

We have attempted to show that it makes sense to study the occurrence of cognitive activities as a function of the moment at which they occur. It has been shown that the orchestration of cognitive activities is related to characteristics of text quality in the case of writing and related to how well students read short stories. Strictly speaking, we cannot infer causal relations from many of these descriptive studies and correlations. However, it was also shown that the orchestration of those activities, which is related to the writing of good texts, is at least partially responsible for differences between conditions in educational experiments. This seems to support the idea that at least some of the process–product relations are causal. Nevertheless, there are still many more questions to be studied before the presented observations will be linked in a coherent theory on writing and reading processes. One thing that stands out is that the moment processes are carried out must be the backbone of any such theory.

We have stressed the importance of going beyond the black-box model in process studies; we need to know how (sub)processes change during the writing and reading process, because these changes have been shown to differentiate between didactical approaches. There is an opportunity and need to open the traditional black box in many experimental studies; we would like to know which processes change due to experimental manipulations. Showing that a didactical principle works through educational research is one thing; showing *why* it works and *what effect* it has on the cognitive operations performed by students, is another—truly a Gold Standard.

We have tried to show that multilevel modeling is worthwhile. This type of modeling allows the analysis of process data while taking into account the hierarchical nature of the data, thereby allowing identification of interindividual differences in intra-individual change during task execution. These interindividual differences can be related to characteristics of output quality. Such relations are the foundation for new experimental and didactical research.

References

Alamargot, D., & Chanquoy, L. (2001). *Through the models of writing*. Dordrecht, The Netherlands: Kluwer.

Andringa, E. (1995). Strategieën bij het lezen van literatuur [Literary reading strategies]. *Spiegel, 13*(3), 7–33.

Bergh, H., van den, & Rijlaarsdam, G. (1996). The dynamics of composing: Modeling writing process data. In C. M. Levy & S. Ransdell (Eds.), *The science of writing* (pp. 207–232). Mahwah, NJ: Lawrence Erlbaum.

Bergh, H., van den, & Rijlaarsdam, G. (1999). The dynamics of idea generation during writing: An online study. In G. Rijlaarsdam (Series Ed.) & M. Torrance & D. Galbraith (Eds.), *Knowing what to write: Conceptual processes in text production* (Vol. 4 in Studies in Writing, pp. 139–160). Amsterdam: Amsterdam University Press.

Bergh, H., van den, & Rijlaarsdam, G. (2001). Changes in cognitive activities during the writing process and relationships with text quality. *Educational Psychology, 21*(4), 373–385.

Braaksma, M. A. H., Rijlaarsdam, G., Bergh, H., van den, & Hout-Wolters, B. H. A. M., van. (2004). Observational learning and its effects on the orchestration of writing processes. *Cognition and Instruction, 22*(1), 1–36.

Breetvelt, I., Bergh, H., van den, & Rijlaarsdam, G. (1994). Relations between writing processes and text quality: When and how? *Cognition and Instruction, 12*(2), 103–123.

Breetvelt, I., Bergh, H., van den, & Rijlaarsdam, G. (1996). Rereading and generating and their relation to text quality: An application of multilevel analysis on writing process data. In G. Rijlaarsdam, H. van den Bergh, & M. Couzijn (Eds.), *Theories, models and methodology in writing research* (pp. 10–21). Amsterdam: Amsterdam University Press.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Chatfield, C. (2004). *The analysis of time series: An introduction* (6th edn.). Boca Raton, FL: Chapman & Hall/CRC Press.

Coffman, W. E. (1966). On the validity of essay tests of achievement. *Journal of Educational Measurement, 3*(2), 151–156.

Couzijn, M., Bergh, H., van den, & Rijlaarsdam, G. (2002, July). *Writing processes and text quality: Effects of L1/L2*. Paper presented at the 8th international conference of the European Association for Research in Learning and Instruction, Writing Special Interest Group, Stafford, UK.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*(11), 671–684.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*(2), 116–127.

Dijk, T. A., van, & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Flower, L. S., & Hayes, J. R. (1980). The dynamics of composing: Making plans and juggling constraints. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing: An interdisciplinary approach* (pp. 31–55). Hillsdale, NJ: Lawrence Erlbaum.

Flower, L. S., & Hayes, J. R. (1981). The pregnant pause: An inquiry into the nature of planning. *Research in the Teaching of English, 15*(3), 229–243.

Galbraith, D. (1999). Writing as a knowledge-constituting process. In G. Rijlaarsdam (Series Ed.) & M. Torrance & D. Galbraith (Eds.), *Knowing what to write: Conceptual processes in text production* (Vol. 4 in *Studies in writing*, pp. 139–164). Amsterdam: Amsterdam University Press.

Goldstein, H. (1979). *The design and analysis of longitudinal studies*. London: Academic Press.

Goldstein, H. (2003). *Multilevel statistical models* (3rd edn.). London: Hodder Arnold.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1–27). Mahwah, NJ: Lawrence Erlbaum.

Hayes, J. R. (2005). New directions in writing theory. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 28–40). New York: Guilford.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing: An interdisciplinary approach* (pp. 3–30). Hillsdale, NJ: Lawrence Erlbaum.

Hoeksma, J. B., & Koomen, H. (1991). *The development of early mother-child interaction and attachment*. Unpublished doctoral dissertation, Free University, Amsterdam.

Hoeven, J., van der. (1997). *Children's composing: A study into the relationships between writing processes, text quality, and cognitive and linguistic skills* (Vol. 12 in *Utrecht studies in language and communication*). Amsterdam: Rodopi.

Janssen, T., & Braaksma, M. A. H. (2007a). Lezen in de diepte; Leren interpreteren van verhalen door vragen stellen [In-depth reading: Learning to interpret stories by generating questions]. In D. Schram & A. Raukema (Eds.), *Lezen in de lengte en lezen in de breedte: De doorgaande leeslijn in wetenschappelijk perspectief*. Amsterdam: Stichting Lezen.

Janssen, T., & Braaksma, M. A. H. (2007b). Literatuur leren lezen door vragen stellen; Effect op verhaalwaardering [Learning to interpret literature by asking yourself questions: Effect on story appreciation]. *Levende Talen Tijdschrift, 8*(3), 11–19.

Janssen, T., Braaksma, M. A. H., & Couzijn, M. (in press). Self-questioning in the literature classroom: Effects on students' interpretation and appreciation of short stories. *L1–Educational Studies in Language and Literature*.

Janssen, T., Braaksma, M. A. H., Rijlaarsdam, G., & Bergh, H., van den. (2005, August). *Flexibility in reading literary texts: Differences between weak and strong adolescent readers*. Paper presented at the 11th conference of the European Association for Research in Learning and Instruction, Nicosia, Cyprus.

Kamalski, J. (2007). *Coherence marking, comprehension and persuasion: On the processing and representation of discourse*. Unpublished doctoral dissertation, Utrecht University, The Netherlands.

Kozijn, R. (2006). *Integration and inference in understanding causal sentences*. Unpublished doctoral dissertation, Tilburg University, The Netherlands.

Land, J. F. H., Sanders, T. J. M., & Bergh, H., van den. (2008). Effective tekststructuur voor het vmbo Een corpus-analytisch en experimenteel onderzoek naar tekstbegrip en tekstwaardering van vmbo-leerlingen voor studieteksten [Effective text structure for Lower Vocational Training: A corpus-analytical and an experimental study on the effects of text structure on reading comprehension]. *Pedagogische Studiën, 85*(2), 76–94.

Land, J. F. H., Sanders, T. J. M., Lentz, L. R., & Bergh, H., van den. (2002). Coherentie en identificatie in studieboeken. Een empirisch onderzoek naar tekstbegrip en tekstwaardering op het vmbo [Coherence and identification in study books: An empirical study in the understanding and appreciation of texts in study books in lower vocational training]. *Tijdschrift voor Taalbeheersing, 24*(4), 281–302.

Mulder, G. (2007). *Understanding causal coherence relations*. Unpublished doctoral dissertation, Landelijke Onderzoekschool Taalwetenschap (LOT), Utrecht, The Netherlands.

Pressley, M., & Afflerbach, P. P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Lawrence Erlbaum.

Pressley, M., & Gaskins, I. W. (2006). Metacognitively competent reading comprehension is constructively responsive reading: How can such reading be developed in students? *Metacognition and Learning, 1*(1), 99–113.

Quené, H., & Bergh, H., van den. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication, 43*(1/2), 103–121.

Quené, H., & Bergh, H., van den. (in press). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language,* doi:10.1016/j.jml.2008.1002.1002.

Rayner, K., & Pollatsek, A. (1994). *The psychology of reading: An interdisciplinary approach*. New York: Lawrence Erlbaum.

Rijlaarsdam, G., & Bergh, H., van den. (1996). Essentials for writing process studies: Many questions and some answers. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 107–126). Mahwah, NJ: Lawrence Erlbaum.

Rosenblatt, L. M. (1938/1995). *Literature as exploration* (5th ed.). New York: Modern Language Association of America.

Snijders, T. A. B., & Boskers, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modelling*. London: Sage.

Tillema-Kortman, M., Bergh, H., van den, Rijlaarsdam, G., & Sanders, T. J. M. (2005, September). *Adaptivity in the handling of cognitive processes during L1 and FL writing and the role of metacognitive control*. Paper presented at the 9th biannual congress of the Writing Special Interest Group, Geneva, Switzerland.

Tillema-Kortman, M., Bergh, H., van den, Rijlaarsdam, G., & Sanders, T. J. M. (2008). *Adaptive strategies during writing in L1 and L2*. Manuscript submitted for publication.

Weijen, D., van, Bergh, H., van den, Rijlaarsdam, G., & Sanders, T. J. M. (2005). Adaptivity: Transferring writing processes between tasks (in L1) and between languages (from L1 to FL English). In C. P. Constantinou, D. Demetriou, A. Evagorou, M. Evagorou, A. Kofteros, M. Michael, C. Nicolaou, D. Papademetriou, & N. Papadouris (Eds.), *Abstracts of the 11th conference of the European Association for Research in Learning and Instruction: Multiple perspectives on effective learning environments* (pp. 1145–1146). Nicosia, Cyprus: Kailas.

Weijen, D., van, Bergh, H., van den, Rijlaarsdam, G., & Sanders, T. J. M. (2008a). *A comparison of process-product relations in L1 and L2 writing*. Manuscript submitted for publication.

Weijen, D., van, Bergh, H., van den, Rijlaarsdam, G., & Sanders, T. J. M. (2008b). *The role of composing episode in first and second language writing*. Manuscript submitted for publication.

Weijen, D., van, Bergh, H., van den, Rijlaarsdam, G., & Sanders, T. J. M. (2008c). *Variation in L1 writing: Cognitive activities and text quality*. Manuscript submitted for publication.

Wesdorp, H. (1974). *Het meten van producti et-schriftelijke taalvaardigheid* [Measurement of productive written language skills; doctoral dissertation]. Purmerend: Muusses.

Chapter 21
Can We Make a Silk Purse from a Sow's Ear?

Daniel J. Mundfrom

The Reverend Jonathan Swift (1801) is widely credited with coining the phrase "you can't make a silk purse out of a sow's ear" (p. 357), although Stephen Gosson appears to have made a similar statement centuries earlier in *Ephemerides of Phialo* in 1579: "seekinge … too make a silke purse of a Sowes eare" (Shapiro, 2006, pp. 619, #272). Regardless of origin, its general meaning implies that if something is not very good to begin with, you cannot do much of value with it. In the context here regarding statistical practices in educational research—and reshaping Swift's statement into a question: *Can we make a silk purse from a sow's ear?*—the implication is that research results that either come from poorly designed studies or use inappropriate techniques to analyze data, or both, have little hope of producing outcomes that will be effective in practice. Although this statement is applicable to research in virtually any context, the focus here is on educational research and its ability, or inability, to inform educational policy and practice in meaningful ways.

Educational research is not new. Educators, psychologists, evaluators, and other professionals have been studying the educational process for a century or more with the goal of improving the practice of education. The foremost international professional organization for promoting, studying, and disseminating research in education is the American Educational Research Association (AERA, n.d.); it was founded in 1916 with the goal of advancing educational research and promoting its application in practice. One would think that with all the educational research conducted year after year the quality of education seen in practice would be continually improving. We would expect to see ever-increasing levels of student performance, higher test scores, better teachers, exemplary schools, and all the by-products and effects of an excellent educational system in society at large. However, one need not look very far or in much depth to conclude that such is not the case—not in the United States and not in most, if not all, other countries around the world. Recent US government legislation underscores this fact. The No Child Left Behind Act of 2001 (NCLB, 2002) and the Education Sciences Reform Act of 2002 (ESRA, 2002)

D.J. Mundfrom
University of Northern Colorado

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

both speak to the need to generate better studies in education that can help to bridge the gap between educational research, policy, and practice.

So why does educational research not lead to better educational practice? With the results of years and years of study, why have we not learned how to provide better education on an ongoing basis? To be sure, to some extent we probably have (see Millar & Osborne, Chap. 3). To imply that there has not been any improvement to education as a result of research studies would simply be untrue. Undoubtedly, there is a myriad of educators around the globe who have discovered tenets or principles that, when applied in their own educational practice, have produced better educated, more learned students. But clearly, the advances have been neither widespread nor sustained throughout recent decades.

The other chapters in Part IV have described a number of statistical considerations and analyses believed to provide insights into education practices and research approaches. They have outlined approaches to make better use of large datasets with modeling and data mining, recognize the possibilities of confounding regression and prediction results with lurking variables, improving measurements and test items, and reanalysis of results from a program of study to produce more acute speculation and tentative relationships. Each of these chapters contributes to the better understanding of relationships within literacy and science education and provides the foundation for informing public policy about education.

One reason for the disconnect between research and practice could be the political nature of educational policy. More and more in the United States, educational policy is formulated in state houses and governor's offices and, as can be seen from the passage of the aforementioned legislation, is alive and growing on Capitol Hill and in the White House. This is not to say that educational policy was not formed and influenced by the federal government prior to the passage of this legislation in 2001 and 2002; for example, Bill Clinton was touted as the *education governor* during his presidential campaigns—with the clear implication that he could bring about reform and improvement in schools. The US Department of Education has existed for years with a cabinet-level director. There have certainly been political efforts to improve education. Many, if not most or all, states now have some form of mandatory, statewide, accountability program that regularly tests students; and they use these data to rate districts and schools—and sometimes individual teachers—on their ability to educate their students. Reform and improvement in education in the 21st century is clearly a political issue. There is no inherent reason why educational research cannot inform this political process and be used to help set the policies that could lead to better educational practice. But for whatever reasons that exist, there does not appear to be evidence to indicate that political attempts to improve education have had any far-reaching, enduring effect on student performance. We do a lot of testing, and we collect a lot of data; but standardized measures of student performance typically do not show consistent widespread gains as a result.

Another reason that educational research has had little sustained impact on student performance could be due to the way in which educational research is conducted. Outside of education and the other behavioral and social sciences, standard research practices involve the use of randomized controlled trials (RCTs) that compare outcomes of individuals who have been treated differently. Most such designs are based on Fisher's Randomization Test (Fisher, 1966) and compare variability among units treated differently (systematic variation) with variability among units treated alike (random variation). Within education, however, research traditionally has involved less rigorous designs. It is more difficult and sometimes not possible to randomly assign individuals to different groups and randomly assign different treatments or interventions to those groups. In recent years, more and more educational research has employed alternative qualitative methodologies that rely less and less on scientific methods and statistical analysis to reach conclusions about how to better inform educational practice. While qualitative methodologies are not inherently inferior to other research techniques, their increased use in education has led to more context-specific studies that do not rely on the basic experimental research tenets that have led to advances in medicine, agriculture, manufacturing, and the natural sciences.

Recent US federal legislation appears to be aimed at swinging the pendulum back in the direction of the collection of rigorous evidence from RCTs (Shelley, 2005). This legislation established the Gold Standard for educational research that emphasizes evidence-based interventions and outcomes that have been shown effective in RCTs and requires studies to establish strong evidence of effectiveness resulting from well-designed, carefully implemented studies. But can the passage of educational reform bills by state or federal governments bring about the needed change? Will the existence of this new legislation improve the design of educational research studies? And even with improved research studies, will the politicized process embrace their findings so that they find their way into educational policy? And even if they do inform policy, will the desired results be realized in practice?

Slavin (2008) examined practices in educational program evaluation and concluded:

> Evidence-based reform has the potential to substantially change the practice of education and to make educational research far more central to education policy. Practitioner-friendly syntheses of research on practical programs play an essential role in establishing the idea that there is evidence worth paying attention to. (p. 13)

Only time will tell if Slavin is right. Rather than attempting to guess how the answers to these questions may be realized, this chapter examines some of the existing practices in educational research to help identify what must change in the design and analysis stages of educational research for there to be any realistic hope that this Gold Standard for educational research can bring about the changes its crafters seek. In other words, if we cannot make a silk purse from a sow's ear, can we find a more appropriate way to make the silk purse?

21.1 Experimental–Observational Units

One of the first lessons learned in research methods is the importance of understanding experimental units and the relationship between them and the design and analysis of the corresponding study. Drawing on the basics of the Randomization

Test (Fisher, 1966), the experimental unit is fundamental to the comparison made between systematic variation and random variation. The experimental unit is often defined as the smallest entity to which a treatment is applied. If a fertilizer or insecticide treatment is applied to a field as a whole, then the field as a whole and not an individual plant is the experimental unit. Similarly, several potted plants on a table in a greenhouse that are all setting under the same grow lamp cannot be considered as separate experimental units—since the treatment (the grow lamp) is applied to them collectively as a group rather than to each pot individually. This concept seems reasonable and understandable in these scenarios.

However, when the situation changes from plants in a field or pots in a greenhouse to students in a classroom and a specific teaching method/technique/strategy is used with the class as a whole, it seems more difficult for researchers to see that each student cannot be considered as a separate experimental unit. Tests are given to each student, and data are available for each student individually—unlike the plants in a field scenario where individual plant data often may not be available. But the situations are the same because the determining characteristic is: To what entity is the treatment applied? To be sure, there are instances in which an intervention is applied to each student separately; in such cases, the student can appropriately be considered as the experimental unit. Such studies, although present in the educational literature, are not the norm and do not comprise the majority of research studies in education.

Typically, a teaching method, specific curriculum, or educational intervention is studied by using it with one or more classes of students and comparing the results in those classes with results from similar classes in which that method, curriculum, or intervention was not used. That is, systematic variation (variation among experimental units treated differently) is measured by comparing the outcomes in the classes that received the treatment with the outcomes in the classes that did not. Random variation is measured by examining differences among the classes treated alike (among those that received the treatment and among those that did not). Differences among individual students within the classes—both those that received the treatment and those that did not—are no more relevant to the analysis of the data from an educational study than are differences among the plants in the field receiving the same fertilizer treatment or the pots on the table under the same grow lamp in the previous examples. The fundamental identification of the appropriate experimental unit cannot be overemphasized. How the study is designed and implemented is crucial to this process; the basic experimental unit must also be the basic unit of analysis. If the treatment is applied to the class as a whole, then the class as a whole must be used as the unit of analysis. Only when the treatment is applied to students individually can the student be used as the unit of analysis. Yet, Slavin (2008) asserted that "many researchers assign schools or classrooms randomly to treatment and control groups but then analyze at the student level" (p. 9).

This principle has important implications for the design and analysis of educational research. Suppose, for example, that a study was designed in which one class of 30 students was taught with Method A and another class of 30 students was taught with Method B. An inappropriate but not uncommon approach to the analysis of data from this design would consider each student as a separate experimental unit; thus, the analysis would consist of 60 experimental units with 59 total degrees of freedom. One degree of freedom is used to compare the two treatments, leaving 58 degrees of freedom for error. Such an analysis would appear to have at least moderate power and finding a significant treatment difference could be likely. But the treatments were not applied to students individually. A class of 30 students is taught en masse so the experimental unit is the class and not the student, resulting in only two experimental units in this study. With only two experimental units, there is only one degree of freedom total (i.e., $n - 1$) and that degree of freedom is used to compare the treatments, leaving no degrees of freedom with which to estimate random error. No inferential test of significance is possible in this study.

Such a study could be improved by adding more classrooms of students that are taught with each method respectively. If two classes are taught with each of Methods A and B, then there are four total experimental units and three total degrees of freedom. Using one degree of freedom to compare the treatments, there are now two degrees of freedom left to estimate random error. Consequently, a statistical test can be conducted; but with such a small number of degrees of freedom, statistical power would be quite low and only substantial treatment effects would be detectable. With data available on 120 students, 60 taught with each method, to some it may seem unfair that such a study would have so little chance of producing useful results. It may be at least partly because of such reasoning that the need exists for reform in how educational research is conducted. Better training and more informed practice would be useful for enhancing the chances of research being designed, conducted, and analyzed with the experimental unit appropriately identified and utilized in the study.

This same basic relationship exists in observational studies, which are also common in educational research. An observational study—one in which the researcher does not create differences between or among the groups by manipulating treatments but rather obtains data by observing one or more outcomes on individuals previously treated with different treatments—does not eliminate the importance of identifying the appropriate unit of analysis. If the treatments still consist of different methods, curricula, or interventions, then those treatments would have been applied to the observational units at some time. The fact that the researcher is not the person administering the treatments does not change the basic relationship that drives the analysis. In this case, the observational units would still be the classrooms of students that received the instruction, etc., as a whole; so the classroom, not the student, is the appropriate unit of analysis.

Finally, this problem is not new; nor is it a recent discovery (Barcikowski, 1981; Blair, Higgins, Topping, & Mortimer, 1983; Donner & Klar, 2000; Levin & Serlin, 1993; Lindquist, 1940; Murray, 1998; Peckham, Glass, & Hopkins, 1969). Yet, still today, it is possible to find such inappropriate analyses in the published literature; and some textbooks still provide examples with the individual student in intact classrooms used as the unit of analysis.

21.2 Multiple Univariate Tests versus a Single Multivariate Test

Sound statistical practice has long recognized the problems inherent in performing a series of tests on a single sample of data. For example, when comparing the means of, say, five groups, a novice researcher may consider performing separate *t*-tests on pairs of the five means taken two at a time. To make all possible comparisons, 10 such *t*-tests would be required. Informed researchers have long known and accepted that a series of multiple *t*-tests, such as described above, is inappropriate and that a single ANOVA comparing the five means is the appropriate analysis in such a situation. Significant effects from the ANOVA justify subsequent comparisons of the means to identify the sources of significance.

It is apparently much less well recognized that multiple univariate tests used in place of a single multivariate test also have problems inherent with their use. With the continued advancements in computing power and the ease with which complex multivariate analyses can be conducted, the importance of identifying and conducting the appropriate analyses cannot be emphasized too strongly. Multivariate datasets are becoming commonplace—not that they did not exist in years past as well. Collecting, storing, and analyzing data has never been easier than it is today. Consequently, larger and larger datasets, with observations–measurements on more and more variables are prevalent (see Anderson, Milford, & Ross, Chap. 13; Meyer, Chap. 15). The need for good practice regarding the appropriate analysis of multivariate data has never been greater. One MANOVA. Several ANOVAs. Does it really matter which choice is made?

It is not hard to see that the hypotheses tested in a single MANOVA are not the same hypotheses tested by conducting several univariate ANOVAs. Comparing the means of several response variables across several groups one at a time is not the same as comparing a vector of means on these same response variables as a whole across the groups. The research questions addressed in these two scenarios are different. So the decision regarding which analytical path to follow is not simply personal preference or ease of analysis—as it could be if we were discussing alternative ways to address the same question. But the fact that the research questions one wishes to address are better suited to several ANOVAs than they are to a single MANOVA does not provide sufficient justification for choosing several ANOVAs for the analysis.

Two similar-sounding research questions and related hypotheses require distinctly different analyses. The problem space has to do with comparing student achievement in five middle schools within a single school district. Each school serves a similar number of students (200 students at each grade level) from similar socioeconomic backgrounds, and the schools had similar achievement histories over the last decade. Year 1 of the statewide assessment program focused on science knowledge; Year 2 focused on general achievement across the curriculum in mathematics, reading, and science. The research questions in both Years 1 and 2 are: Is student achievement in Grade 8 consistent across the five middle schools in the school district? When you have one variable in Year 1 (e.g., a standardized science test score) to compare across several groups (e.g., a sample of eighth graders from each of five schools), an ANOVA is the appropriate statistical procedure. It compares the null hypothesis that the mean science score is the same in all five groups to the alternate hypothesis that at least one of the groups has an average science score that is different from the others. Now in Year 2 when you have three variables (e.g., standardized test scores in mathematics, reading, and science) to compare across these same five groups, using separate ANOVAs for each variable implicitly assumes that each variable is measured in a different sample of individuals. Therefore, you would have three different sets of hypotheses, each one comparing a null hypothesis that the five group means are equal compared to its alternate hypothesis that at least one of the groups has a different mean. Most often, it is not the case that each variable is measured on separate samples of individuals from the five schools being compared; but, in fact, all three variables were measured on the same samples of individuals from the five schools being compared. In this scenario, the null hypothesis is that the vector of three means (one mean for each variable) is the same in all five groups. Therefore, the analysis must consider one pair of hypotheses in regard to the three variables simultaneously, which means a MANOVA would be the appropriate approach. It does not appear to be the case, at least not to this author, that the rationale indicating the appropriateness of the multivariate analysis is a difficult one to fathom.

Multivariate data results when multiple variables are collected on the same sample of individuals. When two or more of those variables are to be considered as the outcome/response variables in an analysis, then the multivariate nature of the data requires a multivariate approach to the analysis. Data collected on multiple variables from the same sample of individuals are inherently related. The relationship among these variables as a consequence of the research design that measured/observed the various characteristics on the same individuals is not necessarily a linear relationship as measured by a correlation coefficient or a consequence of the variables being conceptually related. Consequently, these data have a multivariate nature that can only be accounted for with a multivariate analysis.

If these outcome variables are to be compared across two or more groups, then the *only* appropriate analysis is a MANOVA that incorporates the relationship among these variables into the analysis. The fact that the MANOVA hypotheses may not provide answers to the desired research questions is an unfortunate consequence of poor planning in the original design. Using inappropriate analysis to salvage something from a poorly designed study is no more justified in this multivariate setting than it would be in simpler, more commonly occurring, univariate situations. Testing the vector of means on the response variables across several groups will not always indicate which variables differ individually across those groups. However, if separate analyses of each response variable individually are what is desired, then separate samples of individuals (i.e., one for each response variable) should have been part of the study design.

The analytical strategy presented here for dealing with multivariate data is somewhat different from the view expressed by Huberty and Morris (1989) where a rationale was provided that supports the use of multiple univariate ANOVAs in certain situations. The view expressed here is consistent, however, with that expressed by Thompson (1999) where the use of multiple univariate tests in lieu of the appropriate multivariate test is presented as poor statistical practice. It could well be the case that the inappropriate use of multiple univariate tests with multivariate data reflects the user's lack of understanding of complex data structures and the dependence that exists among variables measured/observed on the same subjects.

21.3 Random Assignment

Another major tenet of comparison-group studies is the notion that the groups are similar in structure and characteristics before some treatment is applied. Whether it is an experimental study in which the researcher creates group differences by treating the groups differently or an observational study in which the researcher enters the picture after the differences are already there, meaningful comparisons to identify differences across the groups that can be attributed to the treatment are only possible if there is some reasonable sense that the groups were not substantially different before any treatments were implemented. No reasonable person would conclude that differences across groups at the end of a study could be attributed to different treatment of the groups during the study if the groups were substantially different from each other before the study began.

In experimental studies, the commonly accepted method of ensuring similar groups prior to treatment is the use of random assignment to create the groups for study. Simply put, random assignment allocates experimental units to different groups on a random basis so that any preexisting differences among these experimental units will be distributed across the groups; and, on average, the groups should be essentially equivalent at the start of the treatment. Random assignment provides a level of control of extraneous variables, that is, variables that are not being studied as a part of the research design but ones that may have some influence on the results of the study particularly if allowed to differ across groups. By assigning experimental units to groups on a random basis for reasonably large samples, it is highly unlikely that all of the experimental units similar to each other prior to treatment implementation would be assigned to the same group. Consequently, whatever effect these similarities across the units may have on the response would tend to average out across the groups and be negated as a result.

In educational research, however, the use of random assignment to create groups is not commonplace and, in fact, might be rather rare. Research conducted in schools often takes place in classrooms that consist of specific students who are placed in those classrooms for specific reasons. In some cases, students may choose the course and section in which they enroll. In other cases, students are assigned to classes at the request of teachers, parents, or administrators for a variety of reasons. In none of these instances can it be said that students were randomly assigned to these classes. If these same intact classes are used in a research study, the basic tenet of substantial similarity prior to treatment is suspect at least, if not unwarranted altogether.

Often, research designs that do not employ random assignment to create groups are called quasi-experimental studies. These designs are susceptible to invalid conclusions because the lack of random assignment does not allow for the same level of control over extraneous variables. Most educational and behavioral science research textbooks devote many pages to experimental research designs and how to control extraneous variables. It is certainly true that research designs can be improved with appropriate attention being paid to the effects that uncontrolled extraneous variables can have on the results. However, if more attention were devoted to the design from the beginning and in particular to the appropriate use of random assignment, it is plausible that less attention would need to be paid to other ways of dealing with extraneous variables and the chances of obtaining meaningful, useful results would be enhanced. Meeting randomization is further complicated under research ethics standards requiring informed consent and voluntary participation (see Anthony et al., Chap. 24).

21.4 Confounding Variables

In the language of research and statistics, confounding refers to situations in which the effect of two or more variables on one or more other variables is so intrinsically entwined so as to be inseparable. Confounding can occur in many contexts (see Wang, Dziuban, Cook, & Moskal, Chap. 19; Meyer, Chap. 15). The extraneous variables discussed in the previous section can be examples of variables that are confounded. Suppose a study is conducted to compare two methods of instruction, A and B. Method A is used in three classes all taught by Teacher I; and Method B is also used in three classes, all taught by Teacher II. In such a scenario, the instructional method and the teacher are confounded so that it is not possible to determine if the performance of the students is a result of the method or of the teacher.

Confounding of variables can have substantial effects on the results of the analysis. Schield (2005) pointed out that in observational studies a confounding variable can make a statistically significant relationship appear to be nonsignificant or to make a statistically nonsignificant relationship appear to be significant. Neither outcome is a good one; care needs to be employed to safeguard against them. Frequently, at least, confounding variables are allowed to influence results because the design does not appropriately account for the multivariate nature of the data being studied. As discussed previously, multivariate data require multivariate analyses to appropriately account for that structure in the analysis. Schield asserted that "failing to teach [the effect of confounding variables] to students dealing with observational data is professional negligence" (p. 2). Schield's comments should not be taken to imply that all observational studies are flawed and that such studies have no place in research—educational or otherwise. There is value to be had from any well-designed and appropriately conducted research. However, whatever the design, the impact that confounding variables may have on the results is important to be taken into account and should not be ignored or underemphasized.

21.5 Overreliance on Software

It has been previously noted that the substantial advances in computing speed and power has had an undeniable influence on the practice of statistics; as well, improvements in and enhanced availability of software to conduct statistical analyses has had a major influence on research practice (see Anderson et al., Chap. 13; Nolan & Temple Lang, Chap. 18; Wang et al., Chap. 19; van den Bergh et al., Chap. 20). These changes are evident in virtually every academic discipline, and education is no exception. The increased ability to perform fast and easy analyses is certainly a positive occurrence in regard to educational research.

But with ease of use also comes the possibility of inappropriate use. Many statistical software packages, such as SPSS, now come with pull-down menus that aid the user in setting up a statistical analysis or specifying a statistical model for the data. In fact, it is possible for individuals who know virtually nothing about statistics or appropriate analytical choices to perform complete statistical analyses that range from simple descriptive summaries to more complex inferential tests. Such a user is likely to conclude that because the software generated the results, the results must be correct—regardless of whether or not the choices were appropriate or if the user even correctly identified data characteristics (nominal, ordinal, interval, ratio) in the analysis. It is easy to code categorical data numerically and use the numerical codes as actual numerical values in a regression analysis. The resulting statistical summaries and inferences, of course, would be meaningless.

Even those software packages that do not have the easy-to-use, pull-down menus can still be used inappropriately by the uninformed user. For example, a programming-oriented package like SAS has default options that automatically kick in if no specific option is specified by the user. While these default options allow the software to run the analyses and generate statistical results, those default options may not be the most appropriate choices for any particular dataset; consequently, the results may not be as clear or as applicable as they could be.

Finally, even software packages that are not statistical in nature, such as Microsoft Excel, often contain the ability for the user to generate statistical results. Extreme care should be exercised in using nonstatistical software to perform statistical analyses. Callaert (2000) identified specific aspects of certain nonstatistical software packages that either produced inaccurate calculations (e.g., a negative R^2 value) or presented material on statistical concepts that was misleading at best and out-and-out incorrect at worst.

It can also be noted that some research journals or journal editors may require that specific analyses be performed before a manuscript can be accepted for publication (e.g., effect sizes). While it may be the case that such requirements, in general, are well intended, it is possible, at least, that such analyses may not be optimal or even appropriate for some particular situation. Caution should be exercised in making such requirements as analyses should always be driven by the design of the study and what best facilitates an answer to the research question.

21.6 Valid Use of Test Scores

Much could be said about the reliability and validity of test scores and that these characteristics are properties of the scores obtained by using instruments and not of the instruments themselves (see Froelich, Chap. 14). Virtually any introductory textbook on educational research or measurement contains ample discussions of these topics. Most recent measurement and research texts accurately define validity as the appropriate use of the scores obtained from the administration of an instrument. Yet the knowledge of these concepts notwithstanding, it is not uncommon to see examples of the scores obtained from instruments (standardized or otherwise) used inappropriately.

It has become standard practice for state or provincial educational agencies to regularly test students regarding their levels of performance in core subject areas (e.g., reading, mathematics, and science, to name a few). Such testing is not new, and several testing companies market specific tests for this purpose. What is new on this front is the use of such test scores to try to hold districts, schools, and individual teachers accountable for the performance of their students. Having nothing against the concept of accountability, it is still incumbent upon the education profession and educational professionals to ensure that such tests' scores are not put to invalid use.

Consider a scenario in which students' performance levels are assessed by the use of a standardized test. If the score is used to infer how well the student is performing, either individually or as compared to other students, then it is likely that such an inference constitutes a valid use of that score. On the other hand, if that score is used to infer how well the teacher teaches, how good the teachers are in a particular school, or the quality of education offered by a particular school district, it is likely that such an inference is an invalid use. Tests designed to measure student performance should not be used to make inferences regarding teacher ability–performance. As educators, we should be appropriately concerned with the ability and capability of teachers. But if we want to measure those characteristics, it should be done with instruments designed to measure teacher ability.

21.7 Closing Remarks

Educational research has been conducted for decades; it will continue to be conducted for decades more. State and federal governments hold a stake and have assumed a role in ensuring the quality of education available to our students. Recent US federal legislation is designed to improve the quality of educational research and attempts to accomplish this outcome by raising the standards of research design. Sound research design is the first step to improved educational research. We must teach and advocate for designs that have the potential to provide evidence of educational success (see Cook, Chap. 17; Nolan & Temple Lang, Chap. 18).

To accomplish this goal, researchers must understand the basic relationship between the experimental/observational unit and the unit of analysis so that studies can be designed with sufficiently many experimental units and their corresponding degrees of freedom so that adequate power is present to detect meaningful differences across groups. Similarly, research questions that involve measuring multiple characteristics on each individual must be followed with a design that incorporates the multivariate nature of the data into the analysis plan. Random assignment—although not always possible—needs to be emphasized as the goal to be met rather than an option that can be avoided if desired. The proper use of random assignment will control many confounding variables and strengthen a study's findings.

But the designs will only be as good as the practices of those who implement them. More rigorous designs employed with shoddy decision-making in regard to measuring characteristics or analyzing data will not produce research that will lead to better educational policy, better educational practice, and better educational outcomes. If we want educational research to truly make a difference in our society, then we must work to eradicate poor statistical practices wherever they may exist. We have much work to do if we truly desire to make that silk purse.

References

American Educational Research Association. (n.d.). *Homepage*. Retrieved July 12, 2008, from http://www.aera.net/

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics*, *6*(3), 267–285.

Blair, R. C., Higgins, J. J., Topping, M. E. H., & Mortimer, A. L. (1983). An investigation of the robustness of the *t* test to unit of analysis violations. *Educational and Psychological Measurement*, *43*(1), 69–80.

Callaert, H. (2000, March). *Teaching introductory statistics and the use of common software packages*. Paper presented at the 2nd biennial Western Statistics Teachers Conference, Greeley, CO.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomized trials in health research*. London: Arnold.

Education Sciences Reform Act of 2002. Pub. L. No. 107–279, 116 Stat. 1940. (2002).

Fisher, R. A. (1966). *The design of experiments* (8th edn.). Edinburgh, UK: Oliver & Boyd.

Huberty, C. J., & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin*, *105*(2), 302–308.

Levin, J. R., & Serlin, R. C. (1993, April). *No way to treat a classroom: Alternative units-appropriate statistical strategies*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Lindquist, E. F. (1940). *Statistical analysis in educational research*. Boston: Houghton Mifflin.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.

No Child Left Behind Act of 2001. Pub. L. No. 107–110, 115 Stat. 1425. (2002).

Peckham, P. D., Glass, G. V., & Hopkins, K. D. (1969). The experimental unit in statistical analysis. *Journal of Special Education, 3*, 337–349.

Schield, M. (2005, August). *Statistical literacy and chance*. Paper presented at the Joint Statistical Meetings, Minneapolis, MN.

Shapiro, F. R. (Ed.). (2006). *The Yale book of quotations*. New Haven, CT: Yale University Press.

Shelley II, M. C. (2005, August). *Education research meets the gold standard: Statistics, education, and research methods after "No Child Left Behind"*. Paper presented at the Joint Statistical Meetings, Minneapolis, MN.

Slavin, R. E. (2008). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*(1), 5–14.

Swift, J. (1801). *The works of the Rev. Jonathan Swift, D.D., dean of St. Patrick's, Dublin*. London: J. Johnson.

Thompson, B. (1999, April). *Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Québec, Canada.

Part V
Public Policy and "Gold Standard(s)" Research

Chapter 22
Speaking Truth to Power with Powerful Results: Impacting Public Awareness and Public Policy

Mack C. Shelley II

This part of the book focuses specifically on the public policy issues of: (a) the ways in which global education funding patterns reflect governmental—and perhaps societal—priorities; (b) the role of research ethics boards in enforcing public policy norms regarding what is appropriate for science and literacy education research; (c) rules and expectations established by national legislative action and by professional associations for maintaining the security of the voluminous sets of data needed for sustained research excellence in science and literacy education research; (d) how qualitative research studies can be employed to provide broader and more lasting impacts on public policy making through systematic research reviews, secondary analysis, comparative case studies, and metasynthesis; and (e) how Gold Standard(s) inform education experts and policy makers about what should be done with research findings. This chapter is intended to elaborate many of the points that have been made earlier in this book and perhaps to foreshadow an action agenda for education researchers and those who seek to influence the shape and direction of public policy. One of the major lines of argument is the need for eclecticism—in methodology, subject matter expertise, and policy agendas. Consistent with that theme of the virtue and necessity of eclectic approaches, and to honor the need for truth in advertising, it may be helpful to know that the author of this chapter is a faculty member with a joint appointment in a department of statistics and in a department of political science, with about 30 years of experience with statistical consulting, and with a background in public policy, program evaluation, and public administration. That background may help explain where this chapter is coming from—as a somewhat eclectic, multifaceted exploration of a topic that is very much at the interface of several disciplines and multiple research methodologies.

M.C. Shelley II
Iowa State University

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

22.1 Speaking Truth to Power

One of the central points of the study of public policy, political science, and public administration is the artistry required to *speak truth to power*. As expressed in particular by Wildavsky (1979), the process by which experts convey the gravamen of their findings to the *powers that be* who make and enforce decisions that may be driven by those research results is an art and craft that reinforces the science and practice of politics. In highly abbreviated form, the essential point is how to reach across the gulf that is created by an unequal distribution of power (researchers having rather little and decision makers having very much more) to transmit understanding to those who are able to compel binding decisions. This involves, among other traits, the refusal to be intimidated by the presence of power, the commitment to pass on knowledge even to an audience that may not be appreciative, and the willingness and artistry to explain inconvenient truths to those who may be shown to be wrong—for example, Mathematica Policy Research showing the possible ineffectiveness of sexual abstinence education programs compared to traditional sex education programs (Trenholm et al., 2008) or the Institute of Education Sciences (IES) concluding that the federally funded and officially endorsed Reading First initiative was no better for student outcomes than alternative literacy programs (Gamse, Bloom, Kemple, & Jacob, 2008).

An important aspect of the increasingly sophisticated evaluation efforts required of scientifically based research standards is the need to strike a balance between stakeholders (such as school district administrators) and the accountability systems that require specialized expertise and that can complicate the process of speaking truth to power (Schmitt & Whitsett, 2008). Cohn (2006) noted that the tradition of scholarly detachment has led to the perception that it is difficult for academics to implement the ideas and advice they have afforded to the policy-making powers that be. He argued that academics can and should make more effective use of the opportunities that are available to them to influence public policy and that policy makers can make better use of scholarly expertise through third-community, public- and private-sector actors who influence or advise policy makers by producing and disseminating usable policy alternatives. These policy advisers include members of the research staffs of government ministries, cabinet committees, central agencies, task forces, investigatory commissions, public inquiries, research councils, private consulting organizations, political parties, interest groups, and think tanks. Cohn emphasized that academics must be sensitive to the need to join in the efforts of advocacy coalitions to situate policy decisions at the political moment when sufficient support exists for a decision to be made.

In the genre of political science, Kingdon (1995) developed a thorough conceptualization of what it takes for an idea whose time has come to make it to the decision-making phase of the policy process. Kingdon's framework uses the metaphors of the policy primeval soup and the confluence of three streams—a political stream related to elections, pressure group actions, and swings in public opinion; a policy stream, in which a policy proposal emerges as the best available alternative; and a problem stream, in which a problem emerges that is seen as important—feeding into the making of public policy by getting an issue onto the policy agenda. Kingdon's perspective emphasizes the essentiality of getting on the policy agenda by making sure a problem and its possible solutions become identified as an issue that requires public-sector attention, discussion, and action. Certainly, education issues generally are high visibility and frequently are caught up in the flow of the currents and cross-currents streaming into, through, and from the policy process. Navigating successfully the shoals and eddies of these streams, and the occasional Odyssean adventures through Scylla and Charybdis, is not for the faint of heart and requires more than the usual degree of commitment to persevere through to success.

Henig (2008) argued that, together with the old image of the ivory-tower aloof academic, "the old model of 'speaking truth to power' in which the scholar as favored advisor whispers into the ear of elite leaders, also is passé; in the age of mass media and the Internet, discourse about research has been democratized" (p. 360). This certainly does seem to be a contemporary assessment of the current state of speaking truth to power; but, far from negating the basic premise of the Wildavsky argument, it modernizes an already well-established perspective on politics, society, and how research interfaces with realities as perceived both within the corridors of power and by the public. Henig surely is correct in noting the need for academic "buffers against ideology and the politicization of the knowledge enterprise [to help maintain] a distinction between research and advocacy, between pursuit of knowledge and pursuit of advantage, between sounding good and being right" (p. 360).

Widespread dissemination and accurate interpretation of the results of education research also depends on contemporary media outlets being staffed by reporters who have sufficient background to know quality results when they see them and who are able to focus on the importance of the findings over the more headline-grabbing controversies that all too often are the natural target of media efforts to reflect or influence policy makers' opinions (Rotherham, 2008). Furthermore, academics need to be aware of the basic constraints, practices, and genre of popular media: 10-second sound bites, brief video clips, and journalistic versions of research reports of interspersed claims, evidence, and narrative that all too frequently imply applications and a degree of certainty that may not have been intended by the original researchers.

The utility of research results certainly needs to be enhanced. Brewer and Goldhaber (2008) argued that:

> since most consumers of the work will not have the time or capacity to judge its quality … [for] the rigor and relevance of educational research … to be increased, we will need a concerted effort from both consumers of research and suppliers who recognize the desperate need for improvement. (p. 364)

Getting the attention of education leaders and convincing them to make productive use of research results surely is enhanced when the research results are consistent, demonstrably relevant to the needs of educational practitioners, and disseminated quickly. That process is facilitated when fostering data literacy is a priority of school leadership and when consensus emerges on the appropriate research design strategy (Fusarelli, 2008; see also Ingersoll, 2008, on out-of-field teaching; and Kim, 2008, on reading research). Kim concluded optimistically that:

> we will be able to establish norms of excellent practice rooted in scientific research and governed by a community of peers. Ultimately, teachers must have access to truth and power if they are to create professional norms that nurture effective instruction and support efforts to help children become proficient readers. (p. 375)

Throughout this book, and perhaps especially in the chapters that constitute Part IV, the authors have addressed a multiplicity of the facets at the interface between power and expertise—where public policy joins with expert judgment and academic expertise to synergize the politics of knowledge (Hess, 2008). Hess argued that, in contrast to health care research, the record of education research is less replete with success stories, and hence "educational research has not earned similar trust or good will, and its advocates have been unsuccessful in making the case that research ought to be funded despite its painstaking pace and uncertain fruits" (p. 356). Henig (2008), going further, noted that "[a]mong policy makers and many scholars, educational research has a reputation of being amateurish, unscientific, and generally beside the point" (p. 357) and thus has less impact than it should, particularly given the internecine methodological disputations that further dispel the idea that education researchers really know what they are doing and that they know how to make proper meaning of the results.

The realization that politics plays a role in the process by which research is filtered and possibly impacts decision making certainly does not surprise the average, randomly selected, social scientist, particularly anyone who may be a card-carrying political scientist. The dimensions of this policy–politics nexus, however, may not be so thoroughly familiar to education researchers or to others who do not reflexively tune in to C-SPAN or other media-generated sources of eye-glaze to those less afflicted with the *can't-help-it* impulse to see and listen to the political process that Iron Chancellor Otto von Bismarck famously likened to sausage-making. Henig (2008) noted that the pressure to produce timely results to fit the dictates of political decision-making schedules:

> is especially the case in politically charged arenas in which groups with tactical interests in advancing or blocking specific policy actions can co-opt the process. Researchers may acknowledge the limitations of their own data and design, but those caveats are often the first things to be stripped from the message as others take it up. In practice, research that aligns with ideological cleavages is more likely to be pushed into the public realm, thus blurring the distinction between advocacy and unbiased analysis. (p. 358)

The final report of the National Mathematics Advisory Panel (NMAC, 2008), based in part on the assessment of 16,000 research publications, provided a recent example of how federal education policy can be impacted by expert panel recommendations. Convened by US President George W. Bush, the panel was formed to advise the administration on how to enhance mathematics education, with members including prestigious professors of mathematics and psychology, a middle school teacher of mathematics, and the president of the National Council of Teachers of

Mathematics. The report concluded, in part, that long-festering debates about what curricular policy to recommend are largely irrelevant:

> To prepare students for Algebra, the curriculum must simultaneously develop conceptual understanding, computational fluency, and problem-solving skills. Debates regarding the relative importance of these aspects of mathematical knowledge are misguided. These capabilities are mutually supportive, each facilitating learning of the others. Teachers should emphasize these interrelations; taken together, conceptual understanding of mathematical operations, fluent execution of procedures, and fast access to number combinations jointly support effective and efficient problem solving. (p. xix)

Similarly, the report found that intense and long-standing policy debates about the relative superiority of teacher-directed or student-directed mathematics instruction miss the point and concluded that:

> [i]nstructional practice should be informed by high-quality research, when available, and by the best professional judgment and experience of accomplished classroom teachers. High-quality research does not support the contention that instruction should be either entirely 'student-centered' or 'teacher-directed'. Research indicates that some forms of particular instructional practices can have a positive impact under specified conditions. (p. 11)

Clearly, high-quality evidence is essential, but not sufficient, in making and justifying instructional decisions. Knowing what to believe, and therefore having a better idea of what to do, is an essential prerequisite for wise public policy making. Synthesizing results across multiple, and often contradictory, studies is a form of high art requiring tools and perspectives that are not readily understandable to many researchers, let alone those who make education policy. To determine which education programs work and, therefore, deserve continued or enhanced support, Slavin (2008a) suggested the following criteria essential for valid program evaluation research: "Clear, thoughtful syntheses in many areas are crucial to providing practitioners, policy makers, and researchers with valid information they can use with confidence to address the real problems of educating all children" (p. 13). As evidenced by recent debates within the education research literature (e.g., Slavin, 2008a, 2008b, and others discussed below), several major efforts to synthesize the current state-of-the-art research record provide the foundation for intentional overviews of research results, including:

- What Works Clearinghouse (US IES, n.d.-b), officially supported by the IES of the US Department of Education (US ED) and now managed by Mathematica Policy Research, Inc.
- Best Evidence Encyclopedia (BEE, n.d.), a collaboration between the Center for Data-Driven Reform in Education in the US ED and Johns Hopkins University.
- Comprehensive School Reform Quality Center (CSRQ, 2006), active from 2003–2006 through the American Institutes for Research.
- The international Campbell Collaboration (Campbell Collaboration, n.d.).
- The United Kingdom's government-supported Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre, n.d.).

These organizations can provide collective results supported by evidence from a broad array of studies, but unfortunately evidence alone does not lead directly to policy decisions. Evidence must interact and react with the decision makers' beliefs, values, and priorities to result in evidence-based policies (see Phillips, Norris, & Macnab, Chap. 27).

22.2 A Theme to Consider: Challenge and Response

Borrowing very loosely from Toynbee's (1934–1961) 12-volume exposition of what he conceptualized as the challenge-and-response cycle throughout recorded human history, the next sections of this chapter lay out what may be regarded as the challenge posed by the present state of affairs of Gold Standard(s) expectations for education research and the response that has come, and that may be expected (or hoped for), from that research community. In doing so, the intention is to provide a broad context within which to consider the implications for education research and public policy agendas.

22.2.1 The Challenge

In the United States, and in many other countries, research funding from government agencies and other sources increasingly has become tied more closely to use of the medical model of randomized clinical trials (RCTs), featuring: (a) randomized assignment of individual subjects or clusters of subjects to treatment or control groups, (b) the need to ensure fidelity of treatment effects over both space and time, and (c) consistent and accurate measurement of well-defined outcomes. This focus on RCT-style interventions recently has been emphasized in the requirements for research in education and in other human sciences. The emphasis on the expressed needs for randomization, control, and measurement has led to a greater need for careful attention to the requirements of focused research by content experts in many diverse aspects of education inquiry and for research methods experts to be willing and available to partner in joint efforts with content specialists. These partnerships are not always easy or straightforward—particularly when there is not a lot of overlap in the substantive knowledge base and the methodological expertise of those participating in these joint ventures.

Knowing how to apply the logic of experimental and quasi-experimental methods has become essential for the successful pursuit of research awards from government sources—in the United States, from public agencies such as the US ED, the National Institutes of Health (NIH), and the National Science Foundation (NSF)—and many other funding sources (e.g., W. T. Grant Foundation, Spencer Foundation). Furthermore, it is important for successful publication of the results from such studies in appropriately high-level outlets. Competition for funding from these and other sources generally has become much fiercer; for example, what once was about a one-in-three reasonable prospect of succeeding with a grant proposal submitted to NSF now is more like a one-in-ten shot in the dark.

22.2.2 The Response

So, how has the education research community begun to address this challenge? The need to deal with the current and future situation has become a major point of discussion among researchers who are content experts in education and the social sciences, often in conjunction with their qualitative and quantitative research methodology colleagues. One recent example is the Ragin, Nagel, and White (2004) NSF-funded volume based on a workshop on the scientific foundations of qualitative research. This publication provides essential recommendations to improve the quality of qualitative research proposals and for evaluating the scientific and substantive merits of such proposals. Among the key questions addressed is: what is an ideal qualitative proposal? Ragin and colleagues also recommended how NSF (and implicitly any other public agency) can support and strengthen high-quality qualitative research, especially in light of the specific resource needs of qualitative researchers that may be understood less well by the reviewers of qualitative proposals than reviewers understand the research needs for more traditional, quantitative submissions.

The Ragin and colleagues' (2004) document provided a substantial set of recommendations for designing and evaluating an ideal qualitative research proposal "to improve the quality of qualitative research proposals and to provide reviewers with some specific criteria for evaluating proposals for qualitative research" (p. 3). It is understood that not all of these challenging targets can be met and that what is sauce for the qualitative goose also is sauce for the quantitative gander, which can be applied to research proposals of all methodological persuasions. The recommendations include the following:

- Write clearly and engagingly for a broad audience
- Situate the research in relation to existing theory
- Locate the research in the relevant literature
- Articulate the potential theoretical contribution of the research
- Outline clearly the research procedures
- Provide evidence of the project's feasibility
- Provide a description of the data to be collected
- Discuss the plan for data analysis
- Describe a strategy to refine the concepts and construct theory
- Include plans to look for and interpret disconfirming evidence
- Assess the possible impact of the researcher's presence & biography
- Provide information about research replicability
- Describe the plan to archive the data (pp. 3–4)

A second set of recommendations addresses ways in which the research grants process can better support and strengthen qualitative research and enhance the productivity of qualitative researchers, taking into account particularly the resource needs of qualitative researchers:

- Solicit proposals for workshops and research groups on cutting-edge topics in qualitative research methods

- Encourage investigators to propose qualitative methods training [professional development and education]
- Provide funding for such opportunities to improve qualitative research training [professional development and education]
- Inform potential investigators, reviewers, and panelists of qualitative proposal review criteria
- Give consideration, contingent upon particular projects, to fund release time for qualitative researchers beyond the traditional 2 summer months
- Fund long-term research projects beyond the traditional 24 months
- Continue to support qualitative dissertation research
- Continue to support fieldwork in multiple sites (Ragin et al., 2004, p. 4)

On the quantitative side of the research methodology spectrum, the American Statistical Association (ASA, 2007), in collaboration with NSF, produced the seminal report *Using Statistics Effectively in Mathematics Education Research*. This publication was the product of 3 years of NSF-funded workshops conducted by the ASA's Working Group on Statistics in Mathematics Education Research, whose membership included leading experts in mathematics education, psychology, measurement, and statistics. Five steps to effective quantitative research in education were noted, which are summarized below.

Step 1: Generate research ideas. Recommendations to researchers include:

- Identify ideas and questions about a topic of interest.
- Determine specific research questions to investigate.
- Build an argument about why this question is worth investigating.
- Make the researchers' beliefs and assumptions about the topic explicit.
- Examine primary and secondary research literature to clarify the researchers' beliefs, biases, and assumptions about the research topic.
- Review existing research and nonresearch literature to determine the current state of knowledge about the questions.
- Determine the concepts and constructs associated with the topic; develop a conceptual framework linking the concepts and constructs.
- Identify research methods (e.g., experimental methods, cognitive models, participant observation) that can provide information about the concepts and constructs.
- Synthesize knowledge about the research question to date.

Step 2: Frame the research program, considering goals and constructs, measurement, and logistics and feasibility.

(a) Goals and constructs recommendations include:

  - Propose a conceptual model linking the constructs.
  - Explore existing data and observations.
  - Identify relevant variables and define them operationally.
  - Use past data and observations to develop potential hypotheses.
  - Determine appropriate research methods.
  - Identify relevant measures or the need for new measures.

- Gather exploratory empirical data to test the research framework.
- Formulate a research question; outline a plan to answer the question.
- Discuss the possibility that measures (e.g., gain scores) may lead to faulty conclusions.
- Provide exploratory and descriptive statistics with appropriate graphs and interpretations.

(b) Measurement recommendations focused on developing and reporting on assessment measures used in education research that have the qualities of validity (the extent to which a measure is meaningful, relevant, and useful for the research at hand), reliability (the extent to which the measure is free of random error), and fairness (the extent to which measures are implemented consistently and validly for all subgroups) include:

- Examine previously used measures; decide if it is necessary to create new ones.
- Provide key details regarding development of new measures and/or selection of off-the-shelf measures.
- Report the relationships each variable has with other variables used in the research.
- Explain how measures align with the goals of the research.
- Determine the sample or population from which measures will be obtained.

(c) Logistics and feasibility recommendations include:

- Consider potential ethical issues and risks associated with the proposed interventions.
- Document and test the procedures to be used in an intervention study.
- Design and conduct a qualitative component to assess measurement difficulties and possible lack of feasibility of the study.
- Investigate how to deal with problems, such as study dropouts and missing data.
- Examine and evaluate threats to internal and external validity.
- Develop trust within the research setting.
- Search for useful common measures that can be related to other research.
- Develop, if necessary, tests to determine interrater reliability and internal validity; refine measures.
- Pilot all instruments in an informal setting; conduct a formal field test or pilot study.
- Develop a plan for the formative evaluation of an intervention.
- Meet institutional review board guidelines, ensuring confidentiality and informed consent.
- Anticipate problems in the field; develop an affordable contingency plan.
- Develop a work plan to coordinate measurement and evaluation within an individual site or among multiple sites.
- Determine any demographic differences between the population and the sample studied.
- Describe the method of sampling, if any.

- Identify the sampling unit and the unit of analysis.
- Describe the treatment and measures in enough detail to allow replication.
- Make sure that adequate time, training, and support services exist to perform the study.

Step 3: Examine the research program. By establishing efficacy, the research program can progress to studies that may be able to establish causal patterns. Recommendations include:

- Specify a study design and the associated data analysis plan.
- Identify subpopulations of interest.
- Define the setting in which the study is to be conducted.
- Identify sources of (extraneous) variability; take steps to control variability.
- Refine measures based on research experience.
- Assess the potential portability of measures to broader contexts.
- Ensure that the intervention received by one subject is independent of the person administering it and independent of the other intervention recipients.
- Provide estimates of statistical parameters as well as the results of hypothesis testing.

Special care must be taken to ensure that statistical results are understandable, correct, and interpreted appropriately. For formal statistical inference, researchers should:

- State the hypotheses clearly.
- Specify a statistical model that addresses the research question.
- Define the population of interest and exclusion/inclusion criteria.
- Describe the characteristics of the study sample.
- Describe how random assignment or random selection was used.
- Describe whether implementation was carried out appropriately.
- Explain measures taken to minimize bias.
- Report statistical power and effect size results.
- Report response rates.
- Provide margins of error or confidence intervals.
- Explain how missing data were handled.
- Describe adjustments to minimize the risk of false positive results from multiple tests.
- Summarize the results of tests of assumptions and diagnostic (e.g., goodness of fit) tests.
- Provide sufficient information to replicate the analysis.
- Consider how to link with other databases.

Step 4: Generalize the research program. This usually involves ramping up to larger studies that randomize classes, groups, or individual subjects to the intervention with appropriate within-study controls on the measurement processes to allow the strongest possible interpretation of causal relationships. Recommendations include:

- Assess the potential portability of measures to multiple institutions in a wide variety of social contexts.

- Design and conduct a multi-institutional randomized study.
- Design and conduct a quasi-experiment.
- Conduct a rigorous statistical analysis of the quantitative results of a multi-institutional study (e.g., a survey, an experiment, an observational study) using statistical methods appropriate to the unit of analysis.
- Specify outcomes: intermediate outcomes (goals) and primary and secondary outcomes.
- Specify how covariates were defined, measured, and used.
- Detail appropriate research designs to test the hypothesis (e.g., experiment, quasi-experiment, matching, repeated measures).

Step 5: Extend the research program. A rigorous, generalized study can be achieved by, for example, syntheses of multiple studies, longitudinal studies of long-term effects, and developing policies for implementation. Ongoing formative evaluations are essential to inform the research team about necessary research adjustments and how to improve measures and procedures. Recommendations include:

- Design and conduct a longitudinal study that allows rigorous statistical inferences over time and long-term improvements in curriculum and student performance.
- Describe the nature of the long-term study (e.g., experimental, quasi-experimental, sample survey, observational).
- Describe the rate of dropouts over time and how this was handled in the analysis.
- Describe how the study maintained measurement integrity over time and in different circumstances.

The linkages connecting methodological research sophistication—whether of qualitative, quantitative, or mixed lineage—with content research expertise and the public policy implications of the results of that research have been drawn out by authors representing a broad range of disciplines and sharing a commitment to ensuring that elected and appointed powers can understand the import of the research and make appropriate use of those findings in formulating public policy decisions. As a case in point, the January/February 2008 issue of the American Educational Research Association's *Educational Researcher* (the contents of which are cited extensively below) offers a full spread of articles revolving around measurement issues that arise in the often tricky business of synthesizing the results of multiple educational program evaluations. This issue features a lead article by Robert Slavin (2008a), with replies by Derek Briggs (2008), Madhabi Chatterji (2008), Mark Dynarski (2008), Judith Green and Audra Skukauskaité (2008), and Finbarr Sloane (2008), with a response by Slavin (2008b).

Slavin's (2008a) argument is that syntheses of research on educational programs have become more important for affecting public policy. Thus, it is increasingly important for such syntheses to produce reliable, unbiased, and meaningful evidence-based interpretations of program results. The number of evaluations of any given program tends to be small, so it is essential to minimize bias in reviewing each study. This is achieved by exercising great care in determining and explaining research design, sample size, any adjustments that may have been made for pretest differences, how long the study lasted, effect sizes, and the number of relevant studies.

Careful research synthesis can result in more meaningful ratings of the strength of evidence for the effectiveness of each program. Particularly for researchers who invest heavily in comparing results across multiple studies and make use of various forms of meta-analysis, this is a must-read opportunity.

Another example of the discussion/debate regarding the role of education research in impacting public policy is afforded in the January 2008 edition of *Phi Delta Kappan* (PDK) on the "Politics of Knowledge," which addresses how educational research may be used to inform policy decisions and foster democratic government. This issue manifests various views—many of which are explored in other sections of this chapter and book—of the education policy process from a number of disciplinary perspectives, including political science, economics, policy studies, urban studies, public affairs, educational leadership and policy studies, sociology, and wonkish think tanks. The present book offers its own contribution to the growing volume of literature on education and policy research provided by these and many other authors and outlets.

22.3 What's all the fuss about, anyway? A Brief Backgrounder

In policy circles, it is pretty much de rigueur to provide a background summary of why we are all gathered together to address any given policy issue. Here is a quick overview, as well as a reminder, of essential points that arise in the debate surrounding Gold Standard(s), building on comments made earlier in this book.

Standards for acceptable and, particularly, fundable research, especially in the context of the US ED, have been affected greatly by two major policy innovations: the No Child Left Behind Act of 2001 (NCLB, 2002) passed on January 8, 2002, and the Education Sciences Reform Act of 2002 (ESRA, 2002) passed on January 23, 2002. The latter of these statutes resulted in creation of the IES (US IES, n.d.-a) in the US ED. Together, these developments have reconstituted federal support for research and dissemination of information in education with ramifications for education research in other countries; they are meant to foster scientifically valid research and have established what often is referred to as the Gold Standard for research in education.

These and other developments denote that greater emphasis in fundable education research now is placed on quantification, the use of randomized trials, and the selection of valid control groups. To meet this challenge, there is an obvious need for experts in research design and research methods to work together with content experts, to apply appropriate methods of measurement, analysis, and interpretation.

NCLB was identified in the legislation as "An Act to close the achievement gap with accountability, flexibility, and choice, so that no child is left behind" (NCLB, 2002, para. 1); hence, the eponymous label of the law, which was officially the 2002 reauthorization of the Elementary and Secondary Education Act of 1965. The NCLB established standards for academic assessments in mathematics, reading or language arts, and science; it required multiple, up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding.

ESRA, or HR 3801, defined *scientifically based research* and *scientifically valid education evaluation* standards to apply rigorous, systematic, and objective methodology to obtain reliable and valid knowledge relevant to education activities and programs, and to present findings and make claims that are appropriate to and supported by the methods that have been employed. In HR 3801, scientifically based research includes systematic, empirical methods that draw on observation or experiment; data analyses that are adequate to support the general findings; measurements or observational methods that provide reliable data; making claims of causal relationships only in random assignment experiments or other designs that substantially eliminate plausible competing explanations; replication or the opportunity to build systematically on the findings of the research; obtaining acceptance by a rigorous, objective, and scientific review; and research designs and methods appropriate to the research question posed.

HR 3801 also specified that scientifically valid education evaluation adheres to the highest possible standards of quality with respect to research design and statistical analysis; provides an adequate description of the programs evaluated and, to the extent possible, examines the relationship between program implementation and program impacts; provides an analysis of the results achieved by the program with respect to its projected effects; employs experimental designs using random assignment when feasible and other research methodologies that allow for the strongest possible causal inferences when random assignment is not feasible; and may study program implementation through a combination of scientifically valid and reliable methods.

Other countries have attempted to enhance educational research using other quality assurance approaches (see Coll et al., Chap. 6). Their approaches are not driven by prescriptive government policy regarding appropriate research approaches, but they do assess the quality of sponsoring institutions, researchers, and proposals in a variety of ways. Funding agencies use evaluation criteria as another tool to facilitate or restrict research approaches (see She et al., Chap. 23).

22.4 How Does the Gold Standard Connect with Public Policy?

Much of the policy debate swirling around implementation of NCLB and the overall US federal government effort to upgrade the quality of education research is related to the creation of IES as the research arm of the US ED through HR 3801. Its mission is to expand knowledge and provide information on the condition of education, practices that improve academic achievement, and the effectiveness of federal and other education programs. Its expressed goal is the transformation of education into an evidence-based field in which decision makers routinely seek out the best available research and data before adopting programs or practices that will affect significant numbers of students (see Hayward & Phillips, Chap. 7).

Perhaps the best articulation of what is meant by the concept and implementation of the Gold Standard for research is provided by the IES's National Center for

Education Evaluation and Regional Assistance (US ED, 2003) user-friendly guide to identifying and implementing educational practices supported by rigorous evidence. The rules of evidence of education interventions come in two levels. The quality of studies needed to establish *strong* evidence requires: (a) RCTs that are well-designed and implemented, following a medical clinical trials model; and (b) that the quantity of evidence needed spans trials showing effectiveness in two or more typical school settings. *Possible* evidence may include: (a) RCTs whose quality/quantity are good but fall short of strong evidence; (b) and/or comparison-group studies in which the intervention and comparison groups are very closely matched in academic achievement, demographics, and other characteristics.

Evaluating whether an intervention is backed by strong evidence of effectiveness hinges on well-designed and well-implemented RCTs, demonstrating that there are no systematic differences between intervention and control groups before the intervention, using measures and instruments of proven validity, and demonstrating the presence of real-world objective measures of the outcomes the intervention is designed to affect. The benchmarks for evaluating whether an intervention is backed by strong evidence of effectiveness include attrition of no more than 25% of the original sample in longitudinal studies, effect size measures of the estimated amount of impact, and *p* values at the traditional level of 0.05 or less, adequate sample size to achieve statistical significance, and controlled trials implemented in more than one site representing a cross section of all schools.

For researchers in search of guidance on the essential quantifiable aspects of research design, an excellent source is the W. T. Grant Foundation’s Optimal Design software (Raudenbush, Spybrook, Liu, Congdon, & Martinez, 2006; W. T. Grant Foundation and University of Michigan, n.d.). Excellent guidance also is provided by Lenth’s (2006) Java applets for determining statistical power (Murphy & Myors, 2003; Schochet, 2005) and sample size. Additional guidance is available online through the What Works Clearinghouse (WWC, US IES, n.d.-b) and Campbell Collaboration (Campbell Collaboration, n.d.).

WWC was established in 2002 by IES to provide educators, policy makers, and the public with a central and trusted source of scientific evidence of what works in education. It reviews and reports on existing studies of interventions (education programs, products, practices, and policies) in selected topic areas that apply standards that follow scientifically valid criteria for determining the effectiveness of these interventions. It also provides technical assistance and a registry of evaluators (US IES, n.d.-c) as well as technical working papers (US IES, n.d.-d). These online assessments and documentation are reviewed by a Technical Advisory Group. As of this writing, WWC has provided detailed results for programs in (a) beginning reading, (b) early childhood education, (c) elementary school mathematics, (d) middle school mathematics, (e) character education, (f) dropout prevention, and (g) English language learning. The most fully elaborated information is available on the first four topics. In each area, WWC evaluates program effectiveness as: meets evidence standards, meets evidence standards with reservations, or does not meet evidence screens. Each specific intervention program is evaluated as having: positive effects, potentially positive effects, mixed effects, no discernible effects, potentially negative effects, or negative effects.

22.5 A Possible Template for Science and Literacy Education Research?

So, in this climate what guidance can be provided to education researchers, particularly those in the fields of science education and literacy research? One possible template for how best to impact the science and literacy education policy areas may be afforded in the aforementioned ASA report (2007). Although focused on mathematics education, it offers some suggestions for research and guidance for actions that may be particularly helpful in the process of trying to speak truth to power.

The focus of the mathematics education template is on how best to cumulate the results of a larger corpus of individual studies to achieve a potentially high-impact summary of programmatic interventions in education. This involves consistent and appropriate use of interventions, observation and measurement tools, data collection techniques, and data analysis methods, and consistent reporting of research results. Doing so facilitates replication (or at least another look at the same problem) and, therefore, makes it more feasible to progress toward the goal of achieving a cumulative discipline, which is commonly seen as a hallmark of *science*. To achieve this goal requires both methodological rigor and methodological diversity, as elaborated, for example, by Raudenbush (2005) and the US National Research Council (US NRC, 2002).

The concept of using larger bodies of studies to help inform policy makers through wider application of both quantitative and qualitative forms of integration and synthesis requires further elaboration. In particular, it is essential to highlight the difference between meta-analysis based on aggregating clinical trials conducted for medical research and meta-analysis conducted using the often more tenuous results of education research. Research protocols, and in many cases measurement procedures and data analysis methods, often are well established and—although often couched in highly technical terminology—frequently are understandable to the general public and are explained and interpreted routinely by mass media outlets. In addition, medical experts are frequent visitors to US government executive-branch agencies and to congressional committees; medical experts generally are given a positive reception, indicating that their expertise is widely understood and respected—if not unchallenged.

In contrast to the apparent near-certainty of the results of medical trials, particularly when aggregated across relatively large numbers of broadly similar studies, the reception often afforded to educational researchers frequently is much less positive. This divergence in the amount of slack given to education—as opposed to medical—researchers by the mass media, the general public, and policy makers may be attributable in large part to the diversity and variety of ideological and methodological positions adopted by educational researchers, exacerbated by disciplinary differences among those who conduct research in education spanning higher education, preK-12, sociology, political science, economics, psychology, statistics, and other areas of expertise that may not speak the same language or use the same procedures or methodologies. Another consequential difference is that most medical clinical trials research is designed to measure the impact of drugs administered in usually carefully controlled environments, such that the analysis conducted on the data resulting from those studies often is not complicated by the need to control statistically for other, potentially confounding, variables. In contrast, even with randomized cluster trials conducted at multiple sites, the analysis of data from education research often needs to be adjusted with covariates and frequently is based on outcome measures that are less precise than what can be achieved under clinical trials laboratory conditions. The greater difficulty in achieving sustained precision of measurement and clear data analysis is compounded further by the lesser amounts of funding available for many education experiments or quasi-experiments; if a standard, medical-style, clinical trials experiment is funded at something like $15 million spread over 5 years, the typical funding for an education intervention study is likely to be much less in total amount and may not be sustained for as long.

Building on the themes of methodological diversity and the need for cumulative findings to maximize the impact of those findings on education-relevant public policy, well-established methods exist in the literature on meta-analysis (e.g., Cooper & Hedges, 1994; Hunter & Schmidt, 2004; Lipsey & Wilson, 2001) and multilevel models (Arnold, 1992; Bock, 1989; Bryk & Raudenbush, 1987, 1989, 1992; Goldstein, 2003; Hedges, 2007a, 2007b; Lee & Bryk, 1989; Raudenbush & Bryk, 1986, 2002) accomplish these goals, by aggregating quantitative results across contexts and across units of measurement (such as individual students, classrooms, or districts). A particularly fruitful line of research is to adjust for aggregate setting effects on student outcomes in cluster randomized trials (Donner & Klar, 2000; Murray, 1998; Raudenbush, 1997; Raudenbush & Liu, 2000), whereby schools or districts constitute the units that are randomized and individual student results are aggregated and compared across those settings of intervention (e.g., Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007; Raudenbush, 1997; Raudenbush & Liu, 2001; Raudenbush, Martinez, & Spybrook, 2007).

Borrowing from the ASA (2007), if research in science education and literacy is to have more effective influence on policy and practice, it must become more cumulative in nature, as suggested for mathematics education. This requires building on existing research to produce a more coherent body of work. Education researchers must be free to pursue problems and questions that are of interest to them. To influence practice, however, the work must be situated within a larger corpus. There is power in numbers—both in the number of studies and in the number of researchers agreeing with each other. Cumulating studies through consistent use of interventions, observation and measurement tools, data collection techniques, data analysis methods, and reporting of research results facilitates replication (or at least another look at the same problem).

Based on the results of pilot studies and the use of appropriate methods for data collection and analysis, the goal is to generalize the findings from a research program. This is accomplished by first establishing the efficacy of the study then determining its portability (Will it work the same anywhere?) and scalability (Will it work the same when we do this big time?). Extending the research program is best accomplished by synthesizing multiple studies through methods of meta-analysis, conducting longitudinal studies of long-term effects using growth-curve models, and developing an implementation policy that can get large-scale funding and political support. Ongoing formative evaluations are needed, to permit mid-course corrections if they are needed. This is the payoff of speaking truth to power successfully.

Research methods expertise comes in extremely handy in this process. As something of a shameless advertisement, it often is a very good idea to add a research methods specialist to a team writing grant proposals. Doing so often helps improve the prospects for obtaining funding. It usually helps with establishing the rigor of the research design and may help get results that may be listened to by the powers that be. The methodological bag of tricks includes, for example, expertise in focus groups, document and content analysis, interview strategies, logic models, experimental and quasi-experimental design, and working with large, complex, and/or messy databases with methods such as data mining (see Wang, Dziuban, Cook, & Moskal, Chap. 19; Ye, 2003). In particular, quantitative methodologists know how to handle the nearly inevitable complications that arise from the presence of missing data through the use of imputation, plausible values, survey weights, poststratification, and other mechanisms.

Complex contemporary methods of data analysis that convey powerful results to policy makers include hierarchical linear modeling (e.g., Arnold, 1992; Bryk & Raudenbush, 1987, 1989, 1992; Cohen, 1988; Goldstein, 1987; Hox, 2002; Lee & Bryk, 1989; Raudenbush & Bryk, 1986, 2002; Raudenbush, Bryk, Cheong, & Congdon, 2001; Reise & Duan, 2002), structural equation modeling (e.g., Bollen, 1989; Bollen & Long, 1993; Byrne, 1998, 2001; Jöreskog & Sörbom, 1996a, 1996b; Loehlin, 2003), meta-analysis, and all sorts of other fancy models. Of particular interest for influencing policy decisions through contemporary data analysis is measuring temporal changes in targeted outcomes. This requires the use of growth-curve modeling (Bollen & Curran, 2005) to measure change in outcomes as a function of time and to predict the rate and pattern of growth. Growth-curve modeling of individual change circumvents the limitations inherent in traditional repeated-measures analysis of variance (Agresti & Finlay, 1997; Howell, 2007; Shadish, Cook, & Campbell, 2002), in which restrictive assumptions (such as sphericity or constant correlations over time) often are not met. Traditional growth-curve modeling ignores individual growth trajectories, which are treated as error, but has difficulty dealing with missing data and inconsistent time periods; these are severe problems because frequently longitudinal studies suffer from relatively high rates of attrition, and it is difficult to sustain repeated measurements at nearly equal intervals.

An enhancement of traditional growth-curve models is provided by the analysis of individual growth curves, in which within-individual change is modeled as a function of time, providing for both linear growth (instantaneous growth rate at intercept) and curvilinear growth (acceleration) in the level-1 (individual-level) model and with predictors of baseline performance, of initial learning rate, and of acceleration in the level-2 (aggregate) model. How to model the time variable is a major methodological issue (see van den Bergh et al., Chap. 20). For example, centering the time variable can dramatically change the interpretation of lower-order coefficients. Variance in the coefficients for individuals may reflect important individual differences in, for example, students' rate of learning and their sensitivity to contextual circumstances. Residuals from such individual growth-curve models reflect individual differences among students, which then can be used as predictors for other analyses. More complex models are possible by incorporating time-varying covariates at level 1, individual covariates at level 2, modeling heterogeneous level-1 variance and autocorrelation, and specifying complex error structures using a hierarchical, multivariate, linear model.

22.6 A Brief Segue

These and other dimensions of research methodology expectations rise to the forefront when the inevitable need arises to put together the research team that must consolidate qualitative and quantitative expertise with content knowledge and to develop the synergies that are essential for successful research proposals. To conclude this chapter, it may be helpful to take note of the practicalities of what must be done and the complexities that need to be addressed in the pursuit of funded research from a major grant opportunity directed toward impacting public policy.

To help make the preceding discussion about methodological research needs more concrete, we examine below the methodological requirements for IES request for proposal CFDA (Catalogue of Federal Domestic Assistance) 84.305A for the Education Research Grants program (US IES, 2008). All of these methodological requirements must be addressed, as specifically as possible. This requires teamwork between content and methods/measurement experts. The hoped-for result is a more vigorous, externally funded, research program.

22.6.1 Measurable Outcomes

> (a) [r]eadiness for schooling (pre-reading, pre-writing, early mathematics and science knowledge and skills, and social development);
> (b) [a]cademic outcomes in reading, writing, mathematics, and science;
> (c) [s]tudent behavior and social interactions within schools that affect the learning of academic content;
> (d) [s]kills that support independent living for students with significant disabilities; and
> (e) [e]ducational attainment (high school graduation, enrollment in and completion of post-secondary education). (US IES, p. 8)

22.6.2 Five Research Goals

(a) Goal One – identify existing programs, practices, and policies that may have an impact on student outcomes and the factors that may mediate or moderate the effects of these programs, practices, and policies;
(b) Goal Two – develop programs, practices, and policies that are theoretically and empirically based …;
(c) Goal Three – establish the efficacy of fully developed programs, practices, and policies …;
(d) Goal Four – provide evidence on the effectiveness of programs, practices, and policies implemented at scale; and
(e) Goal Five – develop or validate data and measurement systems and tools. (p. 9)

22.6.3 Methodological Requirements

- Clear, concise hypotheses or research questions. (p. 51)
- Sample to be selected and sampling procedures to be employed …, including justification for exclusion and inclusion criteria. (p. 59) [Describe strategies to increase the likelihood that participants will remain in the study over the course of the evaluation (i.e., reduce attrition).]
- Detailed research design. … Studies using randomized assignment to treatment and comparison conditions are strongly preferred. … [C]learly state the unit of randomization (e.g., students, classroom, teacher, or school) … [and] explain the procedures for assignment of groups (e.g., schools, classrooms) or participants to treatment and comparison conditions. (p. 59)

Only when a randomized trial is not possible may alternatives that substantially minimize selection bias or allow it to be modeled be employed. Applicants proposing to use a design other than a randomized design must make a compelling case that randomization is not possible. Acceptable alternatives include regression-discontinuity designs or other well-designed, quasi-experimental designs that minimize the effects of selection bias on estimates of effect size through propensity score balancing or regression.

- The power of the evaluation design to detect a reasonably expected and minimally important effect … indicate clearly (e.g., including the statistical formula) how the effect size was calculated. (p. 60)

For clusters or groups of students randomly assigned to treatment and comparison conditions, consider the number of clusters, the number of individuals within clusters, the potential adjustment from covariates, the desired effect, the intraclass correlation (Killip, Mahfoud, & Pearce, 2004, i.e., the variance between clusters relative to the total variance between and within clusters), and the desired power of the design (note that other factors may also affect the determination of sample size, such as using one-tailed versus two-tailed tests, repeated observations, attrition of participants, etc.).

- Measures of student outcomes [including] researcher developed measures and … relevant … standardized measures of student achievement. (US IES, p. 61)
- Fidelity of implementation of the intervention … how the implementation of the intervention would be documented and measured. (p. 61)
- Compare intervention and comparison groups on the implementation of critical features of the intervention [to connect observed differences to treatment effects.] … [A]void contamination between treatment and comparison groups. (pp. 61–62)
- Mediating and moderating variables … that may explain the effectiveness or ineffectiveness of the intervention. … [A]ccount for sources of variation in outcomes across settings (i.e., to account for what might otherwise be part of the error variance). … [D] emonstrate the conditions and critical variables that affect the success of a given intervention. The most scalable interventions are those that can produce the desired effects across a range of education contexts. (p. 62)
- Data analysis. All proposals must include detailed descriptions of data analysis procedures. … Most evaluations of education interventions involve clustering of students in classes and schools and require the effects of such clustering to be accounted for in the analyses, even when individuals are randomly assigned to condition. Such circumstances generally require specialized multilevel statistical analyses using computer programs designed for such purposes. (p. 62)

22.7 Where Do We Go from Here?

The subsequent chapters contributed to Part V of this book span a wide variety of the implications for public policy of expectations/requirements for Gold Standards education research in many different countries and in diverse contexts. The next chapter examines the interplay between the needs for scientifically based research and the provision of research expenditures from the perspectives of education research in the United States, Canada, the European Union, Germany, the United Kingdom, Australia, New Zealand, southern Africa, and Taiwan (Republic of China). A similarly transnational range of views is afforded in the chapter on research ethics, which explores the diversity of policies in different countries and in different institutions regarding human subjects protections in education research and the varying extent to which constraints are imposed on education researchers. Another chapter addresses policies related to data sharing, including data disclosure, confidentiality, and security. Qualitative metasynthesis, applying aspects of meta-analysis from quantitative methodology, is addressed by another set of authors as a means for revealing general patterns in systematic research reviews, metasyntheses, secondary reanalyses, and case-to-case comparisons of qualitative research studies. The part concludes with a call to educators to make better use of the results of science to change social practice.

The chapters in this part, and indeed all the contributions throughout this book, reveal a compelling need for a self-conscious, deliberate, and directed effort to integrate methodology, policy, and advocacy into a coherent approach to speaking truth to power. Knowing what the truth is, when it is ripe for sharing, whom to share it with, and how to convey it with maximum impact are all essential aspects of what is to be done.

References

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd edn.). Upper Saddle River, NJ: Prentice Hall.

American Statistical Association. (2007). *Using statistics effectively in mathematics education research*. Arlington, VA: Author. Available from http://www.amstat.org/research_grants/pdfs/SMERReport.pdf

Arnold, C. L. (1992). An introduction to hierarchical linear models. *Measurement & Evaluation in Counseling & Development*, *25*(2), 58–90.

Best Evidence Encyclopedia. (n.d.). *Homepage*. Retrieved July 7, 2008, from http://www.bestevidence.org/

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*(1), 30–59.

Bock, R. D. (Ed.). (1989). *Multilevel analysis of educational data*. San Diego, CA: Academic Press.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A., & Curran, P. J. (2005). *Latent curve models: A structural equation perspective*. New York: Wiley.

Bollen, K. A., & Long, J. S. (Eds.). (1993). *Testing structural equation models*. Newbury Park, CA: Sage.

Brewer, D. J., & Goldhaber, D. D. (2008). Examining the incentives in educational research. *Phi Delta Kappan*, *89*(5), 361–364.

Briggs, D. C. (2008). Comments on Slavin: Synthesizing causal inferences. *Educational Researcher*, *37*(1), 15–22.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*(1), 147–158.

Bryk, A. S., & Raudenbush, S. W. (1989). Toward a more appropriate conceptualization of research on school effects: A three level hierarchical linear model. In R. D. Bock (Ed.), *Multilevel analysis of educational data*. San Diego, CA: Academic Press.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications and programming*. Mahwah, NJ: Lawrence Erlbaum.

Campbell Collaboration. (n.d.). *Homepage*. Retrieved October 3, 2007, from http://www.campbellcollaboration.org

Chatterji, M. (2008). Comments on Slavin: Synthesizing evidence from impact evaluations in education to inform action. *Educational Researcher*, *37*(1), 23–26.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edn.). Hillsdale, NJ: Lawrence Erlbaum.

Cohn, D. (2006). Jumping into the political fray: Academics and policy-making. *Institute for Research on Public Policy (IRPP) Matters*, *7*(3), 8–36. Retrieved from http://www.irpp.org/pm/index.htm

Comprehensive School Reform Quality Center. (2006). *Homepage*. Retrieved July 21, 2008, from http://www.csrq.org/

Cooper, H., & Hedges, L. V. (Eds.) (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomized trials in health research*. London: Arnold.

Dynarski, M. (2008). Comments on Slavin: Bringing answers to educators: Guiding principles for research syntheses. *Educational Researcher*, *37*(1), 27–29.

Education Sciences Reform Act of 2002. Pub. L. No. 107–279, 116 Stat. 1940. (2002).

Evidence for Policy and Practice Information and Co-ordinating Centre. (n.d.). *Homepage*. Retrieved June 20, 2008, from http://eppi.ioe.ac.uk/cms/

Fusarelli, L. D. (2008). Flying (partially) blind: School leaders' use of research in decision making. *Phi Delta Kappan*, *89*(5), 365–358.

Gamse, B. C., Bloom, H. S., Kemple, J. J., & Jacob, R. T. (2008). *Reading first impact study: Interim report*. (NCEE 2008–4016). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Available from http://ies.ed.gov/ncee/pubs/20084016/index.asp

Goldstein, H. (1987). *Multilevel models in educational and social research*. New York: Oxford University Press.

Goldstein, H. (2003). *Multilevel statistical models* (3rd edn.). London: Hodder Arnold.

Green, J. L., & Skukauskaité, A. (2008). Comments on Slavin: Becoming critical readers: Issues in transparency, representation, and warranting of claims. *Educational Researcher*, *37*(1), 30–40.

Hedges, L. V. (2007a). Correcting a significance test for clustering. *Journal of Educational & Behavioral Statistics*, *32*(2), 151–179.

Hedges, L. V. (2007b). Effect sizes in cluster-randomized designs. *Journal of Educational & Behavioral Statistics*, *32*(4), 341–370.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation & Policy Analysis*, *29*(1), 60–87.

Henig, J. R. (2008). The evolving relationship between researchers and public policy. *Phi Delta Kappan*, *89*(5), 357–360.

Hess, F. M. (2008). The politics of knowledge. *Phi Delta Kappan*, *89*(5), 354–356.

Howell, D. C. (2007). *Statistical methods for psychology* (6th edn.). Pacific Grove, CA: Duxbury.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd edn.). Thousand Oaks, CA: Sage.

Ingersoll, R. M. (2008). A researcher encounters the policy realm: A personal tale. *Phi Delta Kappan*, *85*(5), 369–371.

Jöreskog, K. G., & Sörbom, D. (1996a). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (1996b). *LISREL 8: User's reference guide*. Chicago: Scientific Software International.

Killip, S., Mahfoud, Z., & Pearce, K. (2004). What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *Annals of Family Medicine*, *2*(3), 204–208.

Kim, J. S. (2008). Research and the reading wars. *Phi Delta Kappan*, *85*(5), 372–375.

Kingdon, J. W. (1995). *Agendas, alternatives, and public policies* (2nd edn.). New York: HarperCollins.

Lee, V. E., & Bryk, A. S. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education*, *62*(3), 172–192.

Lenth, R. V. (2006). Java applets for power and sample size. Available from http://www.cs.uiowa.edu/~rlenth/Power/

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Loehlin, J. C. (2003). *Latent variable models: An introduction to factor, path, and structural equation analysis* (4th edn.). Mahwah, NJ: Lawrence Erlbaum.

Murphy, K. R., & Myors, B. (2003). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (2nd edn.). Mahwah, NJ: Lawrence Erlbaum.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the national mathematics advisory panel*. Washington, DC: US Department of Education. Available from http://www.ed.gov/about/bdscomm/list/mathpanel/report/final-report.pdf

No Child Left Behind Act of 2001. Pub. L. No. 107–110, 115 Stat. 1425. (2002).

Ragin, C. C., Nagel, J., & White, P. (2004). Workshop on scientific foundations of qualitative research. Available from http://www.nsf.gov/pubs/2004/nsf04219/start.htm

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*(2), 173–185.

Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher, 34*(5), 25–31.

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*(1), 1–17.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd edn.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. (2001). *HLM5: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International Inc.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199–213.

Raudenbush, S. W., & Liu, X. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, *6*(4), 387–401.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation & Policy Analysis*, *29*(1), 5–29.

Raudenbush, S. W., Spybrook, J., Liu, X., Congdon, R., & Martinez, A. (2006). Optimal Design (Version 1.76) [Software]. Available from http://sitemaker.umich.edu/group-based/optimal_design_software

Reise, S. P., & Duan, N. (Eds.). (2002). *Multilevel modeling: Methodological advances, issues, and applications*. Mahwah, NJ: Lawrence Erlbaum.

Rotherham, A. J. (2008). The translators: The media and school choice research. *Phi Delta Kappan*, *89*(5), 376–379.

Schmitt, L. N. T., & Whitsett, M. D. (2008). Using evaluation data to strike a balance between stakeholders and accountability systems. In T. Berry & R. M. Eddy (Eds.), *Consequences of No Child Left Behind for educational evaluation* (pp. 47–58). San Francisco: Jossey-Bass.

Schochet, P. Z. (2005). *Statistical power for random assignment evaluations of education programs*. (Contract No. ED-01-CO-0038-0009; MPR Reference No. 6046-310). Princeton, NJ: Mathematica Policy Research, Inc. Available from http://www.mathematica-mpr.com/publications/PDFs/statisticalpower.pdf

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Slavin, R. E. (2008a). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, *37*(1), 5–14.

Slavin, R. E. (2008b). Response to Comments: Evidence-based reform in education: Which evidence counts? *Educational Researcher*, *37*(1), 47–50.

Sloane, F. (2008). Comments on Slavin: Through the looking glass: Experiments, quasi-experiments, and the medical model. *Educational Researcher*, *37*(1), 41–46.

Toynbee, A. J. (1934–1961). *A study of history*. London: Oxford University Press.

Trenholm, C., Devaney, B., Fortson, K., Clark, M., Quay, L., & Wheeler, J. (2008). Impacts of abstinence education on teen sexual activity, risk of pregnancy, and risk of sexually transmitted diseases. *Journal of Policy Analysis and Management, 27*(2), 255–276.

United States Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Washington, DC: Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Available from http://ies.ed.gov/ncee/pubs/evidence_based/evidence_based.asp

United States Institute of Education Sciences. (2008). *Education research grants CFDA number: 84.305A*. Retrieved June 17, 2008, from http://ies.ed.gov/funding/pdf/2009_84305A.pdf

United States Institute of Education Sciences. (n.d.-a). *Homepage*. Retrieved June 17, 2008, from http://ies.ed.gov/

United States Institute of Education Sciences. (n.d.-b). *What Works Clearinghouse: Homepage*. Retrieved June 20, 2008, from http://ies.ed.gov/ncee/wwc/

United States Institute of Education Sciences. (n.d.-c). *What Works Clearinghouse: Technical assistance*. Retrieved July 8, 2008, from http://ies.ed.gov/ncee/wwc/tech_assistance/

United States Institute of Education Sciences. (n.d.-d). *What Works Clearinghouse: Technical working papers*. Retrieved July 8, 2008, from http://ies.ed.gov/ncee/wwc/twp.asp

United States National Research Council. (2002). *Scientific research in education*. Committee on Scientific Principles for Education Research. R. J. Shavelson & L. Towne (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

W. T. Grant Foundation and University of Michigan (n.d.). *Building capacity to evaluate group-level interventions*. Retrieved July 21, 2008, from http://sitemaker.umich.edu/group-based/home

Wildavsky, A. (1979). *Speaking truth to power: The art and craft of policy analysis*. Boston: Little, Brown.

Ye, N. (Ed.). (2003). *The handbook of data mining*. Mahwah, NJ: Lawrence Erlbaum.

Chapter 23
Funding Patterns and Priorities: An International Perspective

Hsiao-Ching She, Larry D. Yore, John O. Anderson, Sibel Erduran, Wolfgang Gräber, Alister Jones, Johannes Klumpers, Stephen Parker, Marissa Rollnick, Robert D. Sherwood, and Bruce Waldrip

Research in literacy and science education converge onto *science literacy for all* found in many international reforms (Hand, Prain, & Yore, 2001) and the commonalities in targets (all students), goals (science literacy composed of fundamental literacy and understanding the big ideas in science), and pedagogy (constructivist approaches and authentic assessment) across English language arts and science (Ford, Yore, & Anthony, 1997; Yore, Pimm, & Tuan, 2007). Similar claims apply to mathematics literacy and technology literacy. This convergence and the international move to enhance research quality suggest potential relationships amongst research

H.-C. She
National Chiao Tung University

L.D. Yore
University of Victoria

J.O. Anderson
University of Victoria

S. Erduran
University of Bristol

W. Gräber
IPN Leibniz-Institute for Science Education

A. Jones
University of Waikato

J. Klumpers
European Commission

S. Parker
European Commission

M. Rollnick
University of the Witwatersand

R.D. Sherwood
Indiana University

B. Waldrip
University of Southern Queensland

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

policy, practices, and funding for literacy and science education. Furthermore, such connections should be growing in importance and fiscal priority for funding agencies. In the United States, explicit connections can be seen for research policy and preferred research practice in federal laws—but are there similar connections between policy and practice with research funding of literacy and science education research? Do such relationships exist in other countries?

Two recent policies in the United States illustrate the potential connection. The No Child Left Behind Act of 2001 (NCLB, 2002) reauthorized a number of federal programs aiming to improve the performance of primary and secondary schools by increasing the standards of accountability for states, school districts, and schools as well as providing parents more flexibility in choosing which schools their children will attend. The NCLB Act, which is open for renewal in 2008/09, requires states to develop assessments in language arts, mathematics, and science to be given to all students in certain grades—if those states are to receive federal funding for schools. The Education Sciences Reform Act of 2002 (ESRA, 2002) reauthorized and strengthened the principal education research, statistics, and evaluation activities of the Department of Education. This act funds the national data collection system that allows federal agencies to oversee the entire national education system and promotes a strong, scientifically rigorous research capacity within education. It is believed that such legislation is critically important to the successful implementation of the education reforms and to transform education into an evidence-based field, commonly called the Gold Standard.

Other countries have chosen and implemented various quality assurance approaches for research (see Coll et al., Chap. 6). These approaches are less explicit than the United States' federal law and policy; but they are nonetheless intended to influence the quality, foci, and approaches of educational research. Clearly, these policies and approaches have the potential for influencing funding patterns and priorities and for impacting literacy and science education research. Therefore, the focus of this chapter is to explore whether and how the funding policy and priorities (the *big stick*) would have, or has had, an impact on the quality, quantity, and direction of literacy and science education research funding in different jurisdictions. This chapter explores these potential connections and perspectives radiating out from, and contrasted with, the United States to other countries including Canada, the European Union, Germany, the United Kingdom, Australia, New Zealand, South Africa, and Taiwan (Republic of China). In the sections that follow, currencies are specified in the country's origin except for South Africa, which was converted to USD.

23.1 United States of America

There are two major funding agencies for science education research in the United States: the National Science Foundation (US NSF, n.d.-b) and the United States Department of Education (US ED, n.d.). Some funding does occur through other departments and agencies of the federal government—such as the Department of Energy, the National Aeronautics and Space Administration, and the National

Institutes of Health Office of Science Education (the interested reader may see their respective Web sites for more information: http://www.doe.gov/; http://www.nasa.gov/; http://science.education.nih.gov/)—but the dollar amounts are relatively small compared to the NSF and US ED budgets. While NSF and US ED are both federal government entities, they differ substantially in organizational structure and to some extent in goals and objectives.

The NSF is an independent agency in the federal government. While independent agency heads report to the Office of the President, there are only two members of NSF appointed directly by the President and confirmed by the Senate (director and deputy director). Therefore, the NSF has a reputation in Washington, DC, of being somewhat insulated from the political process. The policymaking and oversight organization is the 24-member National Science Board (US NSB, n.d.), whose members are also nominated by the President and confirmed by the Senate for 6-year terms. NSF's independence makes a difference in the direction that research funding can take.

The US ED has as its head the Secretary of Education, who is a member of the President's Cabinet. The Secretary is nominated by the President and confirmed by the Senate. As would be expected with any cabinet appointment, the Secretary is expected to be a proponent of the President's policies and priorities. The ESRA established the Institute of Education Sciences (US IES, n.d.-a) as the research agency within the US ED. The IES has a Director appointed by the Secretary of Education and a 15-member National Board for Education Sciences appointed by the Director as an advisory group. These serial appointments make this agency more closely tied to political directions in Washington, DC.

However, in the United States, individual states and local governments are the main authorities responsible for educational policy and practices. With these states' rights come the responsibilities for funding and regulating public education. But federal funds for K–12 education and related endeavors (hot-lunch programs, young Americans with disabilities, etc.) and postsecondary scholarships and research carry specified obligations set by the national government and accepted by local and state governments as part of the line item-funding. The passage of NCLB was a major departure from tradition and has not been without controversy. In recent years, some federal research and development (R&D) funding flowed directly to state, urban, and local education authorities through specific teacher-enhancement and systemic-change initiatives; but these governments do not provide significant funding for educational research.

23.1.1 Education and Human Resources Directorate of the National Science Foundation

The Education and Human Resources (EHR) Directorate is one of seven directorates of the NSF; it has major responsibility for science, technology, engineering, and mathematics (STEM) education funding. EHR (US NSF, n.d.-a) has as its goals:

1. Prepare the next generation of STEM professionals and attract and retain more Americans to STEM careers.
2. Develop a robust research community that can conduct rigorous research and evaluation that will support excellence in STEM education and that integrates research and education.
3. Increase the technological, scientific, and quantitative literacy of all Americans so that they can exercise responsible citizenship and live productive lives in an increasingly technological society.
4. Broaden participation (individuals, geographic regions, types of institutions, STEM disciplines) and close achievement gaps in all STEM fields. (Section "Goals")

There have been several changes in the EHR Directorate regarding its organization, priorities, and programs. In terms of science education research, a major change has been the consolidation in 2007 of the two divisions that funded the majority of educational research—Research Evaluation and Communication (REC) and Elementary, Secondary, and Informal Education (ESIE)—into the Division of Research on Learning in Formal and Informal Settings (DRL). This consolidation reduced the number of solicitations (calls for research proposals) coming out of the combined division. The programs of Instructional Materials Development, Teacher Professional Continuum, and the Centers for Teaching and Learning from the former ESIE division were merged into a new program: Discovery Research K-12 (DR-K12, NSF Solicitation 06-593). Programs in Evaluation Research and Capacity Building and in Research on Learning and Education from the former REC division were merged into the Research and Evaluation on Education in Science and Engineering (REESE, NSF Solicitation 07-595) at about the same time.

Analysis of the fiscal years FY2005 actual budget, FY2006 current plan, and FY2007 budget request illustrates that this consolidation actually resulted in a reduced amount of funding being requested for DRL in FY2007 ($160,000) compared to the two former divisions separately on top of a major reduction ($23,600,000) in the FY2006 (US NSF, 2006a, p. 215). A modest ($7.5 million, 3.5%) increase for DRL in the FY 2008 ($222.5 million) was requested (US NSF, 2007); however, Congress used the Consolidated Appropriations Act of 2008 (HR 2764) to fund a variety of agencies, including NSF, for FY2008. This act was not very specific in the amount of funds that should be allocated to individual programs within agencies, making comparisons to previous years more difficult. In the FY2009 budget request for NSF (US NSF, 2008), DRL is listed as having a FY2008 budget estimate of $214.0 million as compared to $208.99 million FY2007 actual budget ($5.01 million, 2.4%) increase—an amount lower than the request. The request to Congress for FY2008 is $226.5 ($12.5 million, 5.8%) increase as compared to the FY2008 budget estimate.

However, the consolidation of the two major divisions may have some positive outcomes. The current solicitations from DRL have developed a stronger set of priorities for programs within the new division. This has been summarized in the latest DR-K12 solicitation (US NSF, 2006c, p. 6) and illustrated in the DRL cycle of innovation and learning that promotes study (clarify, frame, operationalize, theorize, and basic research), design (develop, test, validate, and refine), implement (explore impact in context), evaluate (establish effectiveness and generalize), and synthesize (identify insights, questions, and set agenda). This clarification of what program will take the lead and the evolutionary nature of an R&D agenda may give researchers a better understanding of designing R&D activities, securing funding, and producing higher-quality results.

NSF solicitations, such as the DR-K12 and REESE, have maintained a position that the research methodology must match the problem space and questions under study. For example, in the current DR-K12 solicitation in the research design section:

> The types of claims the researchers hope to be able to make about the materials should be described, and the research design should be linked to the types of claims envisioned. Describe the research design and methodology … and explain why the research design is rigorous. (US NSF, 2006b, p. 14)

This would indicate that no particular research methodology is favored over another, but the writer of the proposal must make a strong case that the design chosen is rigorous. The merit review criteria for all NSF proposals, as established by the NSB, are intellectual merit and broader impacts. In solicitations, two standard paragraphs are usually included to give reviewers guidance in considering these rather broad criteria. In the NSF DR-K12 (08-502) solicitation, they are listed as:

- What is the intellectual merit of the proposed activity? How important is the proposed activity to advancing knowledge and understanding within its own field or across different fields? How well qualified is the proposer (individual or team) to conduct the project? (If appropriate, the reviewer will comment on the quality of the prior work.) To what extent does the proposed activity suggest and explore creative, original, or potentially transformative concepts? How well conceived and organized is the proposed activity? Is there sufficient access to resources?
- What are the broader impacts of the proposed activity? How well does the activity advance discovery and understanding while promoting teaching, training, and learning? How well does the proposed activity broaden the participation of underrepresented groups (e.g., gender, ethnicity, disability, geographic, etc.)? To what extent will it enhance the infrastructure for research and education, such as facilities, instrumentation, networks, and partnerships? Will the results be disseminated broadly to enhance scientific and technological understanding? What may be the benefits of the proposed activity to society? (p. 17)

While particular solicitations may have additional, specific evaluation considerations, these two items are the fundamental review criteria.

23.1.2 Institute of Education Sciences of the US Department of Education

The IES's (US IES, n.d.-b) stated mission and goals are:

> Its mission is to expand knowledge and provide information on the condition of education, practices that improve academic achievement, and the effectiveness of federal and

> other education programs. Its goal is the transformation of education into an evidence-based field in which decision makers routinely seek out the best available research and data before adopting programs or practices that will affect significant numbers of students. (para. 5)

The transformational aspect of the IES has been the most controversial component, and lively debate on the rigor of various research designs has been widely discussed in the literature and at professional meetings.

The most recent solicitation from the IES (Education Research Grants CFDA Number 84.305A, US IES, 2008) is a multidisciplinary solicitation, but it has five goals for all subprograms within the solicitation with the expectation that one or more goals will apply. These goals are:

> 1. Identify existing programs, practices, and policies that may have an impact on student outcomes and the factors that may mediate or moderate the effects of these programs, practices, and policies;
> 2. Develop programs, practices, and policies that are theoretically and empirically based;
> 3. Evaluate the efficacy of fully developed programs, practices, and policies;
> 4. Evaluate the impact of programs, practices, and policies implemented at scale; and
> 5. Develop and/or validate data and measurement systems and tools. (p. 9)

Researchers applying for funds to conduct studies related to these goals have available detailed instructions for issues related to the proposed research. For example, the research design section for goal 3 states: “Studies using randomized assignment to treatment and comparison conditions are strongly preferred” (p. 59), indicating a clear preference for a Gold Standard or randomized controlled trial (RCT) methodology. Although in the methodological requirements section of each goal the phrase “the proposed research design must be appropriate for answering the research questions or hypotheses that are posed” (p. 59) is found, the discussion after this section appears to encourage particular methods over others. This conditioned advice demonstrates a connection between the federal policies and potential funding while the “Review Criteria for Scientific Merit” (p. 78) specified significance (compelling rationale for the significance of the project), research plan (addresses methodological requirements for the target goal), personnel (principal investigator, project director, and other key personnel possess appropriate training, experience, and time to complete the proposed research), and resources (facilities, equipment, supplies, other resources and commitments for the implementation and success of the project).

STEM education research has never been an especially well-funded enterprise in the United States. The overall budget for the research aspects of IES has not grown substantially over the past 3 fiscal years. Moreover, the Research, Development, and Dissemination budget line has remained practically constant at $162.5 million (US ED, 2007). Furthermore, the total funds allocated to the new DRL division at NSF and to the research, development, and dissemination activities at the US ED are only approximately $378 million. This is an extremely small amount compared to the overall combined budget for NSF ($5.91 billion) and the Department of Education discretionary appropriations ($56 billion) of about $62 billion.

Sherwood and Hanson (2008) studied NSF funding allocations and found that only about 2.7% of funds allocated to the ESIE and REC divisions over a 10-year period had been used to fund teacher-education research studies. The majority of funds were used to fund programmatic activities, such as professional development and curricular development efforts. Smaller amounts were used for research studies of student learning in STEM areas and work on student assessments. There have been some promising collaborations between NSF and US ED, such as the Mathematics and Science Partnerships program. However, literacy and science education researchers—as with most educational researchers—will have to compete both among their peers and with other groups seeking support to receive funding for their work.

This brief discussion indicates differences between NSF and IES in their views of the Gold Standard as a major organizing frame for research funding. NSF has mainly limited its guidelines to *the-methods-must-be-appropriate-to-the-question* view; however, in many IES solicitation guidelines, the expectation of RCTs is much higher. The IES of the US ED appears to be more closely linked to the political process, while the EHR Directorate of NSF has greater degrees of freedom within its global mandate. Many of the research projects funded by both agencies since the establishment of the Gold Standard are in their early stages, and it will be of great interest to the literacy and science education communities to see what new knowledge is developed.

23.2 Canada

Canada’s population is approximately 10% that of the United States, but federal funding of $4.3 billion in 2007 on all forms of R&D is only a small fraction of the United States’ funding (CA Department of Finance, 2007, ch. 5). There are three major research funding agencies in Canada receiving about $1.7 billion, including indirect costs: Social Sciences and Humanities Research Council, Canada—$306 million (CA SSHRC, n.d.-b), Medical Research Council of Canada (MRC) and Canadian Institutes of Health Research—$700 million (CA CIHR, n.d.), and the Natural Sciences and Engineering Research Council of Canada—$600 million (CA NSERC, n.d.-a). The Prime Minister of Canada appoints the presidents of these funding agencies for 5-year terms, which is not directly connected to the term of the government. The budget and focus of these agencies reflect a political process and spending priorities of the ruling party, but the loyal opposition party does mediate major swings in both the amount and allocation of funds within the budget. Each agency has specific responsibility for funding curiosity-driven and mission-driven research for member groups: SSHRC includes professional programs; MRC/CIHR includes medicine, pharmacy, nursing, and other health care areas;

and NSERC includes sciences, technology, engineering, and mathematics areas. Collectively from time to time, these agencies identify and fund special R&D envelopes focused on government and society priorities, for example, the Tri-Agency Partnership on Knowledge Syntheses on the Environment.

SSHRC is the main national funding agency for research in education. Canada does not have a national ministry of education; the public education system is under the jurisdiction of the 13 provinces and territories, and with few exceptions, the provincial–territorial ministries do not provide funding for educational research on a systematic, continuous basis. One exception is the province of Quebec in which le Fonds de recherche sur la nature et les technologies du Québec (Fonds NATEQ) funds research in natural sciences and le Fonds québécois de la recherche sur la société et la culture (FQRSC) supports all research contributing to the understanding of people, communities, institutions, and cultures.

The selection procedures (CA SSHRC, n.d.-c) for federal funding of research projects do not contain any criteria related to methodological approach; the criteria are rather general in nature and focus on the record of the researcher making application for funding:

> Record of Research Achievement
>
> The record of research achievement refers to the tangible contributions made by the applicant(s) to the advancement, development and dissemination of knowledge in the social sciences and humanities. The focus of the evaluation is on the most recent six-year period of activity. In evaluating regular scholars, the committee will also take into account the five most significant research contributions (as identified by the applicant) from any period of an applicant's career. (para. 1)

The applicants and coapplicants must be members of the target research communities and hold university research appointments; there is no limit on the number of applications in which a researcher may participate as a coapplicant. The criteria for evaluating grant proposals involve the quality of the proposed research (40%) and the research achievements (60%) of applicants and coapplicants. Research quality is evaluated against a rubric composed of coherence, originality, and sophistication of the problem, research questions, outcomes, and approach. These evaluation criteria discount trendy methodologies. Research achievements involve assessing the quantity, quality, and impact of the applicants' scholarship. Committee chairs and evaluation panel members are given tremendous leeway in terms of their peer evaluations.

SSHRC does allow slight variations in the distribution of the criteria for new scholars, defined as an eligible applicant who has not yet had the opportunity to establish an extensive record of research achievement but is in the process of building one (i.e., less than 5 years since highest degree at application date, held tenure-track university appointment for less than 5 years not including nontenure appointments, or had career significantly interrupted or delayed for family reasons). Applicants adjudicated under the new scholar category have their research achievement and program of research weighted in the overall score such that either a 60:40 or 40:60 ratio will apply, whichever will produce the more favorable overall score.

Annual budget constraints require the committee, program officers, and agencies to manipulate the cutpoints between fundable proposals funded and fundable proposals not funded and the allocations of percentages of research budgets funded for proposals on the margins. Federal funding agencies have started to provide indirect expenses ($315 million in 2007) to universities (distributed proportionately to the institution's success rate) to support their research services and facilitate research cultures. These funds have been welcomed by universities to offset the costs of doing research; unlike the United States, however, they are far less than the actual costs to support and facilitate a research culture.

In looking for patterns in the research funding approved by SSHRC with specific reference to research in literacy and science education, the evidence is simply not available. Since 2003, the approval of research grants for education has been divided into two sections:

- Education 1: Science and mathematics education, educational psychology, reading and writing, special education, bilingual education, English as a second language (ESL), early childhood, physical education and other
- Education 2: Educational administration, adult education, curriculum, distance education, measurement, sociology of education, teacher education and philosophy

The applications, success rates, and funding totals for these two sections are provided in Table 23.1. One pattern that can be discerned is the relatively equivalent funding approved for each category, indicating that research into literacy and science education is certainly not privileged within the SSHRC envelope (Education 1 also contains the high-priority areas of educational psychology, special education, bilingual education, and ESL). Another trend, although not linear and with an exception in 2007, is the general increase in the number of applications and the moderate increase in total funds awarded. Comparison of grants approved and funding reveals that successful proposals receive average annual funding of about $30,000–35,000 for 3 years.

SSHRC programs include (a) Research Grants—Standard and Major Collaborative Research Initiatives, Strategic Research Grants, Community–University Research

Table 23.1 Social Sciences & Humanities Research Council, Canada: Standard research grants success rates and funding

		Projects			
Year	Program	Submitted	Approved	Success rate	Funding ($000 000)
2003	Education 1	129	53	41.1	5.1
	Education 2	103	42	40.1	4.7
2004	Education 1	143	61	42.7	6.6
	Education 2	142	61	43.1	6.6
2005	Education 1	147	59	40.1	6.4
	Education 2	147	58	39.5	5.9
2006	Education 1	170	68	40.0	7.3
	Education 2	158	64	40.5	6.8
2007	Education 1	125	43	34.4	4.0
	Education 2	116	32	27.6	2.9

Alliances, International Community–University Research Alliances—in Partnership with the International Development Research Centre, and (b) Strategic Joint Initiatives—Canadian Initiative on Social Statistics Data Training Schools, Access to Research Data Centres, etc. Occasionally, research envelopes with special or interagency funding are dedicated to specific or mission-driven areas related to literacy and science education; for example, the Canadian Council on Learning (CCL), Canadian Language and Literacy Research Network (CLLRNet), Canadian Centre for Research on Literacy (CCRL), Initiatives for the New Economy (INE), and Centres for Research in Youth Science Teaching and Learning (CRYSTAL).

The CCL (n.d.) was established with a 5-year, $85 million budget commitment by the outgoing Liberal government in 2002. Its funding and project foci are aboriginal learning, adult learning, early childhood learning, health and learning, and work and learning. Each CCL funding center relates to literacy and STEM education, and their standard and special calls for proposals reflect these priorities.

The CLLRNet (n.d.), a Networks for Centres of Excellence member (CA Networks of Centres of Excellence, n.d.), was formed by a group of leading Canadian researchers with a long-standing scientific interest in language and literacy who believe that Canada's competitiveness in the future depends on our children being able to communicate effectively. However, in spite of a positive external evaluation, funding for the next 5-year cycle has not been approved.

The CCRL (n.d.) was the first formally established academic body for research on literacy across the continuum from emergent to third-age literacy. CCRL promotes the consolidation of research interests and expertise to create a new, substantive, interdisciplinary (anthropological, audiological, educational, historical, legal, linguistic, literary, medical, philosophical, psychological, sociological, and speech pathology) research focus within the University of Alberta.

The Education Research Initiative, part of the larger INE (CA SSHRC, n.d.-a), was a strategic collaboration between SSHRC and the Canadian Educational Statistics Council, which contributed over $1 million to empirical research over a 3-year period. The program has not been continued. A focus of the research funded was the use of large-scale educational databases (e.g., the national assessment program in Canada: School Achievement Indicators Program of the Council of Ministers of Education Canada) to identify student and school characteristics related to educational achievement (literacy, mathematics, etc.) in order to enhance educational performance and grow the knowledge economy.

CRYSTAL (CA NSERC, 2006), a $5 million pilot program (2005–2010) funded by NSERC, fosters research in science and mathematics education based on the widespread and growing recognition that improving science literacy and numeracy among youth will help to (a) increase the supply of students qualified for and interested in science, mathematics, and engineering careers and (b) improve the economy. These interuniversity and interdisciplinary centers are composed of one or more faculties of education, science, and engineering, local partners, and schools. Partners were recruited from user communities and focused on public awareness of science, informal learning contexts, teachers, school administrators, and provincial ministries of education. Five regional centers were funded with two emphasizing literacy, mathematics, and science: CRYSTAL Alberta (Alberta CRYSTAL Project, n.d.) and CRYSTAL Pacific (Pacific CRYSTAL Project, n.d.)

Over recent years, funding for research in Canada has increased slightly. Furthermore, federal funding agencies have recognized the cost of building, facilitating, and supporting research cultures within universities by providing funding to offset the indirect costs of research. Within SSHRC and NSERC, and somewhat within CIHR, there has been explicit recognition and support of literacy and science education research. Continued and expanded funding for literacy and science education are tentative, but success in influencing practice and policy will do much to ensure continued support. Recently, NSERC announced the formal recognition of science and engineering outreach–public awareness efforts and has encouraged vice presidents of research and deans of engineering and science to do the same within the university reward systems.

23.3 European Union

The European Parliament, consisting of a growing number of member states, has explored a variety of strategic investments in research, knowledge translation, and other activities to increase public awareness of science and technology, career opportunities, and socioeconomic goals. Many of these activities differ from those used in North America, and some predate the formation of the European Union (EU) and involved nonmember states. Within the context of the activities implemented by the Directorate General for Research, support for science education can be traced back to the launch of the European Contest for Young Scientists in 1989. This popular and highly successful annual event brings together winners of national preuniversity school science competitions from member and nonmember states (countries) that compete for prizes and other awards. Many of the winners go on to pursue highly successful careers in science. The contest, and the nationally funded competitions that underpin it, motivates schools and science teachers in participating countries. Nonetheless, as with any competition, the participants are already highly motivated; they are not truly representative of the student population or the general state of science teaching in Europe.

The Commissioner in 1993, Professor Ruberti, inaugurated the *European Week for Scientific and Technological Culture* that aimed to showcase the best of European science and technology and to demonstrate how science and technology affect citizens' daily lives. The emphasis was placed on fostering a better understanding of what science is and how it is used. Many of the funded activities operated at the interface between informal and formal science education and, therefore, targeted young people at school—but this was rather incidental and not explicitly linked to enhancing the European research base.

The launch of the Fourth Framework Programme (FP4) of the European Community for Research, Technological Development, and Demonstration Activities (1994–1998) saw the inclusion of a new specific program on socioeconomic sciences that covered explicitly, inter alia, research on education and training. Only 2 of the 38 projects and networks supported dealt with science education. The situation was not better under the Fifth Framework Programme (FP5), with none of the 20 activities funded targeting science. However, FP5 did include an action-based topic for raising public awareness of science and technology by promoting dialogue, highlighting the role of the media and science communication, and funding for the "European Science Week" (as it was renamed).

The EU can contribute toward the development of quality education by encouraging cooperation between member states, but the responsibility for content of teaching and the organization of education systems resides with individual member states. This restricts EU financial support to promoting networking and the demonstration of best practice and research through collaborative projects with durations of 2 to 3 years. The European Commission (EC) also supports education in a more general sense through the activities of the Directorate General for Education and Culture and in particular networking between teachers; however, research actions are not funded. Over the course of FP5, the budget allocated for raising public awareness of science and technology activities was increased from 12 million euros (€) to €16.25 million in response to the demand for funding, which doubled annually over the 4 years of FP5. By the end of the program, 54 projects and networks received support of which 22 explicitly targeted young people and their teachers through informal science education-based activities.

Halfway through FP5, the 2000 Lisbon Summit set ambitious targets for sustainable socioeconomic growth in Europe and recognized that these targets could only be achieved through the emergence of a truly knowledge-based society. The targets were linked to specific objectives—including one that highlighted the need to increase the number of educated and employed researchers and science professionals in Europe to levels comparable to those in the United States and Japan. This, in turn, focused attention on activities to make school science more attractive to young people both in terms of its socioeconomic relevance and the analytical skills that learning science confers: skills that are essential to a functional, knowledge-based society. It was also recognized, however, that too many students were turning away from science because curricula were too content-heavy and unappealing.

The FP6 (2002–2006) reflected the recommendations of the Lisbon Summit and offered explicit support for the development and testing of new methods to stimulate young people's interest in science and to promote the dissemination of these results and best teaching practice. It should be noted, however, that there was no explicit education research theme although the role of education was covered in related social science research issues. Over 20 relatively large projects sharing a budget of €18 million (excluding the EU Young Scientist Contest and European Science Week) were funded. Initially, the emphasis was placed on supporting a cluster of projects in order to combat fragmentation. Subsequently, following the findings of an expert group, 16 projects were funded: hands-on science teaching and other methodologies, young people from disadvantaged groups, reinforcing the links between schools and universities, and understanding the differences between girls' and boys' perceptions of science and technology.

The FP7 (EC, 2006) is even more explicit in this respect, calling for the "Creation of an open environment which triggers curiosity for science in children and young people, by reinforcing science education at all levels, including in schools, and promoting interest and full participation in science among people from all backgrounds" (p. 35). This objective is articulated by three key issues within the Science in Society part of the Specific Programme Capacities:

- Supporting formal and informal science education in schools as well as through science centres and museums and other relevant means.
- Reinforcing links between science education and science careers.
- Research and coordination actions on new methods in science education. (EC, 2007b, pp. 28–29)

Mindful of the need to exploit more effectively experience being developed across Europe and to concentrate on actions most likely to have the greatest impact, in 2006 the European Commissioners of Research and of Education and Culture tasked a group of experts to examine a cross section of ongoing initiatives in the field of science education at national and European levels. The objective was to draw from the elements of know-how and good practice that could bring about a radical change in young people's interest in science studies and that could be scaled up at the European level. The working group's report, *Science Education Now: A Renewed Pedagogy for the Future of Europe* (Rocard et al., 2007), made a number of recommendations including the need to promote more widely inquiry-oriented and problem-based science education methodologies in primary and secondary schools and to support teachers' networks.

The 2007 Science in Society call for proposals in the area of science education went some way toward addressing these issues. The following topics were open to action-based proposals: links between science education and research (to promote a better correlation between how science is taught, learnt, and done), teaching methods (with a reference to inquiry-based science education), and images of science (where diversity in terms of student background and career opportunities was stressed). Four project-proposals covering these areas have recently been awarded grant support including a large-scale project. The recently published 2008 Science in Society work program (EC, 2007b) is even more explicit in this area, with the call for proposals on the "dissemination and use of inquiry based teaching methods on a large scale in Europe" (p. 29) as well as a notice of the Commission's intention to establish an Internet-based information platform for the dissemination of information and best practice regarding teaching methods. A more speculative topic, open to research projects, is to see whether the introduction of multidisciplinary topics of current interests (such as forensic science) motivates students to take greater interest in core disciplines and career options. The results of this call are currently being evaluated.

The strategy currently being pursued by the Directorate General for Research is to concentrate on supporting European collaborative activities in the area of science education where the methodologies employed have already been successfully demonstrated in a local setting. It is recognized that best practice is intrinsically context-dependent; thus, specific issues will be opened to research projects in the course of the current framework program that promotes scaling, capacity building, and knowledge translation. Nonetheless, impact at the European level will be compromised if there is no learning with regards to the experience being gained. It is hoped that the establishment of an Internet-based information platform, not specifically linked to a particular methodology or project, will create a very real opportunity for teachers and other stakeholders to benefit, Europe-wide, from collective experience.

Over the last 10 years, as part of the European Information and Communication Technologies (ICT) research, support to education research and especially science education research has developed. Under the heading of e-learning or technology-enhanced learning, the EC has funded research investigating how ICT can be used to support learning–teaching and competence development throughout life. Although this was not necessarily requested in the programs, respondents to these calls were focusing on science frequently to investigate the uses of ICT. In 2007 and 2008, one program objective was to contribute to the development of adaptive and intuitive learning systems, resulting from longer-term research efforts that are able to configure themselves according to the understanding and experience of learners' behavior. These systems will identify learner's requirements, make best use of the individual learning and cognitive abilities of the learner, and give meaningful advice and feedback to both learners and teachers.

23.4 Germany

The Federal Republic of Germany (FDR) was formed in 1949 and after joined in 1990 by five states from the former German Democratic Republic (GDS) consists of 16 Länders (states). The FDR is a founding member state and the largest country in the EU in terms of population with 82 million inhabitants situated in the heart of the EU, having common borders with 9 of the current 27 member states. The FDR has actively supported participation in a unified Europe, the development of the Common Market, the introduction of the strong euro, and the enlargement of member states; it contributes about 20% of the operating budget for the EU; and it is actively promoting the realization of the mission the 2000 Lisbon Summit in becoming a genuine European Research Area (ERA).

Germany has a rich tradition of scientific and technological research that extends to the social sciences and education. Science education research reflects a strong historic association with the academic disciplines of biology, chemistry, and physics and places much R&D responsibilities for curriculum and pedagogy on classroom teachers as experts with both content and pedagogical backgrounds. Due to the federal system of government, the research funding system, particularly in the field of education, is highly complex. The universities, funded by their state government, traditionally form the backbone of the German research system; they are involved in basic and applied research as well as in development. A distinctive feature of the research landscape is the existence of a rich variety of out-of-university research institutes dating back to formation of the Kaiser Wilhelm Gesellschaft (the predecessor of the Max Planck Gesellschaft [MPG]) in 1911. Today, these research centers are composed of many focused institutes without teaching responsibilities and are each managed by an excellent researcher. The Max Planck Society (MPG, basic research), the Fraunhofer Society (applied research), the Leibniz Association (WGL, interdisciplinary, combining basic and applied research), the Helmholtz Association (national centers of high-budget research), and the German Research Foundation (DFG, central self-governed organization of German science) are jointly funded by the federal and state governments (GE Federal Ministry of Education and Research [BMBF], 2007b).

Institutes embedded within the large research centers conduct education research. The MPG is composed of 77 institutes, research places, laboratories, and work groups and has one institute designated to education research: the Max Planck Institute für Bildungsforschung [MPI for Human Development] in Berlin. The Fraunhofer Society has 58 research institutions doing applied research in close connection to industry. The Helmholtz Association is the largest science association with 15 national centers and has a yearly budget of €2 billion. The Wissenschaftsgemeinschaft Leibniz (GE WGL, n.d.) serves as the umbrella organization for 82 institutions and has a total budget of €1.1 billion. These institutions have special sections for humanities and education research: Deutsches Institut für Erwachsenenbildung (GE DIE, [Institute for Adult Education], n.d.) in Bonn, Deutsches Institut für Internationale Pädagogische Forschung (GE DIPF [Institute for International Educational Research], n.d.) in Frankfurt am Main, Leibniz-Institut für die Pädagogik der Naturwissenschaften an der Universität Kiel (GE IPN [Leibniz Institute for Science Education], n.d.) in Kiel, and Institut für Wissensmedien (GE IWM [Knowledge Media Research Center], n.d.) in Tübingen.

These institutions enjoy long-term funding arrangements for doing research in a specific field, but there is also project-specific funding available from the federal and state governments. Individual researchers can apply to the institutes for funding research related to the institute's central focus. Project-based funding is mainly through direct grants from the BMBF, indirectly through the DFG. The BMBF normally defines programmatic funding envelopes (mission-driven focus) that researchers can write proposals within their expertise applied to the defined topics seeking funding. The DFG is funded by the federal (58%) and the state (42%) governments with a current budget of €1.3 billion. Historically, proposals were submitted from single researchers; today, there are much more coordinated and collaborative submissions from groups of researchers, graduate colleges, and special research areas.

The DFG promotes high-quality research in all fields of sciences and humanities and the promotion of young researchers with interdisciplinarity and internationality as key elements. Qualified researchers from all disciplines working at German research institutions are eligible to apply for research grants, which can be used primarily to fund staff, scientific instrumentation, consumables, and travel. Each proposal is subject to a complex peer-review process. First, two or three peer-reviewers, recognized experts in their field, assess the quality of the proposal and give an objective appraisal. Proposals concerning science education research are then assigned to the section of education sciences—teaching–learning process and qualification process—not to natural or physical sciences as in former years. The review board, consisting of elected members of the scientific community, evaluate the appraisals by checking the reviews assessing quality, verifying the process of selecting reviewers, comparing the proposal to other proposals, and providing a recommendation to the DFG Senate committee and the grant committee. This sequential assessment and decision-making process attempts to ensure that funding decisions are made in a fair, transparent manner. Criteria for evaluation are the quality of the proposal, the competence of the applicants, institutional opportunities to realize the project, goals, design and work packages, and the proposed use of funds. There is no fixed proportion among these criteria, weighting is referring to each individual project, and no specified research approach is required.

Germany has taken the results of research and evaluation projects seriously and utilizes these results to shape federal and state education policies. The large-scale international studies—Third International Mathematics and Science Study (TIMMS) and Programme for International Student Assessment (PISA)—created a starting point for enormous changes in Germany's education landscape. Compared to international competitors, German students did not achieve a top rating, as expected by the majority of the population, but achieved only an average rating. TIMSS, PISA, and the domestic benchmarking study PISA-E have revealed central deficits in the education system.

> German pupils show below-average performance in central areas, such as reading, mathematics and the sciences. In no other industrialized country is the social background so decisive for success in school and for education opportunities as in Germany. At the same time, the integration of children and young people with a migration background is clearly less successful.
>
> We need a change in the orientation of our education policy. Our school system must lead to a higher performance level and must enable more children and young people to earn higher education qualifications. In schools, the strengths and individual abilities and background of each child must be in the centre. The principle of challenging and supporting must be followed consistently.
>
> The competition for future opportunities for Germany has essentially become an international competition for the quality of education systems. An education reform therefore requires a national effort of all stakeholders and a broad debate in society across ideological barriers. (GE BMBF, n.d., paras. 2–4)

In 1997, former German President Roman Herzog wanted to raise the significance of education and called for public debates on the education system. He complained that education had only been a topic in experts' circles, creating no more than stagnancy. Contrary to former research studies, the TIMSS attracted more attention among politicians, journalists, and the public, which was caused or at least supported by the discussion after the reunification concerning school reform. Greater efforts in three areas were identified to help overcome *TIMSS-shock* (Klieme & Baumert, 2001): (a) development of school-based reforms to improve classroom teaching, (b) enhancing the empirical research in education, and (c) establishing a school system monitoring approach that switched from input- to output-oriented quality management. The whole process of reform and development in

Germany is not only to be seen in its complex, 16-state, federation system but also in the framework of the EU. Research activities at both EU and national levels must be better integrated and coordinated to become as efficient, attractive, and innovative as anywhere in the world. At the 2002 European Summit in Barcelona, heads of state called for an increase in the proportion of European GDP invested in research from 1.9% to 3% to contribute to reaching this goal.

> The EU and Member States have fully recognised that, together with high-quality education and lifelong learning and a supportive environment for innovation, ERA is essential to making Europe a leading knowledge society and thus creating the conditions for long-term prosperity. The ERA concept encompasses three inter-related aspects: a European 'internal market' for research, where researchers, technology and knowledge can freely circulate; effective European-level coordination of national and regional research activities, programmes and policies; and initiatives implemented and funded at European level. (EC, 2007a, p. 5)

Furthermore, at the1999 Bologna Conference, Germany and its European neighbors set out to establish a common European university system by the year 2010 (EC, n.d.). This reform has resulted in the transformation of bachelor's and master's degrees and the introduction of consistent credits recognized throughout Europe. The Bologna process plays an important role in establishing the ERA, helping graduates to enter the European labor market, and allowing European research careers.

The Bund-Länder Commission for Educational Planning and Research Promotion (BLK), a commission with members of the federal and all states' ministries, discussed common research funding and provided recommendations to the governments. The BLK, the federal government, and the Länders agreed upon a yearly increase of support to research institutions to reach the ambitious Lisbon goals. The disappointing 1997 TIMSS results encouraged the BLK Project Group "Innovation in Education" to form a commission focused on "Enhancing the Efficiency of Mathematics and Science Teaching (SINUS)" that was elaborated by experts from science education and pedagogy (Baumert et al., 1997). This expertise formed the base for a nationwide program with 180 schools from 15 states participating. In each case, a set of six schools, one pilot center, and five network schools worked together and produced modules to improve their mathematics and science teaching. This bottom-up program, coordinated by the IPN in cooperation with the State Institute of School Education and Educational Research in Munich and Bayreuth University, was a success. It was highly accepted by teachers and teacher educators and also recognized as an exemplar approach in *Science Education Now* (Rocard et al., 2007). After a 5-year pilot phase in 2003, the BLK launched a SINUS-Transfer program with 13 states and 734 schools participating (€10 million for the first 2 years). A second transfer (2005–2007) has reached 1,800 schools and a complete nationwide implementation starting in August 2007.

The BMBF, the state governments, and the DFG have funded several highly regarded collaborative science education projects. Some of these projects have traditional discipline-specific focus while others focus on interdisciplinary areas, informal learning contexts, and school accountability.

Chemistry in Context (GE Chemie im Kontext [CHiK], n.d.), started in the late 1990s by researchers from the universities in Oldenburg and Dortmund and the IPN, was inspired by the Salters Approach (United Kingdom) and ChemCom (United States) to develop units for a new chemistry curriculum based on theories of scientific literacy, motivation, and situated learning (Nentwig, Demuth, Parchmann, Gräsel, & Ralle, 2007). The CHiK curriculum follows a context-based approach. Rather than using the structure of the discipline, it generates basic chemical concepts from issues relevant to the learners, starting with the learners' ideas and questions. The teaching methodology of the course builds strongly on self-directed and cooperative forms of learning activities in the classroom. Currently, a large-scale implementation of CHiK is underway with working groups of teachers in 14 states organized by the universities in Oldenburg, Dortmund, and Wuppertal and the IPN.

Physics in Context (GE Physik im Kontext [piko], n.d.) attempts to advance students' scientific literacy (2004–2007). New instructional approaches are designed to enhance students' openness toward physical and technical problems around them. Teams of experienced teachers and educational scientists develop context-based instruction, opening new perspectives for physics education. The new approaches are tested in school practice, evaluated, and disseminated. piko is coordinated by the IPN, in cooperation with the University of Kassel, the University of Education in Ludwigsburg, Humboldt University in Berlin, and the University of Paderborn.

Biology in Context (GE Biologie im Kontext [BiK], n.d.) started in 2005 as a collaborative project among the IPN and the universities of Duisburg-Essen, Gießen, Göttingen, Münster, and Oldenburg. BiK aims to develop innovative biology teaching approaches and creative tasks, which are implemented, evaluated, and disseminated nationwide.

Learning Location Laboratory (GE Lernort Labor [LeLa], n.d.) was designed to promote cooperation among schools, universities, and research institutes. LeLa is a center for advice and quality development for extracurricular activities in the integrated mathematical–scientific–technological field and for the support of a network of teachers and long-last use to supplement school programs. These offerings are committed to the common aim of conveying students an authentic picture of science and the working world in the form of practical activities and hands-on projects.

Bildungsqualität von Schule [BIQUA, Quality of Education in Schools] is a unique, interdisciplinary, priority program launched with the participation of more than 20 expert groups of different universities and research institutes, coordinated by the IPN. BIQUA (2000–2005) aimed to identify parameters that have a significant impact on educational processes and to deduce recommendations for enhancing the quality of schooling based on these experiences (Prenzel, 2007). Another example of the new wave in science education research is the interdisciplinary Research Group Teaching and Learning of Science that investigated school education and instruction in chemistry, physics, and biology by different aspects, which interlink to a common perspective and develop models of instruction connected with a graduate school at the University of Essen-Duisburg (University of Essen-Duisburg, n.d.).

The German states, through the Konstanzer Beschluss [Konstanz Resolution], agreed in 1997 upon continuous, comparing measurements for students' achievements as a starting point for system monitoring (accountability). The monitoring involves further PISA studies and studies concerning knowledge of English language, politics, reading, etc. The most radical change to the German school system might be the introduction of national education standards in 2003. The international comparative studies seem to indicate that students from countries with a systematic quality management and output monitoring achieve better results. Klieme et al. (2007) developed an expertise (expert report, analysis of needs, and recommendation for future work) showing how and why the current input-oriented steering mechanism must be changed to an output-oriented control model. They believed that in a federal system this nationwide change would not only improve schools but would also standardize requirements and facilitate comparisons and changes among different school types and states. There are standards developed for German and Mathematics (Grade 4); German, Mathematics, and the first foreign language (Grades 9 and 10); and Biology, Chemistry, and Physics (Grade 10). The states started implementing these standards in 2004/05, but their success will require reliable output data. Thus, the BMBF established a Supporting Program for Educational Research, which has two foci: measures for structural strengthening and support of special research topics (GE BMBF, 2007a). This program should improve quality management in educational research and promotion of young scientists by enhancing international exchange and improving the communication and publication infrastructure.

The EU member states agreed to enhance their public expenditure for education to 3% of their GDP. Germany is currently spending 2.5% (€56 billion) and continues with great efforts to reach the ambitious goal. The Federal Minister of Education and Research (Schavan, n.d.) stated:

> The Federal Government is investing more money in research and development than ever before. An additional six billion Euro have been earmarked for R&D until 2009. The BMBF is coordinating this initiative and will invest its share of 4 billion Euro in excellent research and emerging cutting-edge technologies. I count on the Länder and on industry to also increase their investment in research. We can only achieve the 3-percent objective when working together. (para. 1)

Clear trends are apparent in Germany for increased funding of education research, connections between policy decisions and funding pattern, and moves toward exploring interdisciplinary and informal learning contexts, large-scale mission-driven programs, and the need for data on which to base evaluations, policies, and education decisions.

23.5 United Kingdom

There have been numerous changes in funding and evaluation of research in the United Kingdom. Demands for university accountability first began under the conservative Thatcher government (Lucas, 2006), and governments of the late 1990s and early 2000s have continued many of the policies. Currently, there are seven Research Councils (RC); the Economic and Social Research Council (ESRC) funds research in the social sciences including education (UK RC, n.d.). Each year the councils invest around 2.8 billion pounds (£) in research covering the full spectrum of academic disciplines from the medical and biological sciences to astronomy, physics, chemistry, and engineering; social sciences, economics, environmental sciences, and the arts and humanities (UK Higher Education Funding Council for England, n.d.). The councils invest around £1.3 billion in research of universities, around £500 million in their own research institutes, and around £300 million in access to international facilities for UK researchers. The ESRC has a planned total expenditure in 2007/08 of £181 million for social sciences (UK RC).

The University Grants Committee has assessed research since 1981 by surveying all subject areas in universities: student numbers, resources, balance of subjects, and the quality of individual institutions. The results of this survey formed the basis of *A Strategy for Higher Education into the 1990s* (Great Britain University Grants Committee, 1984). Subsequently, the Research Assessment Exercise was introduced (see Coll et al., Chap. 6) to assess the quality of university research based on peer-judgments and to inform the levels of funding for research distributed to each university as part of its Quality Research budget. The ESRC (UK ESRC, n.d.) states:

> [A]wards ranging from £15,000 to £1.5 million (100% FEC) to eligible institutions allowing individuals or research teams to undertake anything from a small project to a large-scale survey. The choice of research topic is yours provided it falls within the ESRC's remit. Your research proposal need not be relevant to our Strategic Framework. (para. 1)

Individual funding for proposed research studies and applicants are evaluated against four criteria: scientific quality, timeliness, track record of applicants, and value for money. No mention is made of a preferred or specific research methodology in the criteria. The timeliness criterion manifests itself in terms of political importance, which is captured though the policy landscape of national priorities and pressing curricular issues related to research—what gets funded is often within the policy parameters established by government. Frequently, these priorities manifest themselves as mission-driven funding envelopes focused on a sociopolitical or sociocultural agenda and science curriculum and teacher-enhancement projects.

23.5.1 National Priorities in Science Education

The Science, Technology, Engineering, and Mathematics program of the Department for Children, Schools, and Families (DCSF) aims to rationalize and improve the provision of support for students. A 2006 DCSF report focused on how best to (a) support STEM programs (primary to university levels) and (b) streamline the numerous STEM initiatives and implement them more effectively in every learning organization.

> The government wants to increase students' STEM skills in order to:
>
> - provide employers with the skills they need in their workforce;
> - help to maintain the UK's global competitiveness; and

- make the UK a world-leader in science-based research and development. (UK DCSF, n.d., para. 4)

The government published a 10-year investment framework for science and innovation alongside the 2004 Spending Review. It had consulted extensively with key stakeholders in developing this investment framework, including the scientific community, businesses, charities, regional and devolved bodies, and international contacts and received valuable input from individuals and organizations. The *Science & Innovation Investment Framework 2004–2014* sets out the government's ambitions for the next decade, in particular its contribution to economic growth and public services, and the attributes and funding arrangements of a research system capable of delivering these priorities (UK HM Treasury, 2004).

23.5.2 Developments in the Science Curriculum: Scientific Inquiry and Literacy

Through the 1980s, a series of curriculum development projects were funded across the United Kingdom by local education authorities and various educational publishers (Baggott la Velle & Erduran, 2007). These new initiatives occurred in the wake of the 1960s curricula funded by the private Nuffield Foundation when a series of process-based schemes with supporting materials arose (Newton & Gott, 1989). One example of this, widely taken up across the country in the late 1980s, was *Science in Process* (Inner London Educational Authority, 1987). Here the emphasis was on practicing the skills of the scientific process, such as observing, hypothesizing, recording, and reporting.

The 1988 National Curriculum (NC) in Science represented the first attempt to standardize the provision of science education across the country (UK Department of Education and Science, 1988). There were two overarching principles: science 5–16 and a content-based, balanced science curriculum. Primary school children (ages 5–11) had an entitlement for the first time to education in science; and all compulsory school-aged children had to study biology, chemistry, and physics in approximately equal proportions.

Following a consultation (UK NC Council, 1991), the NC was revised and the number of attainment targets was dramatically reduced. The 1995 Science NC (UK Department for Education, 1995) and the 1999 Science NC (UK Department for Education and Employment, 1999) outlined a major new area of the curriculum—Experimental and Investigative Science—commonly known as Sc1. The 2005 General Certificate of Secondary Education (GCSE, 14–16) examination specifications represented a new model underlying the philosophy of science and how science works presented in the 2006 Science NC that includes scientific literacy as a key focus (UK Department for Education and Skills, 2006).

These national priorities and pressing curriculum issues should influence literacy and science educators in terms of curriculum development, assessment approaches, and professional development efforts. Clear indications of funding priorities for related R&D activities are not yet apparent.

23.6 Australia

Australia’s population is about 6% of the United States, and the Australian governments’ funding of research is much less than the United States’. Education is funded both by the federal and state governments. The state governments give more emphasis to public schools, but they do provide some funding for private schools. The federal government funds initiatives to all schools and provides greater per capita funding to private schools, but it does fund educational initiatives to achieve other national socioeconomic priorities including occasional literacy or science education initiatives.

The Australian federal budget supports three major funding research agencies: the Australian Research Council (ARC), the Australian Learning and Teaching Council (ALTC), and the National Health and Medical Research Council (NHMRC). The NHMRC (AU NHMRC, n.d.) provides funding ($529 million in 2008) for all areas of research relevant to human health and medical research but does not specifically fund health literacy, public awareness, and outreach activities.

The political process influences the budget and focus of these agencies and spending priorities of the governing party, but none of these agencies has a specific mission for science and literacy education research. The government’s support for all research is less than 1% of the federal budget (0.8%), with education research receiving just over 0.1% of the limited funds allocated to research. The average education research grant is less than half that received by the sciences, with a typical education grant being about $140,000 over 3 years. The highest level of funding for an educational project is about $500,000 for the 3 years. These grants do not cover items such as researchers’ salaries, medical insurance, or routine equipment (computers, office furniture, etc.). These expenses are assumed under the general university operating grant provided by the federal government, where the size of the operating grant is affected by previous grant success of applicants from each university.

23.6.1 Federally Funded Research Activities

The ARC’s annual budget is about $626 million, including operating costs for the organization. ARC has two major grant programs—Discovery Grants and Linkage Grants—and specialized research centers and initiatives. The ARC supports research from a broad range of academic disciplines and professions but does not fund medical research. The criteria for these programs are similar; the major differences are in the weightings of the selection criteria.

The Discovery grants seek to develop “A strong capability in fundamental research (sometimes called discovery, basic or blue sky research) [that] will result in the development of new ideas, the creation of jobs, economic growth and an enhanced quality of life in Australia” (ARC, n.d.-a, para. 2). The program’s objectives are to (a) support excellent fundamental research by individuals and teams, (b) assist researchers to undertake their research in conducive conditions, (c) encourage research training in high-quality environments, and (d) develop and maintain a broad foundation of high-quality, world-class research. The success rate of education research proposals submitted to this program in terms of number of grants and amount awarded is well less than 1% (0.16%). For 2008, 102 applications were submitted as educational projects; but it is not possible to determine what percentage dealt with preschool, elementary, secondary, or postsecondary education. Included in the number of proposals and successful grants are projects that are classified as educational but do not deal with school or tertiary education. Typically, just over 20 education-focused grants are awarded each year (~20% success rate) with two or three science education grants awarded and a similar number of literacy education grants.

The criteria for evaluating grant proposals involve the track record and capacity of the team (40%), significance and innovation of the proposed research (30%), description of the proposed approach (20%), and the anticipated national benefit of the proposed research (10%). The applicant must be employed by an approved organization, which is usually a university or government research agency. Early-career applicants are reserved a set proportion of the successful applications. Additional grants are available for indigenous researchers and also fellowships to attract high-profile citizens back to Australian universities (~$9 million). Relatively few of these grants deal with educational issues. The selection criteria do not stipulate any preference for research methodology. ARC research proposal evaluations are very rigorous, stress world-class initiatives, and involve international evaluators from the research area involved. An applicant is restricted to a total of two Discovery grants, including current applications, at any particular time.

The Linkage grants include Linkage Infrastructure, Equipment, and Facilities (LIEF), Linkage International (LI), Linkage Learned Academies Special Projects (LASP), and Linkage Projects (LP).

> The ARC's funding schemes aim to encourage and extend cooperative approaches to research and improve the use of research outcomes by strengthening links within Australia's innovation system and with innovation systems internationally … [such as] partnerships between researchers and business, industry, community organisations and other publicly funded research agencies. (ARC, n.d.-b, paras. 1–2)

The LIEF grants are for 1 year only ($33 million in 2008) and are to encourage collaborative arrangements within and between eligible organizations; support the development and sharing of infrastructures, equipment, and facilities; enhance strengths, and ensure access of high-quality researchers to these resources (ARC, n.d.-c). The selection criteria involve the applicants' track record (20%), significance of research (20%), perceived need to access the proposed infrastructure, equipment, or facility (30%), and strength and benefits of the collaboration (30%). No LIEF grants were awarded for education research projects for the past 5 years.

The LI scheme (~$4 million) provides two types of support (ARC, n.d.-d): International Fellowships, available for outstanding postdoctoral research fellows and senior research fellows to work in Australian or overseas organizations for periods of up to 12 months, and Internationally Coordinated Initiatives, jointly funded by ARC and one or more research funding agencies overseas to research partners in international collaborative projects. These grants (a) support the movement of researchers between eligible Australian research organizations and centers of research excellence overseas and (b) foster collaboration and networking between Australia-based and overseas researchers. Educational researchers received two LI grants in 2008 (~$150,000), but neither grant was in literacy or science education.

The LASP funding envelope (~$550,000) seeks to (a) support programs of research or programs that support research undertaken by one or more of the learned academies that capitalize on their unique capabilities and (b) assist programs of research undertaken by institutions (ARC, n.d.-e). They are expected to have broad benefit for research and scholarship in the natural and applied sciences, technological development, applied technology, social sciences, and humanities. No educational researcher or project received support from LASP in 2008.

The LP scheme (~$220 million) supports collaborative R&D projects between higher education organizations and other organizations, including industry (ARC, n.d.-f). Linkage projects aim to encourage long-term strategic alliances, support collaborative research, foster international postdoctoral opportunities, provide industry-oriented research training, and establish a pool of world-class researchers to meet the needs of industry. Proposals for funding must involve a collaborating organization from outside the higher education sector. The selection criteria involve judgment of investigator's track record (20%), significance and innovation (25%), approach and training (20%), national benefit (10%), and commitment from partner organization (25%). Early-career faculty and staff members are treated similarly to those who apply for Discovery grants. Researchers can be involved in a combined total of five applications and projects. In 2007, 19 education projects were funded; only 1 was in literacy education, and none was in science education.

The ALTC (formerly the Carrick Institute; $825 million in 2007/08) has a mission to enhance learning and teaching in Australian higher education. These programs are designed to enhance the quality of teaching and learning, facilitate innovation and collaboration, and disseminate best practice. The grants under this program are competitive and may include a research or evaluation component, but they are not research grants per se (ALTC, n.d.). The evaluation criteria for grants reflect compliance with the council's mission, objectives, and values; transparency, value of research for the money requested, level of potential impact, and future implications and applications.

23.6.2 Private Research Activities

The Australian Council for Educational Research (ACER, n.d.) is a private educational research company based in Melbourne that does not receive any direct federal or state support. ACER generates about $50 million income annually through contract R&D projects and through delivery of products and services, such as the Programme for International Student Assessment (PISA) contract with the Organisation for Economic Co-operation and Development (see Anderson, Milford, & Ross, Chap. 13).

23.6.3 Literacy and Science Education Projects

The individual state or federal departments of education have occasionally provided research funding for science and literacy education projects. Some of the more recent initiatives are briefly described to illustrate the focus and variety.

The Victoria state Science in Schools (SIS) project was the most substantial single research project ($1.7 million) to have been undertaken in science education in Australia for a number of decades. The Department of Innovation, Industry, and Regional Development awarded Deakin University this grant for research and management with substantial in-kind funding for a regional project officer and school involvement from the participating organizations. This funding reflected a political priority for establishing the state of Victoria as a leader in technological innovation. SIS itself was expanded to include mathematics and technology and became the basis for a successful ARC-funded project.

Spotlight on Science (SOS, 2003–2006) was supported under a Queensland state government mandate to increase the numbers of students studying science and to attract and retain skilled science teachers ($14 million). The SOS project (State of Queensland Department of Education, Training and the Arts, n.d.) involved activities for students and professional development (PD) for science teachers to address the science-teacher shortage and the mandate that all students study science to Year 10. The PD helped teachers develop course outlines and partnerships with industry and was facilitated with the establishment of a Centre of Excellence.

Primary Connections (Primary Connections, n.d.), a nationwide, professional learning program for primary teachers (2005–2008), linked literacy and science learning with a grant from the federal Department of Education, Science, and Training (DEST) and the Australian Academy of Science ($4.8 million). The program was developed through consultation with key curriculum stakeholders at state and federal levels; it includes professional learning modules for teachers and 19 exemplar science units across the primary science curriculum. The Primary Connections modules and PD have been adopted by all state and territory jurisdictions.

The Primary Pre-service Teacher Awards for Excellence in Science Education was initiated by the federal government in 2007 as a recognition program for exemplary students in the final stages of their preservice primary teacher education program (AU DEST, 2007). The goal was to strengthen the scientific literacy of qualifying primary teachers and to allow their enthusiasm to stimulate improvements in primary students’ interest in, and learning of, science.

The National Centre of Science, Information and Communication Technology, and Mathematics Education for Rural and Regional Australia (SiMERR) provides a national forum for addressing issues relating to science, ICT, and mathematics education as they concern rural and regional communities (National Centre of Science, n.d.). Through a combination of strategic research, network building, and practical support, this project aims to identify the needs of geographically and professionally isolated teachers and to enhance their efforts to assist students realize their academic potential in these disciplines. Coordinated by the University of New England and supported with a $5 million government grant over 3 years, this project involves hubs in each state and territory to establish links with teachers, education providers, and relevant professional and community organizations, and to identify research opportunities and priorities.

The lack of long-term, ongoing research funding for science and literacy education in Australia has limited the effects of many R&D projects focused on changing education policy, school priorities, and classroom practices. Many of the funding opportunities that do exist are not a result of governmental policy or priorities, which would allow long-term research, small-scale trials, and scaled implementation. The quality and quantity of literacy and science education research are largely the result of a system in which all kinds of education research compete for very limited funding within a context of changing governments, pressing issues, and societal priorities.

23.7 New Zealand

New Zealand, notable for its geographic isolation, has a population of approximately 4.3 million; this population is mostly of European descent, and the indigenous Māori is the largest minority. Polynesian and Asian people are also significant minorities, especially in the cities. The government's annual operational research, science, and technology investment for 2007 was $657 million. However, science education receives only a very small proportion of these funds. It is difficult to quantify precisely the amount available to scientific literacy as it may be included in other initiatives. The main providers for science education research are two government ministries with distinctly different goals and approaches: Ministry of Education and Ministry for Research, Science, and Technology.

The Ministry of Education (NZ MOE, n.d.) tends to provide funding for operational aspects, such as curriculum and PD, as well as funding for evaluation of student scientific literacy, such as PISA. Within these practical and functional areas of curriculum and instruction, particular aspects will be prioritized. Currently, the MOE is funding (~$1.5 million per year) initiatives related to Learning Experiences Outside the Classroom (LEOTC), which provides funding to support educational activities related to scientific organizations, museums, etc. (Moreland, Jones, & Cowie, 2006). These extracurricular experiences enable students to link learning experiences between previsits, visits, and postvisits with curricular expectations when assisted by teachers and education officers. The collaborations between teachers, education officers, parents, and students resulted in improved learning related to the quality of site exhibits, especially hands-on and real-life experiences. Students also developed positive attitudes about these learning experiences. The LEOTC proposals for funding are evaluated against links to curriculum including scientific literacy where appropriate, quality of the educational program (the staff, opportunities for student learning, etc.), possible learning outcomes, and effective linkages with schools.

The Ministry of Research, Science and Technology (NZ MRST, n.d.) funded the NZ Science Mathematics and Technology Teacher Fellowship, administered by the

Royal Society of New Zealand (~$4.5 million per year), which seeks to raise the profile of science, mathematics, social sciences, and technology within the wider community and to promote a knowledge society. One way of achieving public awareness of science and technology is to provide teachers with new experiences and understanding outside the classroom that enable them to become more effective educators. These fellowships (up to 52 per year) allow teachers to engage in scientific endeavors for up to 1 year with scientific host organizations. The major intentions of these authentic science experiences are to enhance teachers' scientific literacy and their advocacy for science education in that they are expected to share their learning with other teachers. During their fellowship, teachers are able to fully immerse themselves in the discovery and transformation of knowledge and become more skilled in the communication of science and technology. The evidence supports claims that teachers return to the classroom rejuvenated and better able to enthuse their students about the career possibilities of science, mathematics, and technology in New Zealand and to further the government's goals in moving toward a knowledge society. The criteria for selecting these teacher–fellows include links to program's aims, teacher's abilities, and personal attributes, potential impact on students and wider community, connections with host organization, and potential increase in teacher's science knowledge.

The MRST has provided major funding (~$2 million per year) for scientific literacy through the notion of engaging or reengaging in science. Two learning hubs (ICT portals) have been funded—New Zealand Biotechnology Learning Hub and New Zealand Science Learning Hub—to make contemporary science and technology research accessible to school students and teachers. Specific aims and approaches were based on research results and included (a) raising awareness of how science and technology concepts included in the school curriculum relate to the modern world, (b) demonstrating effective strategies to transform research organizations' science stories into classroom experiences, and (c) developing an online, digital framework to promote communication between the R&D sector and schools. The portals provide alternative and contemporary information sources to supplement or replace outdated instructional resources and to illustrate industry stories of science and technology as potential pedagogy.

This governmental initiative reflects the unique situation in which education research results were used to inform the development of the policy, strategic plan, implementation, and evaluation of these portals. This research included focus groups with industry representatives and classroom-based case studies. The focus group interviews conducted with representatives from 6 government-funded institutes, 5 universities, and 11 private biotechnology companies identified four main themes: how to engage and nurture student curiosity and interest; how to develop and foster scientific and technological literacy for responsible citizenship; existing initiatives between the industry sector and schools; and how an online framework can be used to support science and technology education. The classroom case studies suggested that (a) teachers believed access to relevant resource materials and experts were important for enhancing the learning program; (b) there was a need for alternative assessments; (c) teachers needed to understand the nature of science and technology, the underlying concepts, and their implications on learning and teaching; and (d) the science and technology problems and hands-on activities needed to be embedded in real-life contexts to engage students with the inherent complexities. Therefore, the portals were designed to provide both access to experts in a sustainable, multimedia format with regular updates reflecting the needs of modern learners and also layered information to accommodate different educational needs, levels, and interests. Reactions from key stakeholders (teachers and industry) have been extremely encouraging, and classroom trials continue to inform development and functionality, indicating the importance of research-policy partnerships.

The difficulties in New Zealand are that there are very few blue-sky research opportunities and the mission-driven funding is about government priorities. The exception would be the MOE's Teaching and Learning Research Initiative (TLRI) funding (~$2 million per year) that establish 1-, 2-, or 3-year research partnerships focused on building knowledge about teaching and learning and the use of this knowledge to improve outcomes for learners. These practitioner–researcher partnerships are annually informed by needs surveys of experts and are designed to maximize the value and usefulness of research results. The TLRI Overview states that this funding envelope aims to "build a cumulative body of knowledge linking teaching and learning, enhance the links between educational research and teaching practices, … [and] grow research capability and capacity in the areas of teaching and learning" (NZ TLRI, n.d., para. 1). Table 23.2 summarizes the expressions of interest received, short-listed, and granted for 2003–2007. Of the projects that have been funded, 10% have included aspects of science, with significantly more being funded for language and literacy particularly focused on enhancing classroom practices (see http://www.tlri.org.nz/ for funded projects).

The MOE also provides small amounts of funding for science centers and PD through School Support Services, which are mainly focused on enhancing teacher practices. Overall, New Zealand has not had significant amounts of funding for research in scientific literacy in the past. However, with the recent development of science engagement and increased fundamental research, there are renewed opportunities for expansion. New Zealand, unlike other countries, does not have

Table 23.2 New Zealand Teaching and Learning Research Initiative: Number of expressions of interest received, short-listed, and grants awarded (2003–2007)

Year	Received	Short-listed	Awarded
2003	180	30	13
2004	72	24	18
2005	67	22	12
2006	52	24	12
2007	40	14	9
Total	409	114	64

large pools of education research funding and researchers, let alone funding for, and capacity in, science education.

23.8 South Africa

Any consideration of funding priorities and research in South Africa needs to be set against the background of socioeconomic factors and general education priorities. The country has emerged from a turbulent history of apartheid where education lay at the center of the struggle. The 1976 Soweto riots by school students marked a turning point in the struggle, which led to instability in the school system that prevailed until the first democratic elections in 1994, leaving a whole generation of children with inadequate education. The AIDS pandemic has also had a negative effect on the education system and society with the challenges posed by child-headed households and teachers affected by the illness.

The South Africa government embarked on ambitious reforms to the school curriculum and restructuring in the higher education system that would challenge the most well-developed education system and research culture, let alone a system with a shortage of key skills in science, mathematics, and literacy education. At present, there are severe shortages in mathematics and science teachers and Kindergarten–Grade 3 teachers proficient in the majority languages. This teacher shortage is in the context of the closure of many teacher education institutions and teacher preparation becoming the responsibility of the higher education system.

South Africa has participated in several international studies, such as TIMMS, and has the dubious distinction of being the worst performing country in many of these surveys, even when compared to other countries in Africa that are less developed. South Africa, like many countries, is establishing a research culture that supports quality research (see Coll et al., Chap. 6). Against this background, several clear priorities arise for research. Several bodies fund science education research; some are funded directly by government and others by foreign foundations and the corporate sector. Since the government determines education policy, it is the priorities of government agencies that play a key role in determining funding agendas while the foundations and corporations respond to these mission-driven priorities and other nongovernmental initiatives.

23.8.1 Federal Research Funding Agencies

The National Research Foundation (SA NRF, 2007) rates and funds researchers in tertiary institutions and other higher educational organizations, such as museums, and other statutory bodies. The Department of National Education also rewards publication by university academics in recognized journals by paying approximately $10,000 per publication to the institution. The NRF has several grant schemes that reflect their policies, especially those related to building capacity and achieving demographic equity in the number and level of researchers. One category of research grants, known as Thuthuka, directs special funding toward women and black researchers—although some white men under the age of 45 may be funded.

The largest funding scheme, known as the Focus Area Programme, supports researchers and proposals that either involve black and women researchers or include graduate students from these groups in a wide range of disciplines, including "Education and the challenges for change" (SA NRF, n.d.-a). This focus area is divided into seven themes that reflect the education priorities: (a) restructuring in higher education/further education and training; (b) policy implementation studies; (c) science, technology, and mathematics education (STME); (d) human resource development/teacher education and development; (e) curriculum, pedagogy, and assessment; (f) language issues and literacy; and (g) HIV/AIDS in education. Two disciplinary areas—language, and mathematics and science education—are defined as separate themes, indicating their importance. The STME theme is further subdivided into curriculum reform and systemic research, assessment, exemplary practice, the nature of scientific knowledge, the role of language in learning these disciplines and science, technology and mathematics literacy, learner's thinking, and learning and the classroom climate. Amounts available are small; awards above $100,000 are rare, but some do reach $200,000.

NRF guidelines and criteria for proposals emphasize that scientific quality is of paramount importance, the impact of the research should not be localized to a particular department within an institution, institutions should bear the costs of specific curriculum reconstruction, and the educational theoretical underpinnings must be made explicit and contain a research-training component. Proposals are judged in terms of quality and relevance by panels of peers using a two-step procedure (SA NRF, n.d.-c). The 2007 quality (scientific merit) criterion of the primary judgment specifies six categories: exceptional, excellent, very good, good, fair, and poor. The exceptional rating requires that the proposal be judged to have the highest scientific/research merit and be at the forefront of research in the field, while the poor rating requires that the proposal be judged to have one or more scientific flaws. Proposals rated fair and poor are dropped from further consideration, and those rated good are less likely to receive funding due to limited budget. The second step of the evaluation applies relevance criterion involving judgments regarding alignment of the proposal with the objectives for the Focus Area Programme, human resource capacity development in terms of general equity and redress of inequities, and the strategic importance of proposed outcomes. The 2008 review process is set to change with a greater emphasis placed on the evaluation of researchers and with increased funding, but this clarification has not occurred for the education focus. The NRF also has other funding opportunities with national and international partners (SA NRF, n.d.-b).

At present, the NRF does not express a preferred research approach; and the majority of proposals in science education are smaller-scale qualitative studies. This is lamented by the funding agency, which would like to see more balance in the types of research designs and data collection. NRF policy places a strong emphasis on research training, specifically at the Ph.D. level, as a driver for building research capacity; therefore, much of the money allocated is for graduate student bursaries for tuition fees. Grant money for operating expenses is limited and does not cover major equipment or teaching release, which is critical since education professors' teaching loads are relatively high; without such funding, it is difficult to carry out large-scale studies. The greatest challenge facing science education researchers relates to the fact that most graduate students are generally experienced teachers doing part-time studies, leading to higher dropout rates and slower progress.

23.8.2 Nongovernmental and Statutory Research Funding Agencies

Other bodies providing funding for science education research include the statutory body, the Human Sciences Research Council (HSRC), nongovernmental organizations, such as the Joint Education Trust, and the corporate-funded Centre for Development Enterprise (CDE). Some researchers also obtain overseas funding from countries such as the United States (Carnegie), Norway, Canada, and the United Kingdom. Unlike the NRF, local bodies, such as the HSRC and the CDE, usually identify mission-driven rather than curiosity-driven issues; and they design and carry out the research using either full-time employees (HSRC) or consultants (CDE). The funding available from these agencies and their modus operandi allow larger-scale research, such as the TIMMS assessment carried out by the HSRC.

The program within HSRC that conducts most of the science education research is known as the Education, Science, and Skills Development program (SA HSRC, n.d.). The focus areas relevant to literacy and science education include: (a) science, technology, and innovation in developing countries; (b) language and literacies studies, language policy, and implementation; (c) science, mathematics, and technology education research; (d) monitoring and evaluation of educational improvement from national level to the classroom level; (e) student achievement in mathematics and science; (f) trend analysis of student performance for the purpose of system-level planning; and (g) out-of-school programs for mathematics and science. Much of the contract research finds its way into nonrefereed reports produced for policymakers that are either made available on the Internet or for sale. For example, two CDE reports (Clynick & Lee, 2004; Simkins, Rule, & Bernstein, 2007) were launched with great publicity with the hope of influencing government policy regarding the improvement of the teaching of mathematics and science. The HSRC produces a combination of refereed and nonrefereed publications, which are often done in partnership with research units and higher education institutions. Like the higher education community, HSRC researchers also write books (e.g., Reddy, Kanjee, Diedericks, & Winnaar, 2006) that are published by the organization's press and publish in articles in refereed journals.

23.8.3 Research Output and Focus

An examination of the refereed journal output by the South African science education community is illuminating, as it gives an idea of what research is being done at the level of local and international publication. Two recent studies (Rollnick, Adler, & Setati, in press; Venkatkrishnan, Adler, Rollnick, Setati, & Vhurumuku, in press) surveyed research carried out in South Africa that was published in local and international science education and education journals (2000–2006) so as to assess its impact on policy. They identified 104 articles: 33 in local science education journals, 35 in local education journals, and 36 in international science education journals; the large majority of the 131 authors were from advantaged institutions, only 25% were of African descent, and 40% were females.

The studies were predominantly empirical (75%) with some conceptual and document analysis studies. The majority of articles were about teaching and learning issues (66%) with smaller percentages on nature of science (12%, including studies on indigenous knowledge) and teacher education (15%). Only two articles explicitly considered policy although many classified as teaching and learning addressed curriculum issues. What was more startling was the large percentage of studies carried out at the tertiary level: 40% on science and engineering students, and 14% on professors, instructors, and teachers; 23% of the studies were at the secondary school level, 8% at middle school level, and only 6% at the elementary school level. The emphasis at tertiary level was attributed to the large number of researchers working on science bridging programs at university, combined with the greater ease of access to research subjects at the tertiary level. The paucity of studies at the elementary school level and rural settings was cause for concern.

Finally, there was a dearth of studies specifically related to HIV/AIDS awareness education and insufficient large-scale studies, which could be attributed to the research funding leading to difficulties for researchers to do such scaled studies and the part-time status of the graduate students. Hence, large-scale studies are mainly carried out by statutory bodies for paying clients or corporations using consultants where the priority is to complete the research rather than to publish it in peer-reviewed journals. One trend that runs across funding policy, research studies, and publications is the push for equity for underserved and underrepresented peoples and to redress social justice issues.

23.9 Taiwan (Republic of China)

The Ministry of Education (TW MOE, n.d.) and the National Science Council (TW NSC, n.d.-a) are two funding agencies for science education research in Taiwan (Republic of China). These agencies complement and supplement one another in terms of goals and funding initiatives and reflect direct connections to public policies. The science education research funded by MOE tends to be more practical, goal-oriented, and linked to the political process, such as the Grades 1–9 science curriculum reform and development, the new secondary school curriculum, and the Grades 1–12 science teacher projects. The MOE also funds a 5-year, $50 billion (NTD) project (2006–2011, $10 billion per year) that promotes university research covering the full spectrum of academic disciplines across engineering, social science, science, and arts and humanities. In 2006 and 2007, about 70% of $10 billion is designated to the top four universities (National Taiwan University, National Chen Kung University, National Tsing Hua University, and National Chiao Tung University), with the other 30% of the $10 billion allocated to the other seven universities.

NSC is the major funding agency for science education research and has been part of the Executive Yuan (executive branch of the central government of Taiwan) since 1959. The NSC organization involves 15 research and administration units (TW NSC, n.d.-b). There are five discipline-oriented departments involved in research: Department of Engineering and Applied Sciences, Department of Humanities and Social Sciences, Department of Life Sciences, Department of Natural Sciences, and Department of Science Education. The mission of the Department of Science Education (DSE) is to (a) promote high-quality science education research that can provide the foundations for public science education practice and policy actions and to outline future science education policy directions, and (b) promote public scientific literacy and public understanding of science research through funding public science education projects. Those missions definitely guide the NSC funding patterns and priorities.

In 2002, the MOE and NSC held the first National Science Education Meeting of science educators, science teachers, and citizens to identify the central issues in science education (TW DSE, 2006). The White Paper (only available in Chinese) flowing from this meeting clearly stated that science education is about science literacy for all citizens to develop their creativity and science attitude. Furthermore, science education needs to emphasize inquiry, argumentation, thinking habits, and problem solving. Science education research and perspectives had great influence on this White Paper that identified several important research issues:

- Establish science education research evaluation and award system.
- Encourage longitudinal studies in order to build a research database, develop science education theories, and produce findings that will influence and justify classroom teaching practices.
- Emphasize research on science teacher education, certification, and evaluation as a foundation for a science teacher certificate.

The White Paper became an important guideline for the NSC's funding policy, priorities, and practices illustrating the research-policy connections.

23.9.1 Department of Science Education

The Director of the Department of Science Education is appointed from the highly qualified research professors in the universities. The Director reports to the Minister of the NSC who in turn reports to the Executive Yuan, a member of government appointed by the Premier of Executive Yuan. The six academic fields in the DSE are: Mathematics Education, Science Education 1 (science curriculum, learning, and evaluation), Science Education 2 (science teaching and teachers), Information and Computer Education, Medical Education, and Applied Science Education. The Director appoints a professor to lead each field for 3 years, who is responsible for the regular annual research project funding process and recommending research directions. The call for research proposals follows the mission of the department and the current interpretation of the White Paper on Science Education.

In addition to the regular research projects, there are many special calls for research proposals at different times to meet pressing issues and priorities of the NSC. For instance, the MOE and DSE are working together to reform the pre-service science teacher education program through 3-year, research-based projects (2006–2009); the NSC allocated about $100 million into five departments to support cross-department digital learning projects (2003) and $200 million to promote public awareness of the projects. The popular science funding focuses on recent scientific developments in Taiwan, which are broad ranging and often written by scientists rather than journalists. The projects' products are presented in many formats, including books, television documentaries, and magazine articles. The DSE also allocated ~$10–77 million per year to organize "Science Week" since 2000, which became an annual event and was renamed the "Science Festival" in 2006, in order to share findings of scientists and different science activities (e.g., variety of science, and 50 years of Taiwan technology).

The number of research proposals ranged from about 940 to 1,078 during 2000–2006, while the budget increased from ~$650 million to ~$825 million (Table 23.3). Funding increased to ~$1.053 billion in 2007. The percentage of projects funded average 54.9% and ranged between 49% and 62%. These data clearly reveal that the applications and funding have increased, but the acceptance rate is reasonably steady. The funding, application, and award patterns indicate the growing importance and priority assigned to science education research in Taiwan.

Table 23.3 Taiwan Department of Science Education: Funding, applications, and acceptance patterns (2000–2006)

Year	Amount of funding($ 000)	Number of applications	Success rate
2000	650,473	940	58
2001	486,012	1,013	62
2002	566,298	1,069	52
2003	683,825	1,064	55
2004	786,208	1,072	49
2005	799,684	1,041	55
2006	824,866	1,078	53

The DSE's regular research project budget is focused on curiosity-driven inquiries and open applications by researchers. This budget is divided into four evaluation categories:

- Special Research Project—Level A (top 5%), which covers postdoctoral fellows, graduate assistantship, full-time assistantship, conference travel expenses, and funding for the project director. The budget in this category is around $2 million.
- Special Research Project—Level B (top 6–10%), which covers graduate assistantship, full-time assistantship, conference travel expenses, and funding for the project director. The budget in this category is around $1–1.5 million.
- Regular Research Project (top 10–45%), which covers graduate assistantship, full-time assistantship, conference travel expenses, and funding for the project director. The budget for this category is about $750,000.
- Encouragement Research Project (between top 45% and 55%), which provides a very limited budget for research and no funding for other expenses. The budget for this category is less than $500,000.

In recent years, the DSE has started to call for large-scale and longitudinal studies to meet the science education White Paper. These calls address three areas: block projects regarding science teaching and learning, longitudinal projects regarding students' learning of science and mathematics, and longitudinal studies regarding citizens' scientific literacy, interest, and understanding. The proposals require consideration of theory and content development, classroom teaching practice, and a program of studies across several years. The funding for projects in this envelope is about eight to ten times the normal funding for regular projects.

The NSC's policy on funding and evaluation are based upon the following criteria:

- Research outcomes and applicants' ability (40%), which is composed of the number of publications by the applicants (20%), the quality of publications (10%), and the applicants' research ability (10%).
- Research proposal quality (60%), which consists of assessment of alignment with the call for proposal, value of the proposal, creativity of the project, significance and theoretical foundation of the proposal, design and methodology of the proposal, and proper budget.

These criteria and their weightings clearly indicate that the DSE considers research outcomes and ability most important for determining whether the project can be funded. In order to demonstrate their research outcomes and ability, the publication list within 5 years is the most relevant documentation. Different journals would count different scores; for instance, *Social Sciences Citation Index*® (SSCI®) journals would score higher than other journals. NSC emphasizes that research approaches must be appropriate for the research questions and goals of the proposal. Compared to MOE calls for proposals, the NSC has greater degrees of freedom for the researchers' problem space, research question, and research methodology.

23.10 Closing Remarks

Education is part of the social support system of most countries and an essential strategy in their governments' socioeconomic agenda. Several countries align their education R&D funding priorities with economic, social justice, and technological growth goals. Therefore, several agencies or ministries at the federal and state–provincial level fund mission-driven as well as curiosity-driven literacy and science education research activities. Some of these jurisdictions interpret their charges to include R&D and outreach (public awareness) activities involving public–private partnerships and socioeconomic initiatives to improve technological development, human resources, and industries. These mission-driven efforts have involved investments in curriculum frameworks, teacher enhancement, and science and technology promotions and celebrations. The European Parliament—consisting of a growing number of member states, developed countries, and developing countries—has explored a variety of strategic investments in research, knowledge translation, and other outreach activities to increase public awareness of science and technology, career opportunities in these areas, the need to build knowledge societies, and socioeconomic and social justice goals.

Funding envelopes (duration and amounts) appear to influence the type of research actually possible with curiosity-driven (blue skies, discovery, pure research activities) programs seen to be for a shorter term and smaller amounts, while mission-driven (curriculum development, implementation of reforms, professional development) programs are seen to be for a longer term and larger amounts. Curiosity-driven grants tend to be for 1–3 years and not large enough to support the evolution from exploratory, pilot studies to small-scale experiments and on to truly large-scale Gold Standard (RCT) research or to address and remediate many systemic problems. The German institutes and some research-oriented UK universities appear to be the exceptions since they are funded for longer periods to focus on specific research areas or to allocate their own funding. Mission-driven grants tend to be larger amounts available for longer periods (4–10 years). The US systemic initiatives and Australian, German, and UK curriculum development and implementation and teacher enhancement projects are examples of such programs.

The United States has federal policies regarding education and education research, but the funding practice of its NSF recognizes the need for research approaches that match the development of the problem space, specific nature of the research questions, and availability of investigative technologies (data collection and analysis techniques). The US ED appears to be somewhat more closely tied to the random control trials, random clinical trials, or random field trials design advocated in the Gold Standard. Other countries appear to stress rigorous and appropriate research approaches aligned with the problem space, research questions, and established knowledge base rather than focus on the *popular brand name* of the research methodology. The evaluation process and criteria emphasis of these countries is on the applicants' established records of productivity and publications, suggesting a belief that quality approaches can be predicted by past performance rather than actual impact; there does not appear to be significant attempts to assess the applicants' broader research agenda, development of their program of study, and the evolution of their designs to reflect a changing problem space and more acute research questions. The Taiwan NSC is one agency of very few that has started to encourage consideration of the program of research and the evolutionary agenda in one funding envelope.

Many countries value the influence of evidence-based research and generalized results on public policy and instructional practice, but there are few concerted attempts by the funding agencies (governmental and nongovernmental) to require dissemination of research results to broader audiences and end user groups composed of policymakers, teacher educators, school administrators, teachers, parents, and other stakeholders. There are limited examples of research results informing, refining, or implementing policies and a variety of strategies to increase public awareness of science and technology. Influencing end users outside of the normal academic discourse communities requires knowledge of the power structures, decision making, communication strategies, and prolonged effort not normally associated with academic research.

Over the past decade, there have been notable initiatives in what has become known as knowledge mobilization of educational research. These endeavors involve the development of accessible information portals for educational research to make it more available and informative to potential users—particularly in the policy community. In Canada, the United Kingdom, the United States, and elsewhere, there have been funded centers for systematic review of educational research that attempt to synthesize research findings so they are more coherent, focused, and accessible for policymakers. These research syntheses are focused on clearly described, policy-relevant areas of education, have a clearly articulated protocol of review and analysis, and have explicit criteria for the inclusion of research into the analyses—the research has to be of high quality and of an empirical nature. The most active and longest-standing centers for systematic reviews are the Campbell Collaboration (Campbell Collaboration, n.d.) at the international level, the Evidence for Policy and Practice Information and Co-ordinating Centre in the United Kingdom (EPPI, n.d.), and the *What Works Clearinghouse* in the United States (US IES, n.d.-c). In Canada, there have been a number of initiatives in the area of knowledge mobilization—the Canadian Centre for Knowledge Mobilization (CCKN, n.d.) at the University of Waterloo; the Canadian Council on Learning (CCL, n.d.) has a directorate of knowledge mobilization; the INE’s (CA SSHRC, n.d.-a) Educational Research Initiative, such as the Correlates of Learning Outcomes project and the NSERC CRYSTAL program (CA NSERC, n.d.-b).

Unfortunately, knowledge mobilization, scaling, capacity building, and policy-influence processes are not well understood and vary across governments and problem areas. Many research-funding agencies are promoting scaled studies and implementation projects to build research and leadership capacity in literacy and science education. These calls frequently involve systemic perspectives and broad focus across education organizations. System-level research is complex; many models involving resources, professional development, classroom practice, and student performance oversimplify the relationships amongst components, the contextual factors (political agenda, community expectations, union demands, etc.), and research demands.

Large-scale, system-wide, reform models do not always reflect the reality of education systems composed of systems and subsystems nested within one another. This complexity becomes greater when projects involve multiple partners: the federal/national, states/provincial, districts, schools, classrooms, students, parents, and other stakeholders. Many researchers use a mechanical metaphor when thinking of the education system with well-defined relations rather than an ecosystem metaphor with less predictable relationships; the latter metaphor provides better insights into the difficulties of scaling and capacity building. Scaling is not simply a multiplier factor related to the number of participating school districts, schools, teachers, and students—but rather an exponential increase in complexity, costs, and effort.

Building capacity, likewise, is underestimated in its complexity. Many countries and universities are trying to increase their research activities and productivity. On the surface, this appears to be simply a problem of numbers; but in reality, it involves cultural and support issues as well as the number of high-quality researchers. The research culture needs to consider the balance and recognition of the range of activities and responsibilities assumed by professors. This means that alignment must be achievement between expectations and rewards, between time allocation and tasks, and between differential assignment of tasks and productivity. Support systems need to be in place to allow researchers to focus on research and to limit their exposure to administrative demands and other tasks. Research administrative service needs to (a) include seed funding to develop high-quality, fundable research projects and proposals, and (b) address required considerations like ethics approval, technical writing, and budget development and management. Academic and professional associations can do much to help developing countries (e.g., in Africa, Asia, Central and South America, and the Middle East) and emerging universities and research organizations to establish high-quality researchers and research cultures. Summer schools and apprenticeships for researchers, support networks, and mentorships between established researchers and new researchers, and shared ethics standards and codes of conduct are a few examples of existing activities found in the literacy and science education communities.

The patterns and practices of research funding for literacy and science education has great potential and some established promise as science literacy is specifically mentioned in several white papers, commission and taskforce reports, and federal inquiries. The literacy and science education communities need to redouble their efforts to establish clear links between language as a cognitive tool, science understanding, science literacy, and the ability and willingness to participate in the public debate about science, technology, society, and environment issues to reach informed decisions and take sustainable actions.

Acknowledgement Several authors of this chapter serve or have served in various capacities within the funding agencies, ministries, and nongovernmental organizations. However, the views presented in the chapter are their own and do not necessarily reflect the views of the any agency, ministry, or organization.

References

Alberta CRYSTAL Project. (n.d.). *Homepage*. Retrieved June 19, 2008, from http://www.uofaweb.ualberta.ca/edpolicystudies/crystalalberta.cfm

Australia National Health and Medical Research Council. (n.d.). *Homepage*. Retrieved June 19, 2008, from http://www.nhmrc.gov.au/

Australian Council for Educational Research. (n.d.). *Homepage*. Retrieved June 19, 2008, from http://www.acer.edu.au/

Australian Department of Education, Science and Training. (2007, February 7). *New awards to promote science in primary schools* [Press release]. Retrieved from http://www.dest.gov.au/Ministers/Media/Bishop/2007/02/b001070207.asp

Australian Learning and Teaching Council. (n.d.). *Homepage*. Retrieved June 19, 2008, from http://www.altc.edu.au/carrick/go/home

Australian Research Council. (n.d.-a). *ARC Profile: Discovery*. Retrieved June 19, 2008, from http://www.arc.gov.au/about_arc/arc_profile.htm#discovery

Australian Research Council. (n.d.-b). *ARC Profile: Linkages*. Retrieved June 19, 2008, from http://www.arc.gov.au/about_arc/arc_profile.htm#linkage

Australian Research Council. (n.d.-c). *Linkage infrastructure, equipment and facilities*. Retrieved June 19, 2008, from http://www.arc.gov.au/ncgp/lief/lief_default.htm

Australian Research Council. (n.d.-d). *Linkage international*. Retrieved June 19, 2008, from http://www.arc.gov.au/ncgp/lx/lx_default.htm

Australian Research Council. (n.d.-e). *Linkage learned academies special projects*. Retrieved June 19, 2008, from http://www.arc.gov.au/ncgp/lasp/lasp_default.htm

Australian Research Council. (n.d.-f). *Linkage projects*. Retrieved June 19, 2008, from http://www.arc.gov.au/ncgp/lp/lp_default.htm

Baggott la Velle, L., & Erduran, S. (2007). Arguments and developments in the science curriculum. *School Science Review, 88*(324), 31–39.

Baumert, J., Bayrhuber, H., Brackhahn, B., Demuth, R., Durner, H. Fischer, H.E., et al. (1998). *Gutachten zur Vorbereitung des Programms: Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts, Materialien zur Bildungsplanung und zur Forschungsförderung*. [Expertise for preparing the program for increasing the efficiency of teaching mathematics and the sciences. Materials for Educational Planning and Research Promotion]. Bonn, Germany: Bund-Länder Commission for Educational Planning and Research Promotion. Available from http://www.bik-bonn.de/papers/heft60.pdf [in German]

Campbell Collaboration. (n.d.). *Homepage*. Retrieved October 3, 2007, from http://www.campbellcollaboration.org

Canada Department of Finance. (2007). *Budget 2007: Aspire to a stronger, safer, better Canada*. Retrieved June 18, 2008, from http://www.budget.gc.ca/2007/bp/bptoce.html

Canada Natural Sciences and Engineering Research Council. (2006). *Centres for Research in Youth, Science Teaching and Learning (CRYSTAL) selected*. Retrieved June 18, 2008, from http://www.nserc.gc.ca/about/initiatives/initiatives_articlejune_2_e.htm

Canada Natural Sciences and Engineering Research Council. (n.d.-a). *Homepage*. Retrieved June 18, 2008, from http://www.nserc.gc.ca/index.htm

Canada Natural Sciences and Engineering Research Council. (n.d.-b). *Use of grant funds – Centres for Research in Youth, Science Teaching and Learning (CRYSTAL)*. Retrieved June 20, 2008, from http://www.nserc.gc.ca/about/initiatives/crystal_funds_e.htm

Canada Networks of Centres of Excellence. (n.d.). *Homepage*. Retrieved June 18, 2008, from http://www.nce.gc.ca/

Canada Social Sciences and Humanities Research Council. (n.d.-a). *Apply for funding: Policies governing INE programs*. Retrieved June 18, 2008, from http://www.sshrc.ca/web/apply/policies/regulations_ine_e.asp

Canada Social Sciences and Humanities Research Council. (n.d.-b). *Homepage*. Retrieved June 18, 2008, from http://www.sshrc.ca/web/home_e.asp

Canada Social Sciences and Humanities Research Council. (n.d.-c). *Standard research grants: Regulations governing grant applications*. Retrieved June 18, 2008, from http://www.sshrc.ca/web/apply/program_descriptions/standard_e.asp

Canadian Centre for Knowledge Mobilisation. (n.d.). *Homepage*. Retrieved June 20, 2008, from http://www.cckm.ca/index2.htm

Canadian Centre for Research on Literacy. (n.d.). *Homepage*. Retrieved June 18, 2008, from http://www.uofaweb.ualberta.ca/elementaryed/ccrl.cfm

Canadian Council on Learning. (n.d.). *Homepage*. Retrieved June 18, 2008, from http://www.ccl-cca.ca/CCL/Home/index.htm?Language = EN

Canadian Institutes of Health Research. (n.d.). *Homepage*. Retrieved June 18, 2008, from http://www.cihr-irsc.gc.ca/e/193.html

Canadian Language and Literacy Research Network. (n.d.). *Homepage*. Retrieved June 18, 2008, from http://www.cllrnet.ca/

Clynick, T., & Lee, R. (2004). *From laggard to world class: Reforming maths and science in South Africa*'*s schools* [CDE Research no. 13]. Johannesburg, South Africa: Centre for Development and Enterprise. Available from http://www.cde.org.za/article.php?a_id = 50

Education Sciences Reform Act of 2002. Pub. L. No. 107–279, 116 Stat. 1940. (2002).

European Commission. (2006). *Decision No 1982/2006/EC of the European Parliament and of the Council*. Retrieved June 20, 2008, from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri = OJ:L:2006:412:0001:0041:EN:PDF

European Commission. (2007a). *The European research area: New perspectives* [Green paper]. Retrieved June 18, 2008, from http://ec.europa.eu/research/era/pdf/era_gp_final_en.pdf

European Commission. (2007b). *Science in society work programme 2008: Capacities*. Retrieved June 18, 2008, from http://cordis.europa.eu/fp7/wp_en.html#capacities

European Commission. (n.d.). *The Bologna process: Towards the European higher education area*. Retrieved June 18, 2008, from http://ec.europa.eu/education/policies/educ/bologna/bologna_en.html

Evidence for Policy and Practice Information and Co-ordinating Centre. (n.d.). *Homepage*. Retrieved June 20, 2008, from http://eppi.ioe.ac.uk/cms/

Ford, C. L., Yore, L. D., & Anthony, R. J. (1997). *Reforms, visions, and standards: A cross-curricular view from an elementary school perspective*. (ERIC Document Reproduction Service ED406168).

German Biologie im Kontext. (n.d.). *Homepage*. Retrieved June 18, 2008, from http://bik.ipn.uni-kiel.de/typo3/index.php?id = 3 [in German]

German Chemie im Kontext. (n.d.). *Homepage*. Retrieved June 18, 2008, from http://www.chik.de/ [in German]

German Federal Ministry of Education and Research. (2007a). *Rahmenprogramm zur Förderung der empirischen Bildungsforschung* [Framework concept to strengthen empirical educational research]. Retrieved June 18, 2008, from http://www.bmbf.de/pub/foerderung_der_empirischen_bildungsforschung.pdf [in German]

German Federal Ministry of Education and Research. (2007b). *Research at a glance – The German research landscape*. Retrieved June 18, 2008, from http://www.uni-mainz.de/downloads_nsm/bmbf_research_at_a_glance.pdf

German Federal Ministry of Education and Research. (n.d.). *Innovation in education*. Retrieved June 18, 2008, from http://www.bmbf.de/en/1076.php

German Institute for Adult Education. (n.d.). *Homepage*. Retrieved June 18, 2008, from http://www.die-bonn.de/portrait/english/index.htm

German Institute for International Educational Research. (n.d.). *Homepage*. Retrieved June 18, 2008, from http://www.dipf.de/ueber_uns_e.htm

German IPN–Leibniz Institute for Science Education. (n.d.). *Homepage*. Retrieved June 18, 2008, from http://www.ipn.uni-kiel.de/index_eng.html

German Knowledge Media Research Center. (n.d.). *Homepage*. Retrieved June 18, 2008, from http://www.iwm-kmrc.de/www/en/index.html

German Leibniz Gemeinschaft. (n.d.). *Homepage*. Retrieved June 19, 2008, from http://www.wgl.de/?nid = ubu&nidap [in German]

German Lernort Labor. (n.d.). *Homepage*. Retrieved June 18, 2008, from http://www.lernort-labor.de/ [in German]

German Physik im Kontext. (n.d.). *Homepage*. Retrieved June 18, 2008, from http://www.uni-kiel.de/piko/ [in German]

Great Britain University Grants Committee. (1984). *A strategy for higher education into the 1990s: The university grants committee*'*s advice*. London: HMSO.

Hand, B., Prain, V., & Yore, L. D. (2001). Sequential writing tasks' influence on science learning. In G. Rijlaarsdam (Series Ed.) & P. Tynjälä, L. Mason, & K. Lonka (Eds.), *Writing as a learning tool: Integrating theory and practice* (Vol. 7 of Studies in Writing, pp. 105–129). Dordrecht, The Netherlands: Kluwer.

Inner London Educational Authority. (1987). *Science in process: Be scientific*. Oxford, UK: Heinemann.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., et al. (2007). *Zur Entwicklung nationaler Bildungsstandards – Expertise*. [Concerning the development of national education standards]. Bonn, Germany: Federal Ministry of Education and Research. Available from http://www.bmbf.de/pub/zur_entwicklung_nationaler_bildungsstandards.pdf [in German]

Klieme, E., & Baumert, J. (2001). *TIMSS - Impulse für Schule und Unterricht: Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumentation* [Impulses for school and instruction: Evidences, reform initiatives, praxis reports, and video documentation]. Bonn, Germany: Federal Ministry of Education and Research.

Lucas, L. (2006). *The research game in academic life*. London: Society for Research into Higher Education & Open University Press.

Moreland, J., Jones, A., & Cowie, B. (2006). Developing pedagogical content knowledge for the new sciences: The example of biotechnology. *Teaching Education, 17*(2), 143–155.

National Centre of Science, Information and Communication Technology, and Mathematics Education for Rural and Regional Australia. (n.d.). *Homepage*. Retrieved June 19, 2008, from http://www.une.edu.au/simerr/

Nentwig, P. M., Demuth, R., Parchmann, I., Gräsel, C., & Ralle, B. (2007). *Chemie im Kontext:* Situating learning in relevant contexts while systematically developing basic chemical concepts. *Journal of Chemical Education, 84*(9), 1439–1444.

New Zealand Ministry of Education. (n.d.). *Homepage*. Retrieved June 18, 2008, from http://www.minedu.govt.nz/

New Zealand Ministry of Research, Science and Technology. (n.d.). *Homepage*. Retrieved June 18, 2008, from http://www.morst.govt.nz/

New Zealand Teaching and Learning Research Initiative. (n.d.). *About the TLRI*. Retrieved June 19, 2008, from http://www.tlri.org.nz/about.html

Newton, D. P., & Gott, R. (1989). Process in lower school science textbooks. *British Educational Research Journal, 15*(3), 249–258.

No Child Left Behind Act of 2001. Pub. L. No. 107–110, 115 Stat. 1425. (2002).

Pacific CRYSTAL Project. (n.d.). *Homepage*. Retrieved June 19, 2008, from http://www.educ.uvic.ca/pacificcrystal/main.html

Prenzel, M. (Ed.). (2007). *Studies on the educational quality of schools: The final report on the DFG priority programme*. Münster, Germany/New York: Waxmann.

Primary Connections – Linking science with literacy. (n.d.). *Homepage*. Retrieved June 19, 2008, from http://www.science.org.au/primaryconnections/

Reddy, V., Kanjee, A., Diedericks, G. A. M., & Winnaar, L. D. (2006). *Mathematics and science achievement at South African schools in TIMSS 2003*. Cape Town, South Africa: HSRC Press.

Rocard, M., Csermely, P., Jorde, D., Lenzen, D., Walberg-Henriksson, H., & Hemmo, V. (2007). *Science education now: A renewed pedagogy for the future of Europe*. Luxembourg, Belgium: European Commission. Available from http://ec.europa.eu/research/science-society/document_library/pdf_06/report-rocard-on-science-education_en.pdf

Rollnick, M., Adler, J., & Setati, M. (in press). Who are the mathematics and science education researchers in South Africa? *African Journal of Research in Mathematics, Science and Technology Education*.

Schavan, A. (n.d.). *Supporting research – Opening future markets*. Retrieved June 18, 2008, from http://www.bmbf.de/en/99.php

Sherwood, R. D., & Hanson, D. L. (2008). A review and analysis of the NSF portfolio in regard to research on science teacher education. *Electronic Journal of Science Education, 12*(1), 20–38. Retrieved from http://ejse.southwestern.edu/volumes/v12n1/v12n1.pdf

Simkins, C., Rule, S., & Bernstein, A. (2007). *Doubling for growth: Addressing the maths and science challenge in South Africa's schools* [CDE Research no. 15]. Johannesburg, South Africa: Centre for Development and Enterprise. Available from http://www.cde.org.za/article.php?a_id = 264

South Africa Human Sciences Research Council. (n.d.). *Education, science and skills development*. Retrieved June 19, 2008, from http://www.hsrc.ac.za/ESSD.phtml

South Africa National Research Foundation. (2007). *Homepage*. Retrieved April 17, 2008, from http://www.nrf.ac.za/

South Africa National Research Foundation. (n.d.-a). *Focus areas: Education and the challenges for change*. Retrieved June 19, 2008, from http://www.nrf.ac.za/focusareas/educate/

South Africa National Research Foundation. (n.d.-b). *Funding*. Retrieved June 19, 2008, from http://www.nrf.ac.za/funding/

South Africa National Research Foundation. (n.d.-c). *Monitoring and evaluation*. Retrieved June 19, 2008, from http://evaluation.nrf.ac.za/index.htm

State of Queensland, Australia, Department of Education, Training and the Arts. (n.d.). *Science state – Smart state spotlight on science 2003–2006*. Retrieved June 19, 2008, from http://education.qld.gov.au/publication/science/sciencestate.html

Taiwan Department of Science Education. (2006). *Mission*. Retrieved April 17, 2008, from http://www.nsc.gov.tw/sci/mp.asp?mp = 1

Taiwan Ministry of Education. (n.d.). *Homepage*. Retrieved June 20, 2008, from http://english.moe.gov.tw/mp.asp?mp = 1

Taiwan National Science Council. (n.d.-a). *Homepage*. Retrieved April 17, 2008, from http://web.nsc.gov.tw/default.asp?mp = 7

Taiwan National Science Council. (n.d.-b). *Organization*. Retrieved June 20, 2008, from http://web.nsc.gov.tw/ct.asp?xItem = 14932&CtNode = 3417

United Kingdom Department for Children, Schools and Families. (n.d.). *Science, technology, engineering and mathematics programme*. Retrieved June 30, 2008, from http://www.dfes.gov.uk/stem/

United Kingdom Department for Education. (1995). *Science in the national curriculum*. London: HMSO.

United Kingdom Department for Education and Employment. (1999). *Science: The national curriculum for England*. London: HMSO.

United Kingdom Department for Education and Skills. (2006). *Science: The national curriculum for England*. London: HMSO.

United Kingdom Department of Education and Science. (1988). *Science for ages 5 to 16*. London: HMSO.

United Kingdom Economic and Social Research Council. (n.d.). *Homepage*. Retrieved June 30, 2008, from http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/opportunities/current_funding_opportunities/research/

United Kingdom Higher Education Funding Council for England. (n.d.). *Homepage*. Retrieved June 19, 2008, from http://www.hefce.ac.uk/

United Kingdom HM Treasury. (2004). *Science & innovation investment framework 2004–2014*. Retrieved June 19, 2008, from http://www.hm-treasury.gov.uk./spending_review/spend_sr04/associated_documents/spending_sr04_science.cfm

United Kingdom National Curriculum Council. (1991). *Science in the national curriculum: A report to the Secretary of State for Education and Science on the statutory consultation for attainment targets and programmes of study in science*. London: Author.

United Kingdom Research Councils. (n.d.). *Homepage*. Retrieved June 19, 2008, from http://www.rcuk.ac.uk/default.htm

United States Department of Education. (2007). *Fiscal year 2008 budget summary*. Retrieved June 17, 2008, from http://www.ed.gov/about/overview/budget/budget08/summary/edlite-section2f.html

United States Department of Education. (n.d.). *Homepage*. Retrieved June 17, 2008, from http://www.ed.gov/index.jhtml

United States Institute of Education Sciences. (2008). *Education research grants CFDA number: 84.305A*. Retrieved June 17, 2008, from http://ies.ed.gov/funding/pdf/2009_84305A.pdf

United States Institute of Education Sciences. (n.d.-a). *Homepage*. Retrieved June 17, 2008, from http://ies.ed.gov/

United States Institute of Education Sciences. (n.d.-b). *Mission and goal*. Retrieved June 17, 2008, from http://ies.ed.gov/director/

United States Institute of Education Sciences. (n.d.-c). *What works clearinghouse: Homepage*. Retrieved June 20, 2008, from http://ies.ed.gov/ncee/wwc/

United States National Science Board of the National Science Foundation. (n.d.). *Homepage*. Retrieved June 17, 2008, from http://www.nsf.gov/nsb/

United States National Science Foundation. (2006a). *FY 2007 budget request to Congress*. Retrieved June 17, 2008, from http://www.nsf.gov/about/budget/fy2007/toc.jsp

United States National Science Foundation. (2006b). *Program solicitation NSF 08–502*. Retrieved June 17, 2008, from http://www.nsf.gov/pubs/2008/nsf08502/nsf08502.pdf

United States National Science Foundation. (2006c). *Program solicitation NSF 08–506*. Retrieved June 17, 2008, from http://www.nsf.gov/pubs/2008/nsf08506/nsf08506.pdf

United States National Science Foundation. (2007). *FY 2008 budget request to Congress*. Retrieved June 17, 2008, from http://www.nsf.gov/about/budget/fy2008/index.jsp

United States National Science Foundation. (2008). *FY 2009 budget request to Congress*. Retrieved June 23, 2008, from http://www.nsf.gov/about/budget/fy2009/toc.jsp

United States National Science Foundation. (n.d.-a). *About education and human resources*. Retrieved June 17, 2008, from http://nsf.gov/ehr/about.jsp

United States National Science Foundation. (n.d.-b). *Homepage*. Retrieved June 17, 2008, from http://www.nsf.gov/

University of Essen-Duisburg. (n.d.). *Research group and graduate school: Teaching and learning of science*. Retrieved June 18, 2008, from http://www.uni-essen.de/nwu-essen/dox/13.1178.hgF9v.H.En.php

Venkatkrishnan, H., Adler, J., Rollnick, M., Setati, M., & Vhurumuku, E. (in press). Mathematics and science education research, policy and practice in South Africa: What are the relationships? *African Journal of Research in Mathematics, Science and Technology Education*.

Yore, L. D., Pimm, D., & Tuan, H.-L. (2007). The literacy component of mathematical and scientific literacy. *International Journal of Science and Mathematics Education, 5*(4), 559–589.

Chapter 24
Research Ethics Boards and the Gold Standard(s) in Literacy and Science Education Research

Robert J. Anthony, Larry D. Yore, Richard K. Coll, Justin Dillon, Mei-Hung Chiu, Cynthia Fakudze, Irene Grimberg, and Bing-Jyun Wang

Curiosity-driven research has traditionally investigated problems, issues, and challenges through a variety of research designs to match the research foci without many formal constraints. The character of those designs has been the venue of the researchers, to some degree the funding agency, and the research setting. The creative challenge for the researcher has been consideration of the nature of the problem and research question, development of the problem space, and the monetary, instrumental, and contextual resources available. Increasingly over the last 10–15 years, another presence has joined the research team—the Research Ethics Board (REB), Research Ethics Committee (REC), or Institutional Review Board (IRB). REBs (we use REBS, RECs, and IRBs interchangeably in this chapter) play a mandatory role in reviewing and permitting research conducted under the agency of funding bodies and educational or research institutions in many countries. Over this same time, REBs have become widely accepted as a necessary and reasonable process to ensure that ethical standards of research are maintained and to avoid the potential for litigation resulting from

R.J. Anthony
University of Victoria

L.D. Yore
University of Victoria

R.K. Coll
University of Waikato

J. Dillon
King's College London

M.-H. Chiu
National Taiwan Normal University

C. Fakudze
University of Cape Town

I. Grimberg
Montana State University

B.-J. Wang
Yuan Ze University

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

faulty research designs and procedures. However, some researchers contend that the unified research ethics regulations, or *common rule*, for all disciplines overemphasize biomedical inquiries, risks, and norms—leaving much of the uniqueness of social sciences, education, and professional practices and their associated research methods lacking consideration. While the value of REBs is recognized, it is also evident that their procedures and practices are not stable or neutral in their impact on researchers, the potential research topics that are undertaken, and the research designs utilized. These effects and the array of differential influences can be seen on every campus and organization where research ethics reviews operate and, as described in this chapter, in Africa, Asia, Canada, Europe, New Zealand, and the United States.

Explorations of these effects have begun to appear in the academic communities. At the 2nd Island Conference, researchers from many countries came together to discuss contemporary issues in literacy and science education research in light of current national policies that impact this research—in particular, REBs on a global scale and the Gold Standard for research in the United States. In this international setting, researchers had the opportunity to reflect on these policies, the policies' influence on their own research, and implications for future research. Increasingly, the consequences of these policies are starting to be found in the education literature (Sieber, 2006). Some of the issues that have arisen include differences in the interpretation of the domains of power that REBs have over research and special consideration of peoples embedded in law or traditions. For example, although REBs are governed by broadly phrased guidelines for the ethical conduct of research in Canada (Canadian Institutes of Health Research [CIHR], Natural Sciences and Engineering Research Council of Canada [NSERC], & Social Sciences and Humanities Research Council of Canada [SSHRC], 1998), the interpretation of these guidelines is left in the hands of the individual REBs. Thus, REBs may adopt practices and policies of review that differ significantly from setting to setting and even within REBs from researcher to researcher (Anthony, 2004). The authority that REBs take with regard to the review and approval of research can vary widely and thereby differentially impact research. Likewise, national policies allow for *local options*; and the interpretations, procedures, and practices are moving targets (Sieber, 2007).

This chapter provides a theoretic background for research ethics and elaborates critical issues, deliberations, and recommendations flowing from the 2nd Island Conference and other related conferences based on the original deliberations. These critical issues are used as a template for (a) international and aboriginal–indigenous peoples' perspectives and practical resolutions regarding the critical dimensions of research ethics and review procedures and (b) future considerations and other related ethical issues for literacy and science education research.

24.1 Background

Historically, research ethics gained most of its public attention and scrutiny from medical, pharmaceutical, military, and biotechnological research while research in the humanities and social sciences was disregarded. Recent considerations of human rights, privacy, and equality issues have increased attention on social sciences research; however, much can be learned from the ethical issues of the high-profile areas. The first research ethics issue emerged from the post-World War II Nuremberg Tribunal for war criminals, which developed into the Nuremberg Code (Nuremburg Code, 1948) to protect participants in experiments on the human body and explicitly established the importance of informed consent and voluntary participation. The Clinical Center of the National Institutes of Health (NIH) in the United States used these ideas as foundation and developed policy for protecting human beings as experimental subjects. In 1964, the World Medical Association (WMA) announced the Declaration of Helsinki (WMA, 2004) that specified the ethical principles for medical research involving human subjects. These principles have been amended several times, but four (of 32) principles have application to this chapter:

> (5) In medical research on human subjects, considerations related to the well-being of the human subject should take precedence over the interests of science and society.
>
> (10) It is the duty of the physician in medical research to protect the life, health, privacy, and dignity of the human subject.
>
> (20) The subjects must be volunteers and informed participants in the research project.
>
> (22) In any research on human beings, each potential subject must be adequately informed of the aims, methods, sources of funding, any possible conflicts of interest, institutional affiliations of the researcher, the anticipated benefits and potential risks of the study and the discomfort it may entail. (WMA)

In 1979, the US National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research released the Belmont Report (US NCPHSBBR, 1979), which provided guidelines for research ethics that were not mentioned explicitly in the Nuremberg Code and the Declaration of Helsinki and that were applicable to educational research. The Belmont Report established three basic ethical principles—respect for persons, beneficence, and justice—as the cornerstone for regulations involving human subjects.

Recently, funding agencies have used the *big stick* approach to mandating research ethics review of projects and institutions receiving support (CIHR et al., 1998; US National Research Act of 1974; UK Economic and Social Research Council, 2006) while international research associations have focused on their members' conduct regarding professional behavior, research inquiries, knowledge construction, and ownership and intellectual properties (American Educational Research Association, 1992; American Psychological Association, 2002; British Educational Research Association, 2004; International Reading Association, 2008; National Science Teachers Association, 2007; Strike et al., 2002). Unlike the high-profile health and modality research ethical restrictions, the concerns stimulating these actions have frequently been based on anecdotal records of negative events with little empirical exploration and evidence or documented resolution of potential difficulties (Pritchard, 2002; Sieber, 2006). Most of the issues arising in these cases fall under a combination of legal, moral, and ethical considerations.

Legal considerations involve violation of civil and criminal law, and include the unauthorized use of someone's ideas, violation of copyright, fraudulent use of authority and privileged positions, inappropriate conduct with underage people and clients, libel, and other infringements of public or professional statutes. Moral considerations are founded in the less well-defined standards of public, cultural, and professional values and virtues (e.g., good–evil, honesty–dishonesty, integrity–deceit, right–wrong, responsible–irresponsible, etc.). di Norcia (2006) stated:

> Given a large enough sample, one would expect moral values and conduct to range from serious but rare deviance (evil and immoral), to average commonplace conformity (moral minimal and perhaps satisfactory), through creative and insightful ethical problem solving …, to intense and rare commitment (moral heroism, sainthood). (p. 2)

Ethical considerations involve a set of principles derived from legal and moral consideration that include, but are not limited to, customs, habits, conduct, etc. Collectively, consideration of these attributes as they apply to research is a perplexing and critical issue. Sieber (2006) stated:

> To illustrate the speed [and importance] with which [the research ethics] field of study may change, a few months ago it would have been foolish to suggest a situation in which one society's questioning the legitimacy of a particular line of scientific inquiry would help to motivate another society to dominate that field of research and announce a series of scientific victories—that turned out to be fraudulent. But now we know that in the void left by the hesitancy of the U.S. to embrace stem cell research, South Korean scientists took the international lead and prematurely declared a breakthrough purported to cure disabilities and disease. (p. 1)

The legal, moral, and ethical ramifications of this ill-advised action was felt by South Korea, Seoul National University, and a leading scientific journal, *Science* (see Kennedy, 2006, for the retraction of the violating article).

24.1.1 Codes of Ethics and Standards of Professional Conduct

Codes of ethics and standards of professional conduct are intended to be proactive devices to heighten awareness and avoid problems. The American Psychological Association (APA, 2002) established four general principles—beneficence and nonmaleficence, fidelity and responsibility, integrity, and respect for people's rights and dignity—of ethical practice that were incorporated into standards of practice and conduct for their members' various responsibilities and research activities. Frequently, ethical misbehaviors related to these codes and standards involve not-so-serious "infractions of falsification, fabrication, and plagiarism" (de Vries, Anderson, & Martinson, 2006, p. 43). Cohen (2005) reported that about 13 cases reach sanction-level annually at the US Department of Health and Human Service's Office of Research Integrity. de Vries and colleagues suggested that more often "misconduct generally is associated with more mundane, everyday problems in the work environment …, [falling] into four categories: the meaning of data, the rules of science, life with colleagues, and the pressures of production" (p. 43). Meaning of data concerns relate to general issues of outliers in a dataset and the "line between 'cleaning' data and 'cooking data' [during data interpretation]"

(p. 45) while rules of science involve comingling funds amongst different, ongoing research projects. Life with colleagues in a research community recognizes research as a social endeavor in which research team members collaborate and demand a degree of academic civility and traditional hierarchical authority is deemphasized (Florence & Yore, 2004). The publish-or-perish pressure is an ongoing condition of the academy and, in conjunction with fuzzy boundaries of ownership flowing from collaborations, leads to ill-advised use of data and knowledge claims by an individual or the listing of coauthors not truly involved in the research and knowledge-building processes (Strike et al., 2002).

24.1.2 Principles of Research Ethics

When these legal, moral, and ethical dimensions are applied productively to literacy and science education research, it will address some of the underlying concerns embedded in the Gold Standard by enhancing the "public trust in the research enterprise [that] can be nurtured in ways more fruitful than the conventional default preoccupation with regulatory compliance" (Landwirth, 2006, p. 3). Landwirth suggested that some research institutions have designated ethicists and centers in matchmaker roles to collaborate with researchers to proactively address ethics awareness and issues in the design, conduct, and reporting of their research. "Typically, the [researcher] brings only limited experience with the methods and language of ethical analysis, but a strong intuitive ethical sensitivity" (p. 3). This observation can easily be extended to REC chairs and panel members charged with shaping, monitoring, and enforcing ethics in education research. Many research ethics policies attempt to provide an integrated set of guiding principles in a common framework or common rule for all disciplines. Pritchard (2002) stated that the fundamental principles are:

> respect for persons, beneficence, and justice. The principle of respect for persons underlies the obligation to obtain informed consent; the principle of beneficence demands the maximizing of benefit and minimizing of risk; and the principle of justice requires the equitable distribution of the burdens and the benefits of research. (p. 8)

The solution to the ongoing problems facing REBs was to apply these common rule principles across unique and diverse research domains.

> Applied ethics, whether in field research or any other endeavour, should not necessarily contain anything that is ethically peculiar or unique. They should be nothing more than a particularized version of a universal ethical system or code, where the particulars are a function of the nature of the activities unique to that application. … Applied ethics, therefore, should be the application of general ethical principles to specific activities. (Truscott, 2004, p. 812)

Truscott suggested that these judgments should be based on an explicit set of criteria, cases, and conscious decisions—not intuitive, spontaneous, and emotional judgments. Strike and colleagues (2002) provided such illustrative cases associated with each ethics standard developed by the American Education Research Association.

24.1.3 Research Ethics Policies, Implementation, and Review Boards

The US National Research Act of 1974 (§ 474) established that:

> (a) The Secretary shall by regulation require that each entity which applies for a grant or contract under this Act for any project or program which involves the conduct of biomedical or behavioral research involving human subjects submit in or with its application for such grant or contract assurances satisfactory to the Secretary that it has established (in accordance with regulations which the Secretary shall prescribe) a board (to be known as an 'Institutional Review Board') to review biomedical and behavioral research involving human subjects conducted at or sponsored by such entity in order to protect the rights of the human subjects of such research.

The legitimate focus and purpose of research policies and review boards are "to ensure the ethical treatment of research subjects" (Pritchard, 2002, p. 7). The problems encountered by implementation of research ethics in education and the operations of these boards or committees are increased with the diverse interpretations of research, problems addressed, and approaches utilized as the scope of education–social sciences research moves away from the biomedical tradition. Research in education ranges from (a) traditional, two-group, experimental inquiries to the effects on learners' achievement assessed by an accepted test of a well-established instructional program and (b) a safe, but innovative, alternative instructional program to community-based, participatory research focus on social justice issues and political actions of the least well represented and powerful members of a hierarchical authoritarian community. Ethical approval of these issues and designs involves drastically different considerations of the problem space, research questions, methodology, procedures, and reporting. Some approaches, like the traditional control–experimental group design are driven by hypothetico-deductive processes in which a clearly stated hypothesis and predicted outcomes inform data sources, instrument selection, participant recruitment, data collection and analysis, and reporting the argument and results. Other newer approaches—like community-based participatory actions, practitioner inquiry, action research, and classroom design experiments—are not planned and scripted a priori in the same manner as scientific inquiries and rely on being more responsive to events as they emerge, which enables a further stage of inquiry and research design. Ethics review in well-established, traditional designs are based on the evaluation of the stated purposes and procedures against established criteria reflected in most unified research ethics policies; review of the second category involves projections of the criteria into anticipated scenarios and assessment of the researchers' abilities to ethically address the unexpected, which are not reflected in most common rules (i.e., the researcher is opportunistic and responsive to events as they occur and enacts the next procedure of the inquiry, data collection, or data interpretation based on real-time monitor and regulation) (Moretti, Leadbeater, & Marshall, 2006; Zeni, 2001).

Pritchard (2002) stated that IRB members must:

> rely on the regulatory definition of research, which emphasizes the purpose directing the activity in question. Activities count as research to an IRB only if the activity undertaken

> reflects a deliberate objective of discovering or learning something new that transcends the particular activity. Research concerns the organized search for knowledge applicable to other similar phenomena: 'Research means a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge. (34 Code of Federal Regulations [CFR], 97.102[d])' (p. 4)

He continued:

> Because the IRB's purpose is to ensure the protection of human research subjects, a research activity only falls within the IRB's purview if it involves human subjects, as follows: 'Human Subject means a living individual about whom an investigator (whether professional or student) conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information. (34 CFR 97.102[f])' (p. 4)

This definition implies generalized knowledge claims but does not imply research approach or intent to publish or present publicly.

Some professionals and researchers view the REB's actions as infringements on the academic freedom provided by their institution or employment to pursue problems and questions of interest in an inquiry manner of their choice; as well, they are concerned that they require colleagues to make evaluations and "form opinions about the value of their colleagues' research" (Lopus, Grimes, Becker, & Pearson, 2007, p. 70). Major professional associations and some federal governments make this a mute issue by requiring agreement with a code of ethics as a condition of membership or as a condition of receiving a specific research grant or general institution funding. van den Hoonaard (2006) reported that some researchers seek to avoid such infringements by international collaborations based in places without such regulations and by research inquiries not involving human participants.

Traditional scientific inquiry designs utilizing experiment–control groups assigned by random selection and double-blind studies generally fit the ethics review process better than quasi-experimental, fieldwork, and naturalistic inquiries (de Laine, 2000; Keith-Spiegel, Koocher, & Tabachnick, 2006; Lee-Treweek & Linkogle, 2000; Simons & Usher, 2000). Plemmons (2007) stated, "There is a perception that the IRB does not fairly and accurately assess social/behavioral research protocols, especially ethnographic and participant-observation studies" (p. 71). She believed that the lack of public transparency with IRB actions and deliberations results in less responsive actions and lower applicant satisfaction. An analogy can be drawn with the familiar issue in quantitative research based on statistical analysis to balance Type I and Type II error. In Type I error, the standard for accepting a claim is set too low, thereby allowing inappropriate claims to be accepted; while in Type II error, a credible claim is mistakenly rejected. When REBs are overly zealous in applying an unreasonable threshold for approval, they avoid approving research that may include an element of ethical risk (Type I error). However, the emerging chorus from qualitative researchers points out that such unreasonable standards increase the likelihood of rejecting research that has the potential to make important findings (Type II error).

Brydon-Miller and Greenwood (2006) offered several examples where action research studies have been rendered impotent as a consequence of real and anticipated limits imposed by REBs. Sociology researchers in Canada, like education researchers, fear that the research ethics review procedures initiated in 2001 may influence the type of research questions explored and the research methods utilized toward inquiries that do not involve human subjects or toward quantitative designs (van den Hoonaard, 2006). van den Hoonaard found that between 1995 and 2004 (a) the number of masters' theses involving human subjects decreased by 24%, (b) the number of qualitative studies increased, and (c) the concerns expressed by graduate students and supervisors indicated difficulties with the research ethics review process.

Keith-Spiegel and colleagues (2006) believed that the level of satisfaction researchers express about research ethics approval and the operations of the IRBs was based on the implementation of ethics policies, resident expertise of board members, and procedural attention given to the evaluation of the original grant proposal independent of the ethics approval application and process. They believed researchers' satisfaction with ethics policies and review procedures decreases as research becomes less traditional and the designs move away from the norms of traditional scientific inquiries and laboratories and becomes embedded in sociocultural contexts. They surveyed the satisfaction of educational, biomedical, and social behavioral researchers about justice issues (procedural justice, interpersonal justice, bias, and pro-science sensitivity) and other IRB characteristics (competence, outreach, formal functioning, structure, composition, and upholding the rights of human participants). Analysis of the responses by concerns and types of research conducted revealed significant main effects for domains with justice issues rated more important than other issues. There was no significant main effect for type of research, but social–behavioral researchers assigned greater importance to justice issues than did biomedical researchers.

24.1.4 Practitioner and Classroom-based Research

Simons and Usher (2000) outlined four general considerations as ethical principles are applied to situated inquiries: challenges to universality, sociopolitical dimensions, fairness in disadvantaged contexts, and the diversity of approaches in education research. Maguire (2004) stated, "Whatever the location, the important message that resonates is that researchers need to take into account the effects of their research on participants, on public discourse, and on policy makers" (p. 815). Pritchard (2002) addressed some of the difficulties facing researchers and REBs regarding practitioner–researcher dual roles in teacher research, practitioner inquiry, action research, and reflective practice. He especially considered the purposes of nonpublished, informed practice required by professional certificate and employment, and published knowledge-building "broadly [referred] to the array of activities people carry out as they seek knowledge or understanding while pursuing or improving a social practice in which they regularly engage" (p. 3). The ethical considerations are related to the participants and informants in these activities—not to the researcher's intent. In such cases, ethical approval of the research into the professional activity encounters difficulty when these activities are enacted in the workplace and involve clients, students, and colleagues who become the central foci of ethics review.

Pritchard (2002) unpacked the internal dimensions related to practitioner research issues and identified the following as central ethical considerations: (a) informed consent and free choice; (b) education misconception involving power-over and value of, and to whom; (c) procedural changes, responsiveness, and flexibility; (d) contingency for opportunistic and unexpected results; (e) preserving anonymity or confidentiality of participating institutions and informants; and (f) conflict and reform within the research institution, host organization, and participants. Furthermore, he analyzed the obstructions to effective and efficient address of these ethical issues by review boards. He stated:

> [a]sking questions, slowing things down, demanding to be appeased … [results in negative impressions about] the time and effort needed to assemble IRB submissions, respond to IRB requests, and work through whatever modifications on which the IRB insists. [The] IRB's appetite for paper seems voracious. (p. 7)

An inspection of any online or hard-copy templates for ethics approval will reveal very lengthy, complex applications for rather mundane issues. He concluded that IRBs were (a) overloaded with applications, (b) underresourced to handle the workload, (c) ill-informed about the specifics of the research under consideration, (d) focused on the common rule, (e) limited by their interpretations of the rules, (f) overly concerned about insignificant and improbable risks, (g) emphasized protecting the reputation of the institution and research enterprise, and (h) involved in ethical conflict. He suggested concrete improvements for operations, effectiveness, and efficiency of IRBs including enhanced resources, improved expertise and education outreach for board members and applicants, flexibility, and systemic adjustments and reform.

Some research ethics policies allow for local adaptations and interpretations to address unique features. Unfortunately, the local option can be used to include inappropriate requirements that are not central to the ethical treatment of participants in research. McDonald (2004) pointed out that sometimes all three ethical principles—respect for persons, beneficence, and justice—cannot be fully addressed independently and that resolution may involve maximizing compliance across the collective principle—the ethical treatment of research subjects. For example, informed consent may be unrealistic, therefore "the researcher must take on all the risks entailed in research participant protection, [since] there is no easy use of informed consent to off-load responsibility for research harms on to research participants" (p. 817). Sieber (2007) pointed out that the respect for personal and informed consent was evolving constructs and required innovative procedures: “Perhaps it is time to start thinking outside of the box” (p. 2).

Lopus and colleagues (2007) surveyed university professors’ perceptions about ethics review requirements involving students as participants in classroom-based research. They found ethics approval was required:

> in cases which present only minimal risks, and when the investigation is intended for evaluation of teaching approaches only, and not for publication. … [A] logistic regression analysis [of web-based survey responses] identifies the time it takes to complete the review application, the time it takes to receive a response, and the necessity of revising a project as significant factors in the respondents viewing the REC process as a barrier to research. (p. 69)

They believed such negative experiences with low-risk, classroom-based research could be minimized by applying alternatives available within the regulations of the institutions studied.

24.1.5 Community-based Research

Another growing and innovative area of research not well served by common rule regulations and biomedical-dominated interpretations of research is community-based participatory research, which is more of an orientation rather than a specific research method and about real-time design rather than a priori delineation of problem and procedures. Shore (2007) stated:

> Community-based research has multiple meanings depending upon one’s perspective. For some, it may signal that the research is situated within a community setting and does not speak at all about the degree of participation that the community has in the research process. For others, it signals a dynamic relationship between academic investigators and community representative in carrying out the research. (p. 31)

The relationship is critical when (a) the definition of community-based research (CBR) switches from setting or target to partner and (b) the design process becomes collaborative and responsive, more like technological design than scientific inquiry in which the procedures are dynamic and respond to current events in determining the next step. This approach focuses on community as agent of change and participation to address social justice issues and where all partners learn from one another and express civility and value of one another’s contributions and resources. “The community partners are recognized as having expertise through their [lived] experiences and insider knowledge regarding the culture of the community [and its knowledge stores], while researchers often possess research-related skills” (p. 32). This operational definition of CBR does not fit the regulatory definition of research in the US National Research Act of 1974 since it tends to focus on sociopolitical actions, not knowledge claims; or place-based assertions, not generalizable claims; or the researchers become advocates, not objective participants. However, Shore believed the generalizability issue could be addressed by focusing on the application of place-based claims to other places, communities, and situations.

Much of the insights into ethics and research involving human subjects in a community-based context and partnership must be gleaned from practice as this is a newly emerging area of research. In Canada, there are several approaches toward the development of standards evident. The first is individual, growing out of the personal experience of pioneers of this approach. The collection of papers edited by Leadbeater et al. (2006) includes a series of case studies that report on the ethical dilemmas in designing CBR along with an array of responses from researchers and community partners. The authors are circumspect in generalizing from individual case studies. Nonetheless, in a concluding paper, Moretti and colleagues (2006) reminded prospective community-based researchers that:

> [w]hen we launch community-based research, at least two systems come into contact-and sometimes collide: the university's system and that of the community under study. Each system comes with its own history and procedures for identifying and resolving problems, as well as its own beliefs, hopes, and fears as they relate to the process and outcome of collaboration. (p. 234)

As individuals and institutions become more familiar with issues in CBR, there have been attempts to coalesce the individual cases into intuitional guidelines.

There are procedural and ethical challenges with CBR related to community approval, informed consent, and confidentiality and anonymous participation. Social justice can involve the traditions and operations of the community partner in which the research target focus is on vulnerable, subordinate, or less powerful members of the community. There is general acceptance that ethics approval procedures need to recognize the potential involvement in the injustice of the more powerful members of the community—who, therefore, are in a conflict of interest when it comes to supporting and approving the research inquiry. Furthermore, communities like families, rural schools, and religious groups are tightly defined contexts in which confidentiality is difficult to maintain; therefore, anonymous status of informants–participating partners is highly unlikely.

24.1.6 Aboriginal and Indigenous Participants and Their Knowledge Claims

Respect for persons, beneficence, and justice, which are the fundamental principles of research ethics, and the central goal of research ethics approval—to protect participants from physical and emotional harm—applied to vulnerable, aboriginal, and indigenous participants need to consider political authority, individual and social histories, and cultural factors. In some countries, aboriginal and indigenous peoples are afforded *nation* status; their role must be recognized and infused into any approval or evaluation process. A variety of cultural, linguistic, epistemic, and ontological factors across several minority, aboriginal, and indigenous peoples when considering knowledge about nature, natural occurring events, and science literacy have been documented (Yore, Chinn, & Hand, 2008). The social history of people that led to their current state needs to be remembered since a lack of awareness can perpetuate the same mistakes made earlier. Therefore, the research ethics approval process needs to reflect and protect the value, traditions, and conventions of host participants–partners and clearly recognize their history and their intellectual property rights. Furthermore, unlike traditional scientific inquiries, this type of research frequently involves community-based participatory approaches where ongoing deliberations and adjustments to methods and dissemination are part of the design (Glass & Kaufert, 2007). Most research ethics regulations are based in a biomedical framework, scientific worldview, and inquiry model of the dominant culture and are lacking consideration of alternative worldviews, epistemologies, and cultures.

Glass and Kaufert (2007) attempted to access the unpublished, gray literature of aboriginal and nonaboriginal researchers regarding research ethics. They believed that current research ethics policies not only reflect a Western scientific worldview but also were "based on western liberal democratic political traditions protecting individuals, [and they] place great weight on individual autonomy and … self-determination" (p. 26). Unfortunately, some research ethics policies did not reflect on historical factors and prior engagements between cultures. They stated:

> Aboriginal leaders have become more critical of both past and ongoing research and are interested in playing a more active role in projects within their own communities. They also set a high priority on whether a research project is culturally appropriate and respectful of local knowledge. Key questions for many communities are whether the research assists in building local capacity and is potentially able to solve [problems] the community itself identifies as [priorities]. In many cases, communities have articulated their concerns and are ready and able to participate in [the research ethics review process]. (p. 27)

The emerging interpretation of approval and consent in aboriginal communities normally requires community review or consent and provisions for control and ownership of data and knowledge claims. The need to include others in the review process and to share authority requires reinterpretation of funding agencies' and universities' policies, procedures, and practices. Similar deliberations and policies have occurred in Australia to reflect the indigenous rights of aborigines and Torres Strait Islanders, in Canada to the status and rights of the First Nations, in the United States to reflect the indigenous rights of Native American Indians and aboriginal Hawaiians, in New Zealand to reflect the indigenous rights of the Māori people, and in southern Africa in recognition of the diversity of indigenous cultures in that region.

Gadicke (2005) conducted a research and development project about traditional knowledge and technologies related to water in the Columbia River Basin in Canada. Her ethics approval and development activities fully recognized that she was a guest in the Ktunaxa Nation with limited and respectful access to their stories and knowledge about water and technology. Furthermore, she recognized the territorial boundaries and cultural diversity across the geographic area and the various peoples of the Columbia River. Her approved uses of the traditional knowledge and technologies were for a specific purpose and audience reserving ownership to the First Nations involved.

24.1.7 Best Practices

International experiences with research ethics regulations, implementation, and REB practices revealed varying degrees of satisfaction and the general need for monitoring and adjustments to these policies, structures, and practices in literacy and science education research. Best practices should be a goal of any deliberation and investigation of research ethics involving humans (Sieber, 2006). Keith-Spiegel and colleagues (2006) stated, "The ideal ethics committee appears to be a just body that employs fair procedures, treats investigators with respect, and accords them the opportunity to have a voice when disagreements arise" (p. 78). They suggested that consideration of client service, proactive measures, staff and board members' professional development, and effective communications will improve researchers' perceptions of research ethics and IRBs and may, in fact, improve an institution's research program.

The IRBs and RECs of professional associations and funding agencies should promote thoughtful reflections and empirical investigations into the fundamental foundations, critical principles, operational procedures, and research quality (Sieber, 2006). It appears (a) as if the central focus of research ethics is not always central to IRB procedures and practices, and (b) that IRB actions assign greater risks than actually exist, focus on legal exposure, and privilege some a priori research designs over responsive designs intended to reflect and react to contextual variables and real-time events. Effective IRBs need to stay focused on the central goal "to ensure the ethical treatment of research subjects [and the fundamental ethical principles of] respect for persons, beneficence, and justice" (Pritchard, 2002, pp. 7–8). Levine (2006) believed that IRBs are:

> losing [their] effectiveness in safeguarding the rights and welfare of human subjects [in] that IRBs devote too much time doing work that simply does not need to be done. Several routine practices of IRBs are highly time consuming and, in [his] opinion, not sufficiently productive to warrant their continuation in their present form. (p. 1)

He suggested that these activities and procedures should be empirically evaluated and the results of such inquiries should be used "to persuade federal regulators and other policy-makers to reduce the burdens on the IRBs in a rational manner" (p. 1).

Lopus and colleagues (2007) believed best practices need to develop policy and procedure that expedite review for minimal-risk classroom research and exempt evaluations that are not to be published. Improvements to the ethics review process "with respect to classroom-base studies and others that impose virtually no potential harm to human subjects" (p. 70) and will likely increase the amount of research done. Current policies do not impose a barrier. Rubin and Sieber (2006), along with Lopus and colleagues, pointed out that such expedited reviews are allowed under the US regulations and could be conducted within the disciplinary boundaries in which the research methodological expertise is likely to exist.

Shore (2007) believed best practice involving emergent and responsive designs needed to involve an ongoing progressive process, not a singular event or evaluation. IRBs need to become aware of innovative inquiries and build or recruit expertise among their members to ensure informed deliberations and decisions. She suggested that the three fundamental ethical principles need to be elaborated to include ethics of partnership building, empowerment, self-determination, liberty, and social action. Glass and Kaufert (2007) stated:

> Best practices should include a mandatory formal agreement at an early phase of the relationship between the community authority (aboriginal or non-aboriginal) and the investigators detailing issues of data ownership, interpretation/analysis and publication, with specific mechanisms for managing conflicting interpretations or inappropriate use of data. Parties should agree in advance on their roles and responsibilities, desired outcomes, measures of validity, control of the use of data, funding and channels to disseminate findings. The guidelines or policy statement should protect both researchers and participating communities for unreasonable restriction on access to data or the right to publish findings. (p. 37)

Without such best practices and assurances, it would be inappropriate to expend public research funds or to involve graduate students and untenured faculty members in CBR, research involving aboriginal or indigenous participants, or other innovative research designs.

24.2 Critical Issues

The review of related literature, presentations, and deliberations at the 2nd Island Conference on research ethics and the 2006 National Association for Research in Science Teaching Research Committee-sponsored symposia identified several critical issues related to research ethics, IRB procedures and practices, and Gold Standard-quality research in literacy and science education. These issues involve various configurations of single and multiple policies; local interpretations; local panels; vulnerable, aboriginal, and indigenous peoples; practitioner research; futuristic considerations promoted by the US National Research Council (US NRC, 2002, 2004) to share and enhance the use of datasets, secondary analyses, computer-assisted analysis systems; and other interesting issues (see Yore & Boscolo, Chap. 2). We found a range in the development of research cultures, support for quality research, and research ethics in the inquiries leading to this chapter. Some countries and regions have well-developed policies, procedures, and systems in place to facilitate and support high-quality, ethical research practices. Others do not, leaving researchers to depend on their personal values, beliefs, and knowledge. We find the variation places additional demands on journals and professional and accrediting associations to ensure research ethics.

An example of the *one-size-fits-all* is the Canadian Tri-Council policy that is designed to integrate ethics reviews from medicine, natural sciences and engineering, and humanities and social science research under one policy (CIHR, NSERC, & SSHRC, n.d.-a). Unified IRB policies appear to focus on risk management as much as ethics oversight and thereby assume and assign high risk to all inquiries. Some policies reflect a privileged design (experimental–control design) because this design provides a priori hypotheses, procedures, and data sources while other interpretative and contextual designs reflect a technological approach that responds and reacts to events in real time.

Local panels with different interpretations of the research ethics policies and regulations have limited research experience and expertise with some high-quality alternative designs to the Gold Standard randomized controlled trials (RCTs) approach. A consensus has emerged through discussions that there is:

- Tendency of panel, chair, and staff to focus on risk in every application.
- Tendency to require risk management, limit institutional exposure, and use legal language in information letters and consent forms that convey a higher level of risk than actually afforded in the research proposed.
- Tendency to not consider readers, audience, and potential participants with information and consent forms—immigrants and low-proficiency English/domain language.
- Tendency to not respect cultural norms and societal traditions regarding authority within the community and school in approval process, especially in cross-cultural and international research studies.
- Tendency of local panels to overstep their charge to include research design issues.

24.3 National Perspectives

The following brief perspectives from Canada, New Zealand, southern Africa, Taiwan (Republic of China), the United Kingdom, and the United States illustrate some of these critical issues related to codes of research ethics, REBs, and the Gold Standard(s) for literacy and science education research. Some key issues embedded in these codes of research and professional practice are (a) the dual roles of professional practitioner and researcher, (b) ownership of data and interpretations, (c) recruitment of participants, (d) informed consent, (e) termination of involvement, (f) cultural and indigenous rights, (g) confidentiality and anonymity, and (h) future and unforeseen uses of data. Each of these factors manifests its influences on the development and conduct of research in explicit or implicit ways.

24.3.1 Canada

There are three major, government-sponsored granting agencies in Canada: Canadian Institutes for Health Research, Natural Sciences and Engineering Research Council, and the Social Sciences and Humanities Research Council—collectively referred to as the Tri-Councils. These councils had been independently monitoring ethical guidelines and procedures; but in 1994, the Tri-Council Working Group was developed; its final report established the guidelines that govern ethical reviews in all postsecondary and research institutions in Canada. The Tri-Council Policy Statement (TCPS, CIHR et al., 1998) serves the regulatory function of an ethics code. All institutions that receive funding from any of the granting agencies are required to adhere to the principles and processes outlined in the TCPS. The TCPS ensures centralized authority over every research project in the country that involves human participants through the approval process of the institutional REBs. It has resulted in a burgeoning of an ethics bureaucracy throughout Canada's research infrastructure. However, some features have been identified as in need of review; and an Interagency Advisory Panel on Research Ethics has been struck with purpose of conducting wide-scale consultations with the research communities with the goal of bringing forward proposals for revision (CIHR et al., n.d.-a).

The following sections outline some of the disjunctions between researchers and REBs based on a brief overview of the Canadian experience from the perspective of educational researchers at the University of Victoria. This perspective is focused on the key issues established earlier dealing with a one-size-fits-all ethics policy on the diversity of research, in particular on qualitative and CBR traditions. For example, a policy focus on risk and potential for legal exposure in every application demands complex legalese in information and consent communications with potential participants and very likely lacks respect for cultural and professional norms that are also present in a research context. Further tensions emerged between research applicants and the REB when the approval process called into question issues of research design. The impact of these issues was approached through a year-long process of meetings and negotiations involving the REB and a group of educational researchers in an attempt to collectively develop guidelines for an area that had been identified as particularly problematic, that is, teachers as researchers in their classes. The University of Victoria case study is informative about the potential for a process to arrive at a consensus of perspectives. This case study also provides insight into the ontological and epistemological contrasts that underlie the principles and practices of REBs and the power relationships that are exercised between the scholarly concerns for the design of research and the ethical concerns of REBs.

The motivation for addressing the underlying issues for research in educational settings was especially pertinent as large numbers of graduate students undertake the role of teacher–researcher in their own classrooms while conducting action research and reflective practice. There are three key issues in this case study that reflect upon the more general issues related to the relationship of research and REBs: first, the overlapping responsibilities of graduate research advisers and the REB for oversight of the quality of the research design, in particular, exploring the separation of scholarly concern for the most efficacious research design to be applied from the interests of the REB; second, the problem of distinguishing the dual roles of teacher and researcher in the classroom (see Coupal, 2004); and third, the marked gap in the familiarity and experience of actual classrooms between REB members and the teacher–researchers conducting the research. These contrasting perspectives were especially evident in the interpretation of the *power-over* relationship. For practitioners, there was a clear recognition of the authority (ministry of education, school, teachers' union) of established codes of ethics that govern the ethics of the teacher–student relationship in the classroom while the REB explicitly disregards such professional codes of practice and holds to a different conception of power-over students in the classroom. The contrast in these perspectives is fundamental; the expectation of teachers that students are expected to participate in classroom activities sanctioned by the school curriculum versus the REB's expectation that such participation must be voluntary for research purposes.

Classroom-based teacher research has tended over the last decade to be qualitative in design. This may be a reflection of the enormous diversity between educational settings that inhibits more controlled types of research or simply a reflection of the preferences of the community of educational researchers. In either case, the relationship between REBs and qualitative research has been seen as "an unhappy union" (Ells & Gutfreund, 2006). Whether this unhappiness arises from the TCPS or the various applications of the TCPS is a matter for ongoing discussion (Ells & Gutfreund; McGinn, 2005) of such general concern that the Interagency Advisory Panel in Canada has undertaken a separate consultation document on the issues (Blackstone, 2007).

The process of consultation between the University of Victoria educational researchers and REB does not reveal either a unique or novel approach, other than the critical importance of researchers' active participation and stewardship regarding all components of the research enterprise: quality, funding facilitation, and ethics. Rather, it is another example of the potential of the adage: first you talk, then talk, talk again, and finally talk some more. Over a score of meetings and a dozen draft versions of a guideline, consensus was gradually achieved. The progress of the discussion relied upon the participants' dedication to reach a new level of understanding of common objectives and regard for contrasting viewpoints. The initial guidelines that emerged were recently reviewed and expanded beyond the context of classroom-based research to embrace all dual-role research–practitioners. The guidelines' purpose is:

> to assist graduate students and their supervisors in the Faculty of Education and other applied or professional faculties to better understand some of the specific challenges of practitioner-researchers undertaking research in professional/classroom settings and to outline recommended approaches to ensure that the study to be undertaken involves procedures that are consistent with the current ethical standards of research practice outlined in the Tri-Council Policy Statement on Ethical Conduct for Research Involving Humans (TCPS). (University of Victoria Human Research Ethics Office, 2008, para. 1.1)

The guidelines provide guidance around some of the previously perilous situations that delayed or deflected dual-role research. There is clarification of the scope of responsibility around the researcher's focus on research design and the REB's scrutiny of ethical research, which acknowledges the researcher's primary responsibility for design. This legitimizes use of dual-role research to explore matters of concern to researchers.

The achievement of these guidelines represents a case study in the collaboration of the REB and researchers toward the mitigation of potential areas of conflict. Anthony (2004) reported a case where two collaborating researchers submitted individual applications to conduct the same study design in separate classrooms; one application was approved without revision, the other rejected. The decision of the REB was justified on the basis that different members had reviewed each application. Even though the review policies and guidelines were the same, the decisions were not. While it can be appreciated that individuals hold differing perspectives on professional and research practices, it is also apparent that these differences may lead to decisions that are irreconcilable with the application of a common set of principles. This case study also provides insight into the ontological and epistemological contrasts that underlie REB principles and practices and the power relationships that are exercised between the design of research (academic supervision of research) and the review of research (ethics review board). The supervisors and teacher/researchers collaborated on an understanding of the ethical standards for these parallel studies, but the reviewing members of the REB were not bound by a common understanding of the ethical considerations for the research.

At the University of Victoria, the model from the dual-role practitioner guidelines to clarify differing views of research and research ethics is being explored for CBR, another class of research. CBR is meant as an umbrella term that is inclusive of terms such as collaborative research, participatory research, action research, and participatory action research. Like the process for developing guidelines for dual-role practitioner research, a group of interested researchers initiated a consultative process with the REB about concerns related to CBR. Through this consultative process, another guideline for CBR is arising (Bannister, 2008). The issues under consideration are the interplay between the social action agenda of CBR, which calls for shared responsibility in formulating the research agenda and for conducting and disseminating the results of CBR in an ongoing and collaborative manner, and the expectation of REBs for researchers to disclose the details of research design before ethical approval. Extensive community consultation and collaboration often results in *emergent* research designs where the details of the research process develop throughout the study and are not known to either researchers or participants at the outset. Such emergent projects may involve activities that are not initially viewed as research activities. For example, a scholar may be collaborating with a community to consider advice about how to respond to a community need and, in the process of developing an awareness of the need, there is information gathered that later emerges as research data, which requires ethical approval.

In CBR, the local knowledge and expertise of community participants is often considered integral to the research and learning goals, processes, and outcomes. Such an approach changes the balance of power that is typically assumed by REBs (Bannister, 2008; Coupal, 2004; Minkler, 2004). It also raises considerations about rights, responsibilities, and ownership over processes and outcomes. An example is sorting out ownership of the intellectual property that might arise from research that involves the traditional knowledge of a First Nations community (CIHR, 2007; Schnarch, 2004). There is an expectation at the University of Victoria that indigenous community approval will be obtained for certain research, such as when the research specifically involves or includes individuals from an indigenous population or a particular indigenous community will be a central focus. Clearly, the policy landscape for university research involving aboriginal peoples is in an unprecedented state of uncertainty amid dynamic change at national and institutional levels.

The landscape for REBs is in a process of dynamic review and reconsideration. This is particularly the case as the authorities responsible for developing policies on ethical review struggle with the challenge of including qualitative and emergent research designs within the same review process as traditional forms of controlled research. The one-size-fits-all approach to ethical review in Canada offers both the promise of consistency and equity of ethical standards and a large measure of complexity and uncertainty as the vast research landscape is threaded through the eye of a national policy. The Tri-Councils conducted an open consultation regarding qualitative research in the context of the TCPS in 2007 (Canadian Interagency Advisory Panel on Research Ethics, 2006), and it is expected that a new TCPS will be announced by the end of 2008.

24.3.2 New Zealand

Ethics in educational research is administered by local institutional REBs in New Zealand universities and other postsecondary institutions. There is at present no national policy, system, or authority that controls the conduct of such boards. Ethical issues in educational research are, however, bound by a variety of national legislative requirements regarding privacy and freedom of information; and REBs are expected by their institutions to ensure that the conduct of educational research meets these statutory requirements. The most relevant legislation is the New Zealand Privacy Act (NZ Government, 1993) and the Copyright Act (NZ Government, 1994), although some science educational research projects may make aspects of the Health and Safety in Employment Act (NZ Government, 1992) and the Resource Management Act (NZ Government, 1991) relevant (e.g., surveys of public views about land use or sensitive commercial development projects).

Education researchers are expected to observe copyright issues, with raw data generally considered to belong to participants and interpretation of raw data to belong to researchers. All research is expected to protect the identity of schools, students, teachers, and other participants. Informed consent is a key issue with all participants expected to provide written consent on forms that spell out in detail the nature of the research and the commitment required of them. Participants are asked to allow use of raw data for analysis and interpretation and use of interpreted data in publications and presentations, consistent with the copyright and privacy considerations mentioned above. Addressing these issues would satisfy most legislative requirements under New Zealand law.

Many of the educational research ethics issues in New Zealand are fairly innocuous because of the nature of the research projects. New Zealand educational research at present does not involve much in the way of large-scale, quantitative, interventionist, or experimental techniques. Hence, some issues like sample selection and ethical issues associated with experimenting on students, teachers, and classes are seldom of major concern. Any large-scale, quantitative work would likely be government-initiated and endorsed. Although ethical issues would be subject to scrutiny (e.g., by the ministry concerned), ethical approval would be subject to the local REB of the researcher's agency involved in a given research contract.

Much educational research in New Zealand is case study or interpretive in nature. This takes two forms: exploratory case-study research seeking to understand educational issues or to explore educational issues identified in surveys in more depth. Interviews, either one-on-one or focus groups, is the usual method of choice. Other than protection of identities (as dictated by the Privacy Act) and ownership and use of data, there are not major ethical issues. Such studies also may involve classroom observation although this is less common due to resource constraints (but ministry-based contract research frequently involves classroom observation, depending on the contract). Classroom observational research frequently involves the more invasive modern technologies (e.g., videotaping) and puts the researcher potentially in conflict with the Privacy Act. Hence, even with informed consent, it is seldom deemed appropriate to share video or digital recordings at conference presentations, professional development workshops, or for future research projects.

The second form of educational research now common in New Zealand is action research, often by teachers doing postgraduate study and research projects as part of career or professional development. Again, these research projects are typically small-scale, interpretive-based projects. As action research projects are interventionist in nature, one might expect them to address similar ethical issues to that of experimental studies. However, there is an interesting difference between the New Zealand educational system and that in many other nations. The New Zealand educational system underwent dramatic and far-reaching changes in the 1980s and 1990s. In brief, there was significant devolution of school management including curriculum. There is no national curriculum as such but instead a curriculum framework (NZ Ministry of Education [MoE], 1993a, 1993b) and a series of educational curriculum statements (NZ MoE 1993c, 1996a, 1996b, 1996c, 1996d) that indicate broadly what is to be learned and what achievement objectives must be met, similar to the European tradition that places much more responsibility on the classroom expert—the teacher. The development, evaluation, and implementation of the school curriculum are thus the school’s responsibility, and the flexibility embedded in this system is intended to result in a highly learner-centered education system. Recent research suggests this is indeed the case (NZ MoE, 2002). What this means in terms educational research, particularly action research, is that a teacher has the right—indeed, even an obligation—to alter pedagogy to meet the needs of learners. Hence, teachers conducting action research projects do not need participants’ permission to enact interventions. However, they do need to seek consent for data gathering (e.g., interviews) that would not be part of a change to pedagogy and must address the other research ethics issues identified above (e.g., protection of school and student identity, use of data gathered, etc.). The use of these data and interpretation as the basis for professional and academic publication, including theses in university libraries, would require consideration by the host school authority.

24.3.3 Southern Africa

The development of policies and practices related to research ethics is relatively new in southern Africa and has been largely dominated by concerns about international collaboration and the urgency of research related to HIV/AIDS. While the Medical Research Council in South Africa has been developing guidelines on ethics for medical research since 1997, South Africa's national ethics regulations in the area of biomedical research were enacted within the National Health Act of 2003 (SA Government, 2003, Chapter 9) and elaborated in the national ethics guidelines in 2004 (SA Department of Health, 2004). Standards for ethical biomedical research along with regulations governing the establishment of RECs have emerged. However, ethics policies in the social sciences and humanities are trailing behind and are far less well developed. It has been reported that, where they exist, these committees work without formal legislation and merely follow a set of limited guidelines developed at the individual institution (Johns Hopkins Berman Institute of Bioethics, 2007; Louw & Delport, 2006).

Roberts (2006) reported on several key themes in research ethics in Africa. These included policy differences between countries and institutions, which reveal inconsistencies in the application of best practice for achieving a balance between resonance with global standard practices and consideration of unique elements to adequately address local circumstances. In order to present some sense of the current state of research ethics in southern Africa, four policies that guide the research ethics were selected from different institutions.

24.3.3.1 University of Pretoria

Louw and Delport (2006) suggested that the University of Pretoria ethics policy outlines and establishes a structure by which applications for ethical approval are submitted and reviewed and where identified challenges are resolved. A significant challenge for REC members is to familiarize themselves with the array of documentation within which the committee functions, namely the Constitution of South Africa Act of 1996, the Copyright Act of 1978, the Promotion of Access to Information Act of 2000, the Promotion of Justice Act of 2000, the Research Ethics Guidelines (SA Department of Health, 2004), and the Code of Ethics for Research (University of Pretoria, n.d.). This complex array of regulations greatly encumbers the REB; as a result, its actual operation depends largely on the recommendation of individual department ethics committees (Louw & Delport). The role of the Pretoria Committee is not only to evaluate research proposals but also to educate and assist the faculty to understand, appreciate, and apply the ethics of research (Benatar, 2002).

24.3.3.2 University of Cape Town

The Faculty of Humanities at the University of Cape Town has issued general research ethics guidelines for its departments and schools (University of Cape Town, 2006). These guidelines invest initial approval with the departments and schools but provide a flow diagram for appeals that shows concerned researchers paths to follow when a research proposal has not been approved. It illustrates that every research topic has to gain the approval of a departmental REC before the ethics approval application can be considered at the next level. If a departmental REC fails to reach an agreement or the researcher disagrees with its decisions or disputes the methods used, the applicant is given the option of either reformulating or changing the research topic, design, and application or appealing to the Faculty REC for reevaluation.

24.3.3.3 University of Botswana

The University of Botswana (University of Botswana, 2004) policy on ethics and ethical conduct in research aims to establish (a) codes of practice for research and consultancy activities, (b) mechanisms for ensuring compliance with the ethical standards and values of the university and with the international research society and civil society, and (c) the framework for developing and implementing codes of conduct for ethical behavior. This policy is cross-referenced to other university policies and procedures related to academic honesty, staff disciplinary regulations and procedures, research and development, and intellectual property. The research ethics document lists activities that are deemed to be unethical behavior, such as fabrication or falsification of data, plagiarism, conflict of interest disclosure, authorship, use of research funds, and safeguard of human rights. One issue explicitly mentioned is deception involving the researcher's failure to give potential subjects information that may lead to their refusal to participate in the research. The University of Botswana provides a sample code and principles for individual disciplinary-specific departments in their development, implementation, and regular review of policy, procedures, and practices within the university policy. However, there are further restrictions existing outside of the university; for example, no anthropological research can be undertaken without the approval of the government as stipulated in the Anthropological Research Act of 1976.

24.3.3.4 Human Science Research Council

The Human Science Research Council (HSRC) is the major funding agency for scientific research in South Africa and includes research from the natural sciences, engineering, and social sciences. HSRC has produced a code of ethics aimed at monitoring research that is undertaken with public funds (SA HSRC, n.d.). HSRC provides a mission statement that commits the agency to funding and promoting research to the benefit of all people in South Africa and to supporting societal goals. Furthermore, HSRC suggests that (a) research supported by public funds belongs to the public domain and must withstand public scrutiny, and (b) researchers seeking public funding must honor the trust placed in them and respect the rights and dignity of participants.

The HSRC guidelines include the following principles: respect and protection, transparency, scientific and academic professionalism, and accountability. The principle of respect and protection emphasizes that the pursuit of knowledge should not override the consideration of participants' personal, social, and cultural values and that the research must respect the participants' autonomy, protect their well-being, and obtain informed consent. The principle of transparency emphasizes the need for participants to be clearly briefed on the aims and implications of the research outcomes. The participants have to be continuously kept in the loop concerning the process and progress of the research. The principle of scientific and academic professionalism explicitly accepts the role for codes of conduct outlined and accepted by membership in professional and research associations, the use of status and position for personal benefit, and the goal to achieve quality research and justified results. The principle of accountability requires that research be conducted with and not on identified communities. The researcher should provide potential participants the written focus, conditions and terms, potential deliverables, their commitments, and time schedule for the research; this document will clarify involvements and likely lead to successful completion of the research and quality results. These research ethics and procedural expectations are monitored by a committee composed of leading researchers and HSRC staff members.

Several challenges have been identified regarding the application of research ethics principles that emphasize the need for ethics guidelines and the promotion of high ethical standards in southern African contexts. These challenges require researchers to be cognizant of the far-reaching ethical implications of the sociocultural contexts. Louw and Delport (2006) argued that research in the South African context is especially influenced by cultural and linguistic factors. These factors pose ethical problems with regard to the principles of respect of persons, justice, and beneficence. According to them, the respect of persons is jeopardized when obtaining genuine informed consent by using interpreters—especially when the researcher has limited knowledge about the social systems, cultural values, and beliefs of potential participants. Even written consent remains contentious in the South African context due to low literacy levels.

Some southern African contexts are strongly anchored in the cultural and religious beliefs of the people. For example, a study conducted by researchers in the Department of Chemistry at the University of Swaziland (Amusan, Dlamini, Msonthi, & Makhubu, 2002) on traditional medicines revealed that the people's practices are clouded with secrecy, myths, and metaphysical powers. The participants involved in the study had a strong belief in ancestral spirits, which made it difficult to interpret the data in scientific terms. They found that these data could only be understood when considered within the cultures of the people (Makhubu, as cited by Amusan et al.). Makhubu (1998) argued that participants who were traditional healers felt vulnerable and unprotected since they lacked legal recognition. This lack of legal status put the ownership of their medicines and indigenous knowledge in question. Even without the CBR label, there are ethical issues around the integration of the nonscientists–researchers' knowledge, ownership, and interpretation of data included in the research report.

Louw and Delport (2006) further observed that in the southern African context the ethical principles of beneficence and respect could be violated by the use of measurements (e.g., standardized tests) that are culturally inappropriate as well as lacking validity due to language differences. A further issue is conflicts that arise between traditional methods of knowing, learning, and teaching and those imported from colonial powers (McKeever, 2000). Worldviews and their related views of reality, epistemological beliefs, and ontological assumptions need to be considered and respected as outside researchers gain access, engage traditional knowledge, and make these ideas from a different interpretative framework (Yore et al., 2008).

Most researchers are advantaged in comparison to research participants. This advantage and associated power difference are potentially problematic for application of the principle of justice. Louw and Delport (2006) observed that researchers need to manifestly address this principle because "[t]he political legacies of the apartheid era may still be operating in a given situation and researchers need to be aware of the cultural dynamics and the potential impact on their research endeavors" (p. 60). For example, when addressing ethical issues in conducting educational research in a postcolonial context, McKeever (2000) raised the issue of whether she, as a white person, had a right to research black experience.

The combination of need for research and limited resources makes international collaboration essential for most researchers in southern Africa. Such a situation results in distinctive considerations for the development of ethical standards. On the one hand, increased international research collaboration leads to a consideration of the value of ethical pluralism. On the other hand, collaborative research benefits from clear and explicit, ethical guidelines that are consistent with international standards, which present the specter of ethical imperialism. Benatar (2002) argued that new ways of thinking about the role of RECs is required in developing countries in order to promote progress in authentically grounded research, which may involve hybrid policies and procedures that achieve a balance between established international practices and unique policy elements in consideration of local needs.

24.3.4 Taiwan (Republic of China)

Taiwan has a well-developed research culture in its universities, research institutes, and development centers that has led its modern economic growth in science and technology. The development of research ethics followed a similar track as Western countries, starting in human scientific studies (medical science, biology, etc.) and gradually spreading to other research areas (science education, psychology, sociology, economics, etc.). The development of research ethics reflected the cultural traditions and different emphases, priorities, and relevance of the human benefits and costs. Mature practices and thorough procedures can be found in the medical sciences, biotechnology, and biology; whereas in other research areas, similar policies, practices, and procedures are only starting to evolve. Huge differences and gaps exist at the national, institutional, and individual levels for research ethics in different academic and professional organizations.

Formal research ethics regulations and procedures did not emerge at the explicit level until the last decade when serious and formal concerns were initially expressed in biotechnology and medical sciences. Before that, research ethics were not an issue for most disciplines in the academy. Basically, researchers followed the principles of goodwill and self-regulation, which means research ethics were maintained at their own discretion and with respect to their personal beliefs, professional values, and positive intentions. There were no clear rules for researchers to follow, no official forms to complete, and no standard operating procedures to take. Since the public highly values academics and the traditional thinking was that most of the research studies were for the public good and welfare, there was no urgent need to establish rules and procedures to regulate research practices. In 1997, a slight change occurred when the Public Health Agency issued the Guidelines for Good Clinical Practice (GCP; Shih, Shih, Chen, & Chen, 2005) initiating a series of reactions inside Taiwan's medical research communities that subsequently spread to other academic fields.

The Academia Sinica (TW Academia Sinica, 2007), Taiwan's most prestigious research institute for sciences and humanities, has been instrumental in leading the considerations of research ethics. Informed consent, even if the law requires participation, appears to be a basic principle (Bryman, 2001). The Academia Sinica formed the Human Subject Research Ethics Committee/IRB in 2004, which in turn established regulations and ethical guidelines to conduct research on human subjects in accordance with the Declaration of Helsinki (WMA, 2004) and the Belmont Report (US NCPHSBBR 1979). The Academia Sinica (TW Academia Sinica, n.d.) stated:

> Use of an informed consent document is an important component of the informed consent process. To assure truly informed consent by subjects, the consent document information should be presented in non-technical language that subjects can understand. If the document is not understandable, a claim could be made that the participant did not really know what they agreed to participate in.
>
> To increase the chances that the informed consent document will be understood by most subjects, it is recommended that investigators: write at no higher than an eighth-grade reading level; use simple, straightforward sentences; use commonly recognizable terms and measurement amounts; avoid the use of jargon or technical language; and explain terms that may not be easily understood. If non-Chinese speaking subjects will be enrolled, plan to translate informed consent documents. Likewise, if illiterate or visually-impaired subjects will be enrolled, plan to provide witnessed verbal translations of the informed consent document. (§ Readability of the Informed Consent Document)

The regulations and guidelines for researchers address three basic ethical principles:

> Underlying the federal regulations, state statutes, and University policies for human subject protection are three principles. They are: autonomy, beneficence, and justice.
>
> The principle of autonomy requires us to respect each individual's right to decide freely whether or not to enroll in research.
>
> The principle of beneficence requires that investigators attempt to 'do good' or, conversely, 'do no harm' in the conduct of their research.

> The principle of justice requires that access to research must be equitable, meaning that the risks of research should not disproportionately be borne by the disadvantaged and the benefits of research should not be reserved for the privileged.
>
> The principles are described in detail in a document known as the 'Belmont Report' which is available http://ohsr.od.nih.gov/guidelines/belmont.html online here. (§ Guiding Principles for Human Subject Protection)

Early developments inside the medical sciences and associated public concern forced the government to establish clear regulations. The revised Medical Care Act of 2003 (MCA) required that only teaching hospitals (allied regional medical centers, usually sponsored by respected universities) can conduct clinical trials and that proposals for clinical trials must be submitted to a human research committee composed of medical technologists, law experts, and social workers. This requirement was first enforced by the National Health Research Institute in 1999 and later by the Department of Health, the Executive Yuan in 2000, and the National Science Council in 2001 (Kuo, 2001). In addition, consent forms were required to state specific information, including the objectives and methods of the experiment, possible adverse effects or risks, expected results, alternative treatments; participants could withdraw from the study at any time.

The chief editors of major academic journals were placed under great pressure to attend a series of seminars and courses on the international ethics codes or standards. A 2000/01 survey of all 66 chief editors (65 responded) found that they agreed about the importance of the IRB review and that participants' consent, risk–benefit assessment, and justice in selecting human participants were necessary for intervention studies regulated by the MCA (Shih et al., 2005). Moreover, Shih and colleagues also found that chief editors were more positive toward policies regarding non-MCA regulated intervention studies than were other physicians. However, the actual practice of research ethics was not as encouraging; only 5 (9.1%) required IRB approval of studies involving human participants as a prerequisite for publication. Furthermore, 42 (64.6%) did not present any information on human research ethics or legal protection of human participants in their instructions for submission; 18 (27.7%) mentioned the Uniform Requirements for Manuscripts Submitted to Biomedical Journals (URMSBJ) (see http://www.icmje.org/ for more information); 7 (10.8%) required privacy protection; 1 (1.5%) referred to the Declaration of Helsinki; and 4 (6.2%) simply indicated that participants' consent should be obtained in the journals' guidance to authors. However, the situation is changing in that it is a common requirement for every paper submitted to these journals to have passed the IRB review that normally requires participants' written consent, risk and benefit assessment, and to follow URMSBJ.

A survey with Delphi technique on research ethics in Taiwan was completed in 1999 (Yang, Kuo, Chen, & Chou, 2001). The questionnaire followed the design of Cabana and colleagues (1999) to investigate participants' knowledge, attitude, and practice (test–retest reliability = 0.84). Results from the 172 respondents (400 public health researchers were surveyed for a return rate of 43%) showed that 70.6% agreed that although subjects signed the consent forms it might not really express their willingness to be tested in an experiment. Furthermore, 92.9% of participants considered the importance of confirming the subjects' comprehension of the information and research involved; and 52.8% revealed that as long as researchers provide reasonable explanations, research subjects' oral consent was acceptable. Close to half of the participants (48.3%) agreed that there is no need to have reviews from the IRB if a study is carried out in an educational environment related to educational methods or assessments. However, 68% of participants did not agree that as long as the government conducted the study it did not need to go through the review process.

Science education in Taiwan has never been regulated by laws, public policies, or guidelines for experimental (research) practice; nor is there a written consent requirement for human participants. Researchers differ about how to address research ethics issues. Universities are not consistent in requiring a review similar to those conducted by the IRB for studies in medicine and biological sciences. The requirements for research ethics are recognized as important in the sciences; and science education should not ignore this issue since participants' safety, privacy, and deception are equally important in educational research. The learning records of students are very important datasets for some educational and sociological studies. However, these data are private; and owners of this information would not want other people to have access. Traditions and historical practices are difficult to change, especially in a hierarchical society like Taiwan. This means there are several critical questions about these datasets:

- Who actually owns these data?
- How should guidelines and procedures for accessing and using these data be established?
- Do acceptable procedures for accessing, sharing, and using these data exist?
- Should the collectors of these data be afforded unrestricted use?

These are the important issues that science education needs to explore if participants' rights and welfare are to be ethically addressed since secondary analysis and data sharing are likely to become pressing issues.

An analysis of research studies published from 2001 to 2007 in the *Chinese Journal of Science Education*, Taiwan's top science education journal, found that very few papers mentioned research ethics for participants. Figure 24.1 demonstrates the pattern of consideration related to participants' safety, consent, and privacy (0% for deception). Harm, privacy, and deception were not consistently mentioned; only informed consent was explicitly mentioned across the volumes of the journal with the level of consideration increasing from about 10% (2001) to 21% (2007) of the articles. In those cases, the consent forms were from the teachers—not from the students who were the real participants. Historically, researchers requested permission and signed consent forms from the principal and teachers but did not require students to complete consent forms or provide verbal agreement. It appears to indicate a cultural tradition in education where students were viewed as *possessions* of the schools, principals, and teachers in the early years of the survey.

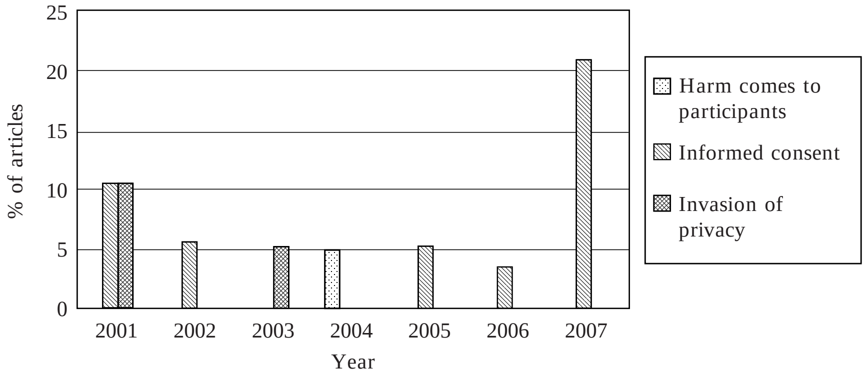


Fig. 24.1 Percentage of *Chinese Journal of Science Education* articles indicating harm, informed consent, and privacy

Publishing research results and the public display of participants' work, places additional ethical demands upon the researchers and especially teacher–researchers. Teacher–researchers occupy dual roles that place different ethical demands on information obtained from students and other participants. Information collected to improve learning and classroom practice does not require extraordinary ethical consideration other than those of caring teachers' regard for student safety and welfare under their professional standards and code of ethics. Using the same information in other professional and academic settings (teacher workshops, conferences, etc.) and purposes (graduate theses or dissertations, journal articles, commercial resources) goes beyond the normal approvals afforded teachers. Therefore, researchers must consider: (a) the integrity in writing research reports or papers while presenting the data and doing data analysis; (b) the ethics of sharing the findings with other members of the research team or graduate supervisor; (c) the fair contributions of each person on the research team or potential coauthors; (d) the ethics of sharing data with other participants; (e) the acknowledgment of participants' contribution to research; (f) the appropriateness and appreciation of sponsors and funding agencies without implying endorsement; and (g) the proper procedures, importance, and requirements for credit among the contributors (participants, research assistants, and researchers). In other words, integrity of doing and publishing research should be taken seriously by researchers and graduate supervisors.

The Declaration of Helsinki (WMA, 2004) stated that the basic principles for medical research are:

> Both authors and publishers have ethical obligations. In publication of the results of research, the investigators are obliged to preserve the accuracy of the results. Negative as well as positive results should be published or otherwise publicly available. Sources of funding, institutional affiliations and any possible conflicts of interest should be declared

> in the publication. Reports of experimentation not in accordance with the principles laid down in this Declaration should not be accepted for publication. (#27)

The ethical choices of researchers as well as procedural decisions are reflected in quality research. Respect for and trust of participants and their data, ethical interpretation of data and sharing research findings, acknowledgment and credit of colleagues, and attributing the success to the right persons, situations, and treatments are fundamental ways to demonstrate integrity and ethics. The fundamental motives for researchers (knowledge builder and teacher) to explore authentic problems and to seek solutions and insights are to make a difference for the current participants and to provide insights for future generations.

Generally speaking, due to their cultural history, Asian intellectuals are afforded high respect from the public; therefore, conducting research with civilians and students seems relatively more convenient in Asia than in Western countries. However, this does not mean that Asian scholars have the right to take advantage of their high status and use or abuse this privilege. Therefore, necessary respect for participants should be taken as the first priority while collecting data. These data carry with them the same respect as the identity of the informants. Researchers must respect participants and protect them from being harmed in the research—data collection, data interpretation, and public display of the results. Society's expectations of integrity, discipline, and self-regulation must be recognized and honored.

24.3.5 United Kingdom

Ethical issues in educational research have been the subject of debate and discussion for decades in the United Kingdom. The seminal book *The Ethics of Educational Research* includes chapters on the ethics of feminist educational research, school-based research, case-study research, and educational ethnography (Burgess, 1989). Burgess identified the key ethical issues in educational research as involving sponsorship, relations between researchers and participants, informed consent, and data dissemination. Simons (1989) addressed the question of whether guidelines could be produced for educational researchers and evaluators, a question that Burgess described as fascinating.

The British Educational Research Association (BERA, 1992) formally adopted ethical guidelines at its 1992 Annual General Meeting (AGM). The history of the guidelines can be traced to a March 1988 invitational seminar convened by the noted researcher John Elliott (Furlong, 2004) that focused on the monitoring of research contracts (Simons, 1995). Simons noted that a code of ethics had been proposed at BERA's inaugural AGM in 1974 but had been rejected. She hypothesized that the code was rejected because of the "possible disrepute of professional code of ethics which may be self-serving of professional interests rather than underpinning values in the public interest" (p. 441). The report of the seminar, entitled Towards a Code of Practice for Funded Educational Research (Elliott, 1989), did not lead immediately to the adoption of ethical guidelines (Simons). In 1991, the BERA Council invited two senior educational academics to propose a set of guidelines, which were adopted at the 1992 AGM (BERA, 1992). This action in the United Kingdom followed similar action in the United States by a few months. Simons suggested that the reason that the BERA adopted the ethical guidelines was due "partly, at least, to the increasing politicization of the research culture" (p. 441).

The politicization continued over the next decade and criticism and debate within the research communities ultimately led to the revision of the guidelines. In the current guidelines' introduction, the BERA (2004) claims that "[a]s a code of practice the guidelines were universally welcomed but also attracted a degree of criticism in relation to their scope and application" (p. 3). A working group of three academics began the task of reviewing and revising the original guidelines; after extensive consultation, the Revised Ethical Guidelines for Educational Research were presented to the 2003 AGM and formally adopted by the BERA Council the following year. In the preamble to the revised guidelines, BERA noted that they "are offered as [a] set of principles and advice [that are not designed to] selectively judge or constrain, directly or indirectly, the methodological distinctions or the research processes that emanate from them" (p. 5). The set of principles includes respect for the person, knowledge, democratic values, the quality of educational research, and academic freedom.

The guidelines are set out under three headings of the researchers' responsibilities to participants, sponsors of research, and community of educational researchers. The 41 responsibilities distributed under these three headings include voluntary informed consent, deception, the right to withdraw, children, vulnerable young and vulnerable adults, incentives, detriment arising from participation, privacy, disclosure, methods, publication, misconduct, and authorship. In general, the guidelines make little reference to external bodies or laws. However, when discussing deception or subterfuge, BERA (2004) "recommends that approval for this course of action should be obtained from a local or institutional ethics committee" (p. 6). When discussing children, vulnerable young people and vulnerable adults, BERA "requires researchers to comply with Articles 3 and 12 of the United Nations Convention on the Rights of the Child" (p. 7; United Nations, 1990). Finally, the guidelines insist that "Researchers must comply with the legal requirements in relation to the storage and use of personal data as set down by the Data Protection Act (1998)" (p. 9).

Simons (1995) noted that the research community welcomed the original guidelines but they "start[ed] from a deficit [due to contextual factors of the period and their adoption] raises further questions about their potential impact and efficacy in the current political climate. … Had they been introduced 20 years ago, their impact might well have been different" (p. 441). She cautioned that, at a time when the UK government stood accused of dubious practices involving the use and publication of education research, the focus of researchers was not on:

> what guidelines we can establish to ensure the best conditions for research but what defensible ethical procedures we can devise to try to ensure that the research we conduct can raise questions independent of political agendas, without political interference in the process and without findings being censored or publication restricted. (pp. 441–442) (also see Ham, 1999)

Simons continued by suggesting that guidelines—such as “Funding bodies should not be allowed to exercise restrictions on publication by default, e.g. by failing to answer requests for permission to publish, or by undue delay” (BERA, 1992, p. 4)—were written as a response to past government actions and with an eye to what was seen as an even more hostile future. She noted that researchers, faced with a sponsor who refuses to publish their work, can adopt several strategies including leaking their findings, publishing letters and articles in the press anonymously, and getting questions asked in the Houses of Parliament. The revised guidelines shifted the emphasis somewhat; and the BERA (2004) held that “[t]he right of researchers to publish the findings of their research under their own names is considered the norm for sponsored research” (p. 11), but then listed six exemptions—including when “[r]esearchers have waived this right in writing” (p. 11)—that allow sponsors to own rather than simply to commission research.

Educational researchers may also find the ethical guidelines published by the British Sociological Association (BSA, 2002) and the British Psychological Society (BPS, 2008) of interest. However, Simons (1995) noted that the BSA and BPS guidelines “were not necessarily seen as appropriate for the relatively recent discipline of educational research that focused on studying education in its own right” (p. 448); these guidelines are still used by many educational researchers.

Despite the existence of the BERA, BSA, and BPS guidelines, the ethical hoops that UK educational researchers had to jump through were barely systematized compared with the situation in other countries. Researchers from some countries often found the somewhat laissez-faire approach to ethical approval exhibited by some UK universities both curious and somewhat disturbing. However, the situation has changed with the recent publication of the Research Ethics Framework (REF) by the major funding agency of social science research, the Economic and Social Research Council (ESRC). The REF, which took effect formally on January 1, 2006, states that the ESRC will only fund research “where consideration has been given to ethical implications, and in those institutions where appropriate arrangements are in place” (UK ESRC, n.d.-a, para. 1). In a sublime piece of understatement worthy of Crick and Watson, the ESRC noted that “[t]he Framework will therefore have implications for applicants to ESRC, research ethics committees within HEIs [higher education institutions] and for those assessing research proposals” (para. 1). This is particularly true as all other main funding agencies of social science research in the United Kingdom support the REF. (For interested readers, background papers relating to the history and background to the REF can be found on the University of York’s website http://www.york.ac.uk/res/ref/documents.htm.)

The interdisciplinary and interagency context of the new framework can be gleaned from the statement that it “is also conscious of the increasing importance of interaction between the social sciences and the natural and medical sciences and the new challenges that these are creating in sensitive areas such as genomics and stem cells research” (UK ESRC, n.d.-b, para. 3). The ESRC noted the importance and need for guidelines and standards that were designed by and for the social sciences research community, rather than the continued adoption and adaption of those established for researchers in medicine.

The REF identifies six key principles of ethical research that must be applied:

- Research should be designed, reviewed, and undertaken to ensure integrity and quality.
- Research staff and subjects must be informed fully about the purpose, methods, and intended possible uses of the research, what their participation in the research entails, and what risks, if any, are involved. Some variation is allowed in very specific and exceptional research contexts for which detailed guidance is provided in the policy guidelines.
- Confidentiality of information supplied by research subjects and anonymity of respondents must be respected.
- Research participants must participate in a voluntary way, free from any coercion.
- Harm to research participants must be avoided.
- The independence of the researcher(s) must be clear, and any conflicts of interest or partiality must be explicit.

The REF is a high-stakes instrument. The ESRC (UK ESRC, 2006) warns that breaches of:

> good ethical practice … will be treated as a very serious matter by the Council. They could result in the immediate suspension of the individual project and other projects based at or under the co-ordination of the contracting institution, and a halt to the consideration of further applications from that institution. (p. 2)

While not seeking to impose a single model and set of procedures, the ESRC (UK ESRC, 2006) "will ensure that its peer review of proposals addresses ethical issues, and engage in dipstick testing of institutions with awards to check that commitments to ethical review have indeed been followed through by institutions" (p. 2). The implication of external audits (dipstick testing) is that the main funding body for social science research in the United Kingdom does not fully trust universities to carry out good ethical practice or avoid conflict of interest in rendering ethics approval of research proposals. This point is further emphasized by the statement: "Before the start of a project, funds will not flow until the administering institution provides written confirmation that the required ethical approval has been received" (p. 2). A further indicator is evident in arrangements for expedited ethical approval in cases "where the potential for risk of harm to participants and others affected by the proposed research is minimal" (p. 3). Expedited review "is carried out by one or more members of a Research Ethics Committee (REC), commonly its chair, and not by a member of the Department due to carry out the research" (p. 3). The ESRC also mandates that "a REC must have at least one academic member from outside the Department conducting the research and at least one appropriately trained lay member" (p. 3). However, there is recognition of the relationship between a researcher's professional (BERA, BSA, BPS, etc.) ethics standards, codes, and guidelines and the REF by this statement: "In the first instance, it is the responsibility of the researcher, or research team, guided by their professional disciplinary standards, to decide whether a project is ethically sensitive" (p. 7).

An institutional response to the new climate within which universities are working can be judged by King's College London's ethics approval system. For educational researchers (which might include all students at magisterial and doctoral level), the first step is to decide on the level of risk that the potential research might involve to participants. This risk assessment involves answering six questions, for example, "Could the study induce psychological stress or anxiety, or produce humiliation or cause harm or negative consequences beyond the risks encountered in normal life?" (King's College London, n.d.). If the answer is yes to any of the questions, then the applicant must apply through the Social Sciences, Humanities, and Law Research Ethics Subcommittee. If the potential risk is assessed as moderate or uncertain, then the application is reviewed by a REB. Most undergraduate and masters students are able to follow a low-risk procedure, which allows expedited consideration of applications.

Some indication of the shift in the importance of ethics in research is the recent focus on ethics in the UK educational press. A recent article in *The Times Higher Education Supplement* began: "Ethical considerations may not be at the top of your priorities when developing a research proposal. But, […], your pet project could have an unforeseen impact on some participants" (Swain, 2006, para. 1). The article continues with an example of how casual some researchers were about seeking ethical approval:

> You've just dashed off an application form to the university's ethics committee and told them to relax. No issues to worry about and consent's certainly sorted out. Those schoolgirls you use in your studies are always dead impressed by the idea that you're a [professor]. (para. 2)

Swain then proceeds to give advice about how to get ethical approval for research. A professor who chairs a research ethics committee at a large UK university comments that "universities have to start promoting a culture of ethics so that when people come to fill in these forms and read instructions they understand the issues underpinning the form and what's wanted" (para. 19). That such a comment should be made in 2006 gives some indication of the prevailing culture with respect to ethical approval for social science research. The report comes with a warning from one of the panel that drew up the REF "getting something through an ethics committee can easily take more than 18 months, especially if revisions are needed" (para. 21).

While the ethical issues involved in doing educational research are broadly the same as they were in the 1980s, the standardization of ethical approval in UK universities has changed beyond recognition. The increasing internationalization of research and the growth in the awareness of the rights of the individual have led to the major funding agency of educational research introducing a research ethics framework that has forced universities to adopt high visibility and what are perceived as heavily bureaucratic systems of ethical approval. It remains to be seen what impact the new procedures will actually have on the education research community, but one thing is certain—the changes are irreversible.

24.3.6 United States of America

In June 1992, the American Education Research Association (AERA) adopted and published its Ethical Standards in *Educational Researcher* (AERA, 1992). The American Psychological Association (APA) followed in December 1992 with its Ethical Principles of Psychologists and Code of Conduct, which were revised in 2002 and effective in June 2003 (APA, 2002). (For interested readers, a comparison of the 1992 and 2002 APA ethical principles and codes of conduct is provided at http://www.apa.org/ethics/codecompare.html, showing line-by-line changes.) The 2002 APA document covers a wide variety of principles and conduct to fully embrace the professional activities of psychologists in practice and research situations: resolving ethical issues, competence, human relations, privacy and confidentiality, record keeping and fees, education and training, research and publications, assessment, and therapy. The principles and codes involving research and publication (#8) are worthwhile to literacy and science education researchers, especially the sections on deception, publication credits, duplicate publication of data, and peer reviewing.

The AERA guiding standards recognized that:

> educational researchers from many disciplines, embrace several competing theoretical frameworks, and use a variety of research methodologies. …The standards [are meant to] remind us that we are involved not only in research but in education. It is, therefore, essential that we continually reflect on our research to be sure that it is not only sound scientifically but that it makes a positive contribution to the educational enterprise. (p. 23)

The six guiding standards address responsibilities to the field, research populations, educational institutions, and the public; intellectual ownership, editing, reviewing, and appraising research; sponsors, policy makers, and other users of research; and students and student researchers. Each major standard was elaborated with 3–12 more explicit standards to guide members' ethical practices in designing and doing quality research and their academic conduct in research environments. Strike and colleagues (2002) provided a series of cases associated with these standards as professional development tools to enhance awareness and improve conduct of educational researchers.

The federal code to protect human subjects provides the foundation for ethics review in the United States (Protection [45 CFR 46], 2005). But the code of research ethics and REBs are not the only consideration in research design and conduct. The recent mandate for scientifically valid research in education and the reorganization of the US Department of Education and establishment of the Institute for Education Sciences (IES) raised serious issues and concerns for both educational practitioners and researchers. On the practical side, for example, educational program and curriculum developers—many of whose services and materials are already widely used by schools—are scrambling to find the expertise and resources needed to evaluate their products in order to meet the requirement of being research-based. Similarly, community and other private, nonprofit, educational organizations situated outside of the university system are not only dealing with the need to conduct evaluation research—for which they may be ill-equipped—but also with the need to find an IRB to review and approve their research plans to ensure that they adequately protect research participants.

Although the present upheavals may ultimately be justified in terms of improved educational practices, there is another, more disturbing aspect to these demands for scientifically rigorous educational research. This is the fundamentally antiscientific nature of these political mandates. The RCT is deservedly accorded the status of being a Gold Standard for answering certain types of questions, but it is not the most appropriate or most rigorous approach to answering all scientific questions—including important questions about program effectiveness. Elevating the randomized experiment to its present status as the standard for producing scientifically important information (with its cousin, the quasi-experiment, begrudgingly tolerated as a distant but at least minimally acceptable alternative) has privileged one scientific paradigm and a subset of the available tools of scientific inquiry. This privileged status is unwarranted and unjustifiable in some research situations, problem spaces, and research questions. Given that this standard is to be applied across the board in the provision of federal funding for the conduct of educational research, the result is to preemptively exclude large areas of legitimate, important, scientific research from consideration for support. To borrow from Elliot Eisner, our demand for scientific rigor is in danger of becoming associated with rigor mortis. Research ethics and review procedures need to reflect the full range of quality research approaches and ensure that they facilitate quality innovative approaches to address the range of critical problems and questions facing literacy and science education.

24.4 Special Considerations: International Students and Indigenous Peoples

Within the general principles of research ethics, each of the national perspectives from Canada, New Zealand, and the United States have special constitutional considerations, policies, or laws regarding research ethics dealing with special classes of research subjects, such as international students and their cultural values and the nations' founding peoples and their knowledge. Established policies and guidelines that regulate research involving Alaska Natives, Australian Aboriginals, First Nations people, and Native Americans include: Alaska Federation of Natives (AFN) Guidelines for Research (AFN, 1993) and Guidelines for Respecting Cultural Knowledge (Assembly of Alaska Native Educators, 2000); Code of Research Ethics developed with the Native Mohawk community of Kahnawake in Canada (Macauley et al., 1998); the Model Tribal Research Code developed by the American Indian Law Center Inc. (AILC, 1999); the guidelines of the Australian Institute of Aboriginal and Torres Strait Islander Studies (Australian Institute of Aboriginal and Torres Strait Islander Studies, 2000); the US Basic Health and Human Services Policy for Protection of Human Research Subjects

(US Department of Health and Human Services, n.d.); and the Principles for the Conduct of Research in the Arctic (US Interagency Arctic Research Policy Committee, 1995). There are growing efforts to afford similar consideration to aboriginal Hawaiians and other indigenous peoples. These concerns and related actions have grown out of past effects of colonization and unauthorized access and use of indigenous people's knowledge, customs, and cultural artifacts (Yore et al., 2008). The following briefs attempt to surface some of the considerations and how these special issues are addressed for international students and indigenous peoples.

24.4.1 International Students and Education Research Ethics

The research center at the University of Waikato in New Zealand has a large number of international students. The educational issues brought by these students add an interesting dimension to the research activity, but at the same time the different educational systems and cultural practices result in some interesting ethical issues. The most common issue is that of seeking informed consent. For New Zealand-based research, informed consent is a must. Participants must know what they consent to and have the right to withdraw from any research project at any time without giving reasons. This is not the case in the educational context for many international students. It is common, for example, like the past practices in Taiwan for school students, for the dean of a teacher training program or officials from the ministry of education to give blanket approval of a research project and essentially require students or teachers to participate in the research as directed by the researcher. In such cases, researchers go along with the official and cultural norms of the particular educational context but insist on adherence to other ethical practices, such as use of information and protection of identities.

24.4.2 Indigenous Peoples and Education Research Ethics

Aborigines, First Nations, and Indigenous Peoples in Canada, New Zealand, and the United States require special consideration when exploring their education, culture, and traditional knowledge systems. At the University of Victoria in Canada, separate research ethics and procedures have been developed with the First Nations regarding inquiries into their culture and their knowledge claims. This requirement is based upon the TCP involving health sciences, humanities and social sciences, and natural sciences and engineering funding agencies that specifically addresses research with aboriginal people (CIHR et al., n.d.-b). The dialogue between REBs and researchers who focus on other areas of research will necessitate continuing consultation and clarification. Discussion of such guidelines can be anticipated to continue not only with researchers with regard to the ethical standards of research but also with First Nations communities who have their own concerns and priorities. The University of Victoria Human Research Ethics Board (HREB) requires that any researcher contemplating a study that includes indigenous peoples complete a separate Indigenous Community Approval section in addition to the standard ethics application. The conditions that govern this approval remain loosely defined:

> Indigenous community approval may be required when the research involves Indigenous people from a community (whether residing in urban or reserve areas), the cultural knowledge and/or resources of Indigenous people, or where individuals speak on behalf of an Indigenous nation. (University of Victoria HREB, 2008, item G)

The CIHR (2007) has proposed guidelines prepared in conjunction with its Institute of Aboriginal Peoples' Health to assist researchers and institutions in carrying out ethical and culturally competent research involving aboriginal people. The intent is to promote health through research that is in keeping with aboriginal values and traditions. The tone of this document is clearly intended to represent an aboriginal perspective on research. This is signaled in the acknowledgment to the proposal:

> The members of the Aboriginal Ethics Working Group (AEWG) would like to acknowledge the Creator and those who came before us without whom this document could not have been written. We also acknowledge the hard work of the many individuals, communities and organizations that generously provided input to this document. In particular we would like to acknowledge the contribution made by the Kahnawake Schools Diabetes Prevention Project with their Code of Research Ethics (www.ksdpp.org). We understand that the English and French languages do not always allow Aboriginal concepts and world views to be effectively communicated across cultures and we do not wish to offend with words that have been written. We do encourage continuous dialogue as Aboriginal ethics are articulated within an academic research context. (p. 11)

REBs, researchers, and potential participants in research face the requirement of deciding whether the research falls into this special category that the TCP and local REBs have identified. Related to this are questions of community. In urban settings, members of many different indigenous groups may be included in research or a small number of indigenous peoples may be included in a larger research study. For such circumstances, the scope of obtaining community consent remains to be clarified. Policies and practices addressing these issues in Canada and the United States vary across the First Nations and Native Americans in specific regions, since negotiations have been between individual indigenous authorities resident in the region.

The reemergence of Indian American self-determination and self-governance in the United States has required research sponsors to consult with tribes, tribal organizations, and national Indian associations, agendas, and guidelines for research focused on Native American issues. Some of the relations between the community and researchers apply to the Alaska Native and Hawaiian contexts as they seek self-determination and protection of their culture, language, rights, and indigenous knowledge. Moreover, in order to be sensitive to the legitimate problems of these communities and for research to have a beneficial impact on these communities, it is necessary that the researcher be familiar with cultural ways and beliefs of the tribes and establish a social relation with members of the community

(Alaska Native Knowledge Network, n.d.; AILC, 1999). This personal relation between the researcher and the subjects implies qualitative research methods, which are sometimes at odds with current research policy and funding agencies. In many cases, cultural conditions posed by the changing distribution of indigenous populations conflict with the ethical, methodological Gold Standard for educational research thereby delaying resolutions and approval of REBs. Paradoxically, this conflict is putting at risk research in areas with critical need of improvement and jeopardizing answers that could be beneficial for the stakeholders of indigenous educational issues.

New Zealand also is in the unusual position, for a previous colony at least, of having a founding document—The Treaty of Waitangi—that underpins much legislation. The Treaty is an agreement signed by the Crown (in the form of the British colonialist governor) and the Māori people (New Zealand's first nation or indigenous peoples). The Treaty itself is actually rather brief and vague in its original form (Treaty of Waitangi, 1840). However, any legalization is expected to adhere to the principles of the Treaty. As one might imagine, this is open to interpretation. Some, for example, take this to mean every governmental authority must have Māori representation or at least consult with Māori on virtually any issue. To illustrate, any Marsden Fund application (New Zealand's premier *blue skies* research fund) must have a suitable statement if the research is deemed relevant to, and cognizant of, the position of Māori—what is termed Māori Responsiveness. The position taken by the Royal Society of New Zealand (2005) on Marsden fund applications illustrates the issue:

> Māori Responsiveness
>
> The Marsden Fund Council acknowledges its obligation to operate the Marsden Fund, Te Pūtea Rangahau a Marsden, in accordance with the Treaty of Waitangi. In order to give effect to its commitment, the Council seeks to achieve greater Māori participation and leadership in Marsden research and, where research projects involve issues of significance to Māori or have significant Māori content, requires that applicants are in consultation with Māori.
>
> The requirement for consultation is not intended to deter researchers but to ensure that the research is well planned, that appropriate etiquette is observed when access to Māori sites, culturally sensitive material and knowledge is sought from their owners, and that Māori intellectual and cultural property rights are respected. As a first step, researchers should seek advice from their institution, many of which have established processes for consultation with Māori.
>
> Consultation with Māori is not expected, and may not be appropriate, for proposed projects where no specific interest for Māori can be identified. (p. 7)

This statement might seem mild, but it means that few applications for science or science educational research would not require consultation, given that almost anything in New Zealand is taken to involve or impact upon Māori. Few applications can afford to ignore such oblique directions.

Presently, there is no explicit requirement to consult with Māori or to have Māori representation on institutional REBs for research involving education; but there seems little doubt that this will eventually become part of the educational research landscape. The opposition to such a requirement is strongly opposed to what it sees as preferential status accorded Māori. Many, if not most, New Zealand schools have Māori children and caregivers. Hence in reality, research that involves schools may routinely involve Māori and potentially require consultation with Māori.

24.5 Closing Remarks

Researchers in literacy and science education and research culture in general accept the need for policies regarding ethics, honesty, integrity, and moral values. However, the various ways that academic administrators and REBs have implemented these policies and generated power structures lead us to raise common questions about several policies and practices that have been recognized worldwide. The central issues relate to codes for research practice and professional conduct that flow from shared values, beliefs, and assumptions about humanity, quality inquiries, and professional responsibility to society. From the collective position of the literacy and science education professionals represented in this chapter, these issues include but are limited to: the link between ethics and quality, the dual roles of educator and researcher; power-over relationships within the academy, professional organizations, and research setting; recruitment of participants; assessment and balance affordance of risk; informed consent, voluntary participation, and termination of involvement; cultural and indigenous rights; confidentiality and anonymity; and ownership of research data, artifacts, and interpretations. Most importantly these attributes are as much central to the *quality of research* as to ethical conduct (Strike et al., 2002; Zeni, 2001). Compliance with the fundamental principles of research ethics—respect, autonomy, and protection of the individual; beneficence of the educator–researcher to do no harm; and to demonstrate justice, fairness, and concern for the vulnerable—are standards that enhance the quality, worth, and creditability of any results flowing from research. Findings that arise from such research have greater likelihood of influencing policy makers and practitioners because of their epistemological integrity: moral and ethical foundation.

Strike and colleagues (2002) noted that epistemological integrity may also include differences of position regarding approaches to research:

> [W]hile intellectual integrity may involve conscientiously applying a self-chosen paradigm, it also seems to require that our paradigm be chosen for appropriate and good reasons. … [R]easonable and competent people often disagree about the appropriate approach to studying education phenomena in ways that have yet to be resolved by evidence and argument … [but], we should select the methodology that is appropriate to the questions we ask. (p. 11)

This recognition of variety in research approaches does not reduce the demand for procedural rigor, compelling arguments, and evidence-based knowledge claims about the problem space and research questions. Clearly, ethical standards can be expected to be no less controversial—requiring an equally diverse discussion of fair, rigorous, and consistent professional judgments and evaluations of research decisions, peer-reviews, and personnel assessments. Furthermore, such ethical standards not only apply to research participants' dignity, sensitivities, privacy, rights, and contributions but also require open, forthright, and broad dissemination of all research results and an appropriate recognition of creative contributions with shared authorship, institutional affiliations, and funding support.

The perspectives provided in this chapter illustrate similarities and differences across diverse research communities and academic cultures in literacy and science education: Canada, New Zealand, southern Africa, Taiwan, the United Kingdom, and the United States. But these perspectives are only a starting point to encourage and support research communities' development of ethical and supportive research cultures and to provide informed feedback to governmental policy makers, funding agencies, and university administrators. Some perspectives described here (United Kingdom—BERA code of ethics, United States—AERA and APA codes of ethics) have long track records of working with research ethics policies and procedures; others (Canada—TCP statement, Taiwan—Academia Sinica) are developing and amending policies, procedures, and practices; while others (New Zealand, southern Africa) are moving toward explicit policies. Research ethics policies are only the first step; the difficulties are in the implementation! The experience of literacy and science education researchers and their professional organizations needs to be applied to designing quality standards for research ethics. It is these researchers and representatives of researchers who have the greatest breadth of experience in applying standards to actual research. Surprisingly, an informal survey of association websites revealed that some literacy and science education research associations have not attempted to contribute their experience in applying ethical standards to research through the establishment of codes of ethics.

Research ethics need to be futuristic and reflect recommendations by expert panels and research associations regarding secondary uses of data and access to datasets by other than the primary researchers. Elsewhere in this book can be found encouragement of data sharing, secondary analysis of both quantitative and qualitative data, rigorous data collection and interpretation involving external reviewers and critics, and encouragement to move high-quality research results into the policy-making arena and instructional development process. These nontraditional uses of data and research results will need to be incorporated into ethics policies, applications, and review processes. This means that REBs and researchers must anticipate data sharing, secondary analysis, and multiple uses when seeking initial ethics approval for their research projects.

We have outlined other growing concerns from various perspectives to be addressed by REBs. These include the following nonexhaustive issues:

- Who is to invigilate the application of ethical standards for research? REBs generally have no monitoring function beyond the initial review of research applications. Will journal editors play an oversight role in monitoring research ethics? How might funding agencies audit actual compliance and conduct?

- Independent research and commercial research groups searching for legitimacy have made use of IRBs for review of for-hire and contract research. Questions of vested interest and limits on dissemination of results are not a prominent feature of existing REBs. How can commercial research be monitored?
- What are the unmapped areas involving community-based research? Without much imagination, one can foresee potential problems in CBR involving graduate students, faculties of graduate studies, and universities. What happens when a CBR team of community partners, graduate students, and faculty members encounters contentious results that the community does not wish to have published? What happens to the graduate student's dissertation? What about the untenured faculty member's potential publications?
- REB deliberations can be too labor-intensive if their charge and efforts are not precisely focused. The inhibiting conditions that led to the development of the University of Victoria Guidelines for Dual-Role Research (University of Victoria Human Research Ethics Office, 2008) and the situation reported in the United Kingdom of taking 18 months for approval of rather low-risk projects demonstrate the inappropriateness of applying the same review procedures across the risk spectrum.
- The protection of researchers and participants is a fundamental principle that can be intelligently applied to low-risk inquiries not requiring comprehensive review as well as the analysis and meta-analysis of public data, public figures as subjects of research, anonymous observations of public activities, and autobiographical approaches. REBs can adopt policies that acknowledge contexts that are clearly of such minimal risk that they are more appropriately considered separately from full review through expedited review processes and waivers.
- REB chairs and panels, professors, independent researchers, and graduate students need professional development regarding the intentions of research ethics, approval procedures, and applications (Strike et al., 2002). These might involve:

  - Risk assessment (low-risk, such as accepted classroom practices, should focus only on the use of data, intended use, and public display).
  - Time and effort savings from the approval process on low-risk projects can be devoted to improved research quality.
  - Bureaucratic structures and organizations need to focus on their charge and not wander into the problem-finding and design processes.
  - REB chairs and members must be selected from the best representatives of the research communities (active and productive researchers) with the appropriate motives.

The unreflective application of one-size-fits-all to problems in the high-risk areas of medicine, pharmaceuticals, military, and biotechnology have been found to override low-risk contexts of normal classroom and professional practices. The preoccupation with risk can instill unreasonable fear in potential participants through the use of complex, legal language in consent forms that is not reflective of the conventional nature of the research involved. The *big-stick* approach of funding agencies in the United Kingdom, the United States, and Canada has mandated a complex administration of REBs that are part of today's political environment and are unlikely to change unless literacy and science education researchers become proactive during the policy development and review processes.

References

Alaska Federation of Natives. (1993). *Alaska Federation of Natives guidelines for research*. Retrieved June 8, 2008, from http://ankn.uaf.edu/IKS/afnguide.html

Alaska Native Knowledge Network. (n.d.). *Resources for compiling and exchanging information related to Alaska Native knowledge systems and ways of knowing*. Retrieved June 8, 2008, from http://www.ankn.uaf.edu/

American Educational Research Association. (1992). Ethical standards of the American Educational Research Association. *Educational Researcher*, *21*(7), 23–26.

American Indian Law Center Inc. (1999). *Model tribal research code* (3rd edn.). Albuquerque, NM: Author. Available from http://www.ihs.gov/MedicalPrograms/Research/pdf_files/mdl-code.pdf

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct*. Retrieved May 31, 2008, from http://www.apa.org/ethics/code2002.html

Amusan, O. O. G., Dlamini, P. S., Msonthi, J. D., & Makhubu, L. P. (2002). Some herbal remedies from Manzini region of Swaziland. *Journal of Ethnopharmacology*, *79*(1), 109–112.

Anthony, R. J. (2004, October). Consistency of ethics review. *Forum Qualitative Sozialforschung/ Forum: Qualitative Social Research*, *6*(1), Art. 5. Retrieved from http://217.160.35.246/fqs-texte/1–05/05–1–5-e.htm

Assembly of Alaska Native Educators. (2000). *Guidelines for respecting cultural knowledge*. Retrieved June 8, 2008, from http://ankn.uaf.edu/publications/knowledge.html

Australian Institute of Aboriginal and Torres Strait Islander Studies. (2000). *Guidelines for ethical research in indigenous studies*. Canberra, Australia: Author.

Bannister, K. P. (2008). *Ethical considerations in community-university research and learning collaborations for the University of Victoria*. Unpublished discussion document for the Human Research Ethics Board, University of Victoria, British Columbia.

Benatar, S. R. (2002). Reflections and recommendations on research ethics in developing countries. *Social Science & Medicine*, *54*(7), 1131–1141.

Blackstone, M. A. (2007). Power dynamics and ethical practices governing artists and their research participants. *NCEHR/CNERH Communiqué*, *14*(1), 11–16.

British Educational Research Association. (1992). *Ethical guidelines for educational research*. Edinburgh, Scotland: British Educational Research Association/Scottish Council for Research in Education.

British Educational Research Association. (2004). *Revised ethical guidelines for educational research*. Retrieved June 5, 2008, from http://www.bera.ac.uk/publications/pdfs/ETHICA1.PDF

British Psychological Society. (2008). *Generic professional practice guidelines*. Retrieved June 9, 2008, from http://www.bps.org.uk/publications/prof-pract/prof-pract_home.cfm

British Sociological Association. (2002, March). *Statement of ethical practice for the British Sociological Association* Retrieved June 9, 2008, from http://www.britsoc.co.uk/equality/Statement + Ethical + Practice.htm

Brydon-Miller, M., & Greenwood, D. (2006). A re-examination of the relationship between action research and human subjects review processes. *Action Research*, *4*(1), 117–128.

Bryman, A. (2001). Ethics in social research. In A. Bryman (Ed.), *Social research methods* (pp. 475–486). London: Oxford University Press.

Burgess, R. G. (Ed.). (1989). *The ethics of educational research*. London: Falmer.

Cabana, M. D., Rand, C. S., Powe, N. R., Wu, A. W., Wilson, M. H., Abboud, P.-A. C., et al. (1999). Why don't physicians follow clinical practice guidelines?: A framework for improvement. *Journal of the American Medical Association*, *282*(15), 1458–1465.

Canadian Institutes of Health Research. (2007, May). *CIHR guidelines for health research involving Aboriginal people*. Retrieved July 14, 2008, from http://www.cihr-irsc.gc.ca/e/29134.html

Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, & Social Sciences and Humanities Research Council of Canada. (1998). *Tri-Council Policy Statement: Ethical conduct for research involving humans* [with 2000, 2002, 2005 amendments]. Retrieved June 5, 2008, from http://www.pre.ethics.gc.ca/english/policystatement/policystatement.cfm

Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, & Social Sciences and Humanities Research Council of Canada. (n.d.-a). *Interagency advisory panel on research ethics*. Retrieved June 6, 2008, from http://pre.ethics.gc.ca/english/

Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, & Social Sciences and Humanities Research Council of Canada. (n.d.-b). *TCPS: Section 6. Research involving Aboriginal Peoples*. Retrieved June 9, 2008, from http://www.pre.ethics.gc.ca/english/policystatement/section6.cfm

Canadian Interagency Advisory Panel on Research Ethics; Social Sciences and Humanities Research Ethics Special Working Committee. (2006, December). *Qualitative research in the context of the TCPS: A follow-up to the Giving Voice to the Spectrum report and a discussion paper*. Retrieved June 6, 2008, from http://www.pre.ethics.gc.ca/english/workgroups/sshwc/consultation07.cfm

Cohen, J. J. (2005). A word from the president: "Research integrity is job one". *AAMC Reporter,* (September). Retrieved from http://www.aamc.org/newsroom/reporter/sept05/word.htm

Coupal, L. (2004, October). Practitioner-research and the regulation of research ethics: The challenge of individual, organizational, and social interests. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, 6*(1), Art. 6. Retrieved from http://217.160.35.246/fqs-texte/1–05/05–1–6-e.htm

de Laine, M. (2000). *Fieldwork, participation and practice: Ethics and dilemmas in qualitative research*. London: Sage.

de Vries, R., Anderson, M. S., & Martinson, B. C. (2006). Normal misbehavior: Scientists talk about the ethics of research. *Journal of Empirical Research on Human Research Ethics*, *1*(1), 43–50.

di Norcia, V. (2006). The ethics in human research ethics [Guest editorial]. *Journal of Empirical Research on Human Research Ethics*, *1*(2), 1–2.

Elliott, J. (1989). Towards a code of practice for funded educational research. *Research Intelligence, 31*, 14.

Ells, C., & Gutfreund, S. (2006). Myths about qualitative research and the Tri-Council Policy Statement. *Canadian Journal of Sociology, 31*(3), 361–373. Retrieved from http://www.cjsonline.ca/cjsindex/vol31.html

Florence, M. K., & Yore, L. D. (2004). Learning to write like a scientist: Coauthoring as an enculturation task. *Journal of Research in Science Teaching*, *41*(6), 637–668.

Furlong, J. (2004, May). The 2008 RAE and beyond [From the President]. *Research Intelligence, 87*, 1–2. Retrieved from http://www.bera.ac.uk/pdfs/RI_87.pdf

Gadicke, J. M. (2005). *Integrating aboriginal knowledge into the elementary science curriculum*. Unpublished master of education project, University of Victoria, Victoria, British Columbia, Canada.

Glass, K. C., & Kaufert, J. (2007). Research ethics review and aboriginal community values: Can the two be reconciled? *Journal of Empirical Research on Human Research Ethics, 2*(2), 25–40.

Ham, V. (1999). Tracking the truth or selling one's soul? Reflections on the ethics of a piece of commissioned research. *British Journal of Educational Studies, 47*(3), 275–282.

Hoonaard, W. C., van den. (2006). Trends in Canadian sociology master's theses in relation to research ethics review, 1995–2004. *Journal of Empirical Research on Human Research Ethics, 1*(4), 77–88.

International Reading Association. (2008, March). *Code of ethics*. Retrieved May 31, 2008, from http://www.reading.org/association/about/code.html

Johns Hopkins Berman Institute of Bioethics. (2007, January 23). *NIH-funded case study: Research ethics committees in Africa report inadequate funding, staffing and training* [Press release]. Retrieved from http://www.bioethicsinstitute.org/web/module/press/pressid/91/interior.asp

Keith-Spiegel, P., Koocher, G. P., & Tabachnick, B. (2006). What scientists want from their research ethics committee. *Journal of Empirical Research on Human Research Ethics, 1*(1), 67–82.

Kennedy, D. (2006, January 20). Acts of God? [Editorial]. *Science, 311*(5759), 303.

King's College London. (n.d.). *Applying to the education and management panel*. Retrieved June 6, 2008, from http://www.kcl.ac.uk/research/ethics/applicants/sshl/panels/em/

Kuo, I. T. (2001). Introduction of human research committee in Taiwan [in Chinese]. *Newsletter for Research of Applied Ethics, 7*(19), 22–24.

Landwirth, J. (2006). [Letter to the Editor]. *Journal of Empirical Research on Human Research Ethics, 1*(1), 3–4.

Leadbeater, B., Banister, E., Benoit, C., Jansson, M., Marshall, A., & Riecken, T. (Eds.). (2006). *Ethical issues in community-based research with children and youth*. Toronto, Ontario, Canada: University of Toronto Press.

Lee-Treweek, G., & Linkogle, S. (Eds.). (2000). *Danger in the field: Risk and ethics in social research*. New York: Routledge.

Levine, R. J. (2006). Empirical research to evaluate ethics committees' burdensome and perhaps unproductive policies and practices: A proposal. *Journal of Empirical Research on Human Research Ethics, 1*(3), 1–4.

Lopus, J. S., Grimes, P. W., Becker, W. E., & Pearson, R. A. (2007). Effects of human subjects requirements on classroom research: Multidisciplinary evidence. *Journal of Empirical Research on Human Research Ethics, 2*(3), 69–78.

Louw, B., & Delport, R. (2006). Contextual challenges in South Africa: The role of a research ethics committee. *Journal of Academic Ethics, 4*(1), 39–60.

Macauley, A. C., Delormier, T., McComber, A. M., Cross, E. J., Potvin, L. P., Paradis, G., et al. (1998). Participatory research with native community of Kahnawake creates innovative code of research ethics. *Canadian Journal of Public Health, 89*(2), 105–108.

Maguire, M. H. (2004). Review of the book: *Situated ethics in educational research. Science Education, 88*(5), 813–816.

Makhubu, L. (1998). Bioprospecting in an African context [Essays on science and society] *Science, 282*(5386), 41–42.

McDonald, M. (2004). [Review of the book: *Danger in the field: Risk and ethics in social research*]. *Science Education, 88*(5), 816–818.

McGinn, M. K. (2005). Ethical and friendly researchers, but not insiders: A response to Blodgett, Boyer, and Turk. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research, 6*(3), Art. 37. Retrieved from http://nbn-resolving.de/urn:nbn:de:0114-fqs0503375

McKeever, M. (2000). Snakes and ladders: Ethical issues in conducting educational research in a post-colonial context. In H. Simons & R. Usher (Eds.), *Situated ethics in educational research* (pp. 101–115). London: Routledge/Falmer.

Minkler, M. (2004). Ethical challenges for the “outside” researcher in community-based participatory research. *Health Education & Behavior*, *31*(6), 684–697.

Moretti, M., Leadbeater, B., & Marshall, A. (2006). Stepping into community-based research: Preparing students to meet new ethics and professional challenges. In B. Leadbeater, E. Banister, C. Benoit, M. Jansson, A. Marshall, & T. Riecken (Eds.), *Ethical issues in community-based research with children and youth* (pp. 232–244). Toronto, Ontario, Canada: University of Toronto Press.

National Science Teachers Association. (2007, June). *NSTA position statement: Principles of professionalism for science educators*. Retrieved May 31, 2008, from http://www.nsta.org/about/positions/professionalism.aspx

New Zealand Government. (1991). *Resource Management Act*. Wellington, NZ: Government Printer.

New Zealand Government. (1992). *Health and Safety in Employment Act*. Wellington, NZ: Government Printer.

New Zealand Government. (1993). *Privacy Act*. Wellington, NZ: Government Printer.

New Zealand Government. (1994). *Copyright Act*. Wellington, NZ: Government Printer.

New Zealand Ministry of Education. (1993a). *National education guidelines*. Wellington, NZ: Government Printer.

New Zealand Ministry of Education. (1993b). *The New Zealand curriculum framework*. Wellington, NZ: Government Printer.

New Zealand Ministry of Education. (1993c). *Science in the national curriculum*. Wellington, NZ: Learning Media.

New Zealand Ministry of Education. (1996a). *Biology in the New Zealand curriculum*. Wellington, NZ: Learning Media.

New Zealand Ministry of Education. (1996b). *Chemistry in the New Zealand curriculum*. Wellington, NZ: Learning Media.

New Zealand Ministry of Education. (1996c). *Physics in the New Zealand curriculum*. Wellington, NZ: Learning Media.

New Zealand Ministry of Education. (1996d). *Technology in the New Zealand curriculum*. Wellington, NZ: Learning Media.

New Zealand Ministry of Education. (2002). *Curriculum stocktake report to Minister of Education, September 2002*. Wellington, NZ: Author.

National Research Act of 1974. Pub. L. No. 93–348, 88 Stat. 342. (1974).

Nuremburg Code. (1948). Retrieved June 7, 2008, from http://ohsr.od.nih.gov/guidelines/nuremberg.html

Plemmons, D. (2007). Studying IRB processes. *Journal of Empirical Research on Human Research Ethics*, *2*(1), 71–72.

Pritchard, I. A. (2002). Travelers and trolls: Practitioner research and institutional review boards. *Educational Researcher*, *31*(3), 3–13.

Protection of human subjects, 45 CFR 46. (2005).

Roberts, L. (2006). *Current practice in research ethics: Global trends and new opportunities for African universities* [Extended executive summary]. Retrieved June 9, 2008, from http://www.research-africa.net/media/pdf/EthicsExecSum.pdf

Royal Society of New Zealand. (2005). *2005 Preliminary research proposal guidelines for applicants*. Retrieved June 6, 2005, from http://www.rsnz.org/funding/marsden_fund/media/2005_Prelim_Guidelines.doc

Rubin, P., & Sieber, J. E. (2006). Empirical research on IRBs and methodologies usually associated with minimal risk. *Journal of Empirical Research on Human Research Ethics*, *1*(4), 1–4.

Schnarch, B. (2004). Ownership, control, access, and possession (OCAP) or self-determination applied to research: A critical analysis of contemporary First Nations research and some options for First Nations communities. *Journal of Aboriginal Health*, *1*(1), 80–95.

Shih, Y.-T., Shih, S.-F., Chen, N.-S., & Chen, C.-S. (2005). Human research protections: Current status in Taiwan and policy proposals [in Chinese]. *Taiwan Journal of Public Health*, *24*(4), 360–373.

Shore, N. (2007). Community-based participatory research and the ethics review process. *Journal of Empirical Research on Human Research Ethics*, *2*(1), 31–41.

Sieber, J. E. (2006). The evolution of best ethical practices in human research. *Journal of Empirical Research on Human Research Ethics*, *1*(1), 1–2.

Sieber, J. E. (2007). Respect for persons and informed consent—A moving target. *Journal of Empirical Research on Human Research Ethics*, *2*(3), 1–2.

Simons, H. (1989). Ethics of case study in educational research and evaluation. In R. G. Burgess (Ed.), *The ethics of educational research* (pp. 114–140). London: Falmer.

Simons, H. (1995). The politics and ethics of educational research in England: Contemporary issues. *British Educational Research Journal*, *21*(4), 435–449.

Simons, H., & Usher, R. (Eds.) (2000). *Situated ethics in educational research*. London: Routledge/Falmer.

South Africa Department of Health. (2004). *Ethics in health research: Principles, structures and processes* [Research ethics guidelines – 2004]. Retrieved June 6, 2008, from http://www.doh.gov.za/docs/index.html [Documents > > Fact Sheets/Guidelines > > Norms, standards, instructions]

South Africa Government. (2003). *National Health Act, No. 61*. Retrieved June 7, 2008, from http://www.doh.gov.za/docs/legislation-f.html

South Africa Human Sciences Research Council. (n.d.). *HSRC code of ethics*. Retrieved June 6, 2008, from http://www.hsrc.ac.za/Corporate_Information-8.phtml

Strike, K. A., Anderson, M. S., Curren, R., van Geel, T., Pritchard, I., & Robertson, E. (2002). *Ethical standards of the American Educational Research Association: Cases and commentary*. Washington, DC: American Educational Research Association.

Swain, H. (2006, August 11). Protect yourself and the subjects. *The Times Higher Education Supplement,* p. 12. Available from http://www.timeshighereducation.co.uk/story.asp?sectioncode = 26&storycode = 204808)

Taiwan Academia Sinica. (2007). *Homepage*. Retrieved June 8, 2008, from http://www.sinica.edu.tw/main_e.shtml

Taiwan Academia Sinica. (n.d.). *Human subject research ethics committee/IRB*. Retrieved July 16, 2008, from http://proj1.sinica.edu.tw/~irb/e-education-1.htm

Treaty of Waitangi. (1840). *Read the Treaty*. Retrieved June 8, 2008, from http://www.nzhistory.net.nz/politics/treaty/read-the-treaty/english-text

Truscott, D. (2004). Fieldwork, participation, and practice: Ethics and dilemmas in qualitative research [Book review]. *Science Education*, *88*(5), 811–813.

United Kingdom Economic and Social Research Council. (2006). *Research ethics framework (REF)*. Retrieved June 5, 2008, from http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/ESRC_Re_Ethics_Frame_tcm6–11291.pdf

United Kingdom Economic and Social Research Council. (n.d.-a). *Homepage*. Retrieved June 9, 2008, from http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/opportunities/research_ethics_framework/

United Kingdom Economic and Social Research Council. (n.d.-b). *Research ethics framework: Homepage*. Retrieved June 9, 2008, from http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/opportunities/research_ethics_framework/

United Nations. (1990). *Convention on the rights of the child*. New York: UN High Commission for Human Rights Office. Available from http://www.unicef.org/crc/

United States Department of Health and Human Services. (n.d.). *Basic HHS policy for protection of human research subjects (45 CFR 46)*. Retrieved July 16, 2008, from http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm

United States Interagency Arctic Research Policy Committee; Social Science Task Force. (1995). Principles for the conduct of research in the Arctic. *Arctic Research of the United States*, *9*(Spring), 56–57. Retrieved from http://arcticcircle.uconn.edu/SEEJ/ethics.html

United States National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979, April 18). *The Belmont report: Ethical principles and guidelines* *for the protection of human subjects of research*. Retrieved June 6, 2008, from http://ohsr.od.nih.gov/guidelines/belmont.html

United States National Research Council. (2002). *Scientific research in education*. Committee on Scientific Principles for Education Research. R. J. Shavelson & L. Towne (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

United States National Research Council. (2004). *Advancing scientific research in education*. Committee on Research in Education. L. Towne, L. L. Wise, & T. M. Winters (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

University of Botswana. (2004). *Policy on ethics and ethical conduct in research*. Retrieved June 6, 2008, from http://www.ub.bw/ord/r_documents.cfm?d = Ethics_Policy_RD04_05H.pdf

University of Cape Town. (2006). *Faculty of Humanities guide to research ethics: Research on human subjects*. Retrieved June 6, 2008, from http://www.humanities.uct.ac.za/research/ethics/

University of Pretoria. (n.d.). *Code of ethics for research*. Retrieved June 8, 2008, from http://web.up.ac.za/sitefiles/File/3653/Code%20of%20ethics%20for%20research(1).doc

University of Victoria Human Research Ethics Board. (2008, June). *Application for ethics approval for human participant research*. Retrieved June 6, 2008, from http://www.research.uvic.ca/forms/index.htm#HREC [HREB > > Applications]

University of Victoria Human Research Ethics Office. (2008, January). *Guidelines for ethics in dual-role research for teachers and other practitioners*. Retrieved June 5, 2008, from http://www.research.uvic.ca/forms/hrec/geidrr.pdf [HREB > > Guidelines]

World Medical Association. (2004). *Declaration of Helsinki: Ethical principles for medical research involving human subjects*. Retrieved June 6, 2008, from http://www.wma.net/e/policy/b3.htm

Yang, C. M., Kuo, N. W., Chen, C. S., & Chou, C. Y. (2001). Situational analysis of research ethics review in Taiwan. *Public Health Quarterly*, *28*(3), 177–187.

Yore, L. D., Chinn, P. W. U., & Hand, B. (Eds.). (2008). Science literacy for all: Influences of culture, language, and knowledge about nature and naturally occurring events [Special Issue]. *L1—Educational Studies of Language and Literacy, 8*(1). Retrieved from http://l1.publication-archive.com/public?fn = enter&repository = 1

Zeni, J. (Ed.). (2001). *Ethical issues in practitioner research*. New York: Teachers College Press.

Chapter 25
Data Sharing: Disclosure, Confidentiality, and Security

David J. Dude, Michelle A. Mengeling, and Catherine J. Welch

A primary mission of educational research is to champion excellence in education. Through the design of research that is generalizable and replicable and also through the successful implementation of research results, higher levels of learning, achievement, and performance for students can be obtained. Given this mission, data and information become essential resources to the success of such research. Data come in many forms—numeric, text, images, audio, and video—and are essential for the translation of research into knowledge, products, and procedures to improve education. Recognizing the costs, both monetary and time, of collecting and maintaining rich data, the concept of data sharing is often promoted with educational research. The sharing of data encourages diversity of analyses and perspectives, promotes new research, and makes possible the testing of new or alternative approaches and methods. Sieber (2006) supported the sharing of data and emphasized the value to secondary users who "can verify, refute, or refine original results, building upon those existing data, develop and test new theories, generalize or extend tentative findings and answer new empirical questions" (p. 48).

Yore and Boscolo (see Chap. 2) indicate that the US National Research Council (NRC) taskforce report *Advancing Scientific Research in Education* (US NRC, 2004) focused on ways to improve the quality of, and substantive foundation for, education research, which in part involved professional development of researchers and funding evaluation panels, and the ethics, structures and procedures for sharing data. Data sharing also promotes the extension and expansion of previous research and helps to facilitate the work of new researchers. The sharing of data permits the creation of new datasets when data from multiple sources can be combined to create an even richer dataset.

Sharing data also provides an opportunity to avoid the duplication of expensive data collection procedures. Through multiple uses of the same dataset, cost efficiencies are realized both in terms of direct costs and timelines. Through secondary uses of datasets, researchers are able to build on and advance the work of others.

D.J. Dude, M.A. Mengeling, and C.J. Welch
University of Iowa

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

Researchers need to recognize that data sharing may be complicated or limited by institutional practice and policy, local institutional review board (IRB) rules, and state/provincial and federal regulations. In 2003, the US National Institutes of Health (US NIH, n.d.-b) implemented a requirement that grants with direct costs over $500,000 incorporate plans for data sharing that protected the identity of individual participants. Consistent with this requirement and regardless of project size or scope, researchers need to be aware of the limitations and restrictions of sharing data and then address these issues within their design plans. All entities involved in data sharing—the originators of the data, data recipients, consultants, and secondary users—must be aware of the policies that govern use of the data and also understand and comply with the infrastructure requirements that enforce these policies.

Government statistical agencies have addressed many of the same issues faced by educational researchers with respect to data sharing. Disclosure review boards are frequently instituted to provide human research protection within a data-sharing arrangement. Zarate and Zayatz (2006) discussed disclosure limitations and the role of disclosure review boards. Using examples from the National Center for Health Statistics and the US Census Bureau, critical elements of disclosure review, which have application to other jurisdictions, are discussed. In-depth knowledge of the data file structure and the population being studied are critical to the success of these review boards.

Given the concern and limitations of data sharing, one solution may be the use of publicly released data. However, finding a balance between the usefulness of publicly released data and protection of the participants is often difficult (O’Rourke et al., 2006). Rodgers and Nolte (2006) addressed approaches to restricting access to shared data. They developed four modes of data sharing ranging from the least restrictive (public data files), to an increasing level of restriction (health data files and restricted data files), to the most restrictive (data enclaves). They also discussed the usefulness of a data enclave that provides a monitored environment for data use.

Above all, the rights and privacy of individuals who participated in research studies or who are part of an existing dataset must be protected. Data intended for broader use must be free of any information that would permit identification of individuals. Issues of data ownership, proprietary data, and governance policies must be addressed. Issues related to disclosure, confidentiality, and maintaining security make data sharing a very critical point in the research process and one to be taken seriously by the researcher.

25.1 Overview

The purpose of this chapter is to assist researchers in understanding these data security issues. We present four primary areas of concern for researchers to consider when they design their research and as they address issues related to the source of their data.

Data Security. Security issues related to the physical, electronic, and administrative security of data are presented and discussed. Key to this section is the recognition that data can be stored in a variety of ways. Issues specific to different types of storage (both physical and electronic) are discussed. Administrative security addresses issues related to the level of access to data.

Sharing of Data. Issues specific to sharing are discussed next. Sharing data both within an organization and between organizations are discussed as well as procedures that researchers may be asked to follow.

Educational Data for Research. Procedures to ensure the appropriate collection of data and the groups that govern these procedures are discussed.

Data Integrity. Finally, the issue of data integrity is presented. The level of confidence that researchers have in their results is directly related to the quality and integrity of the data. Issues related to the editing of data and combining data from multiple sources are discussed.

25.2 Data Security

Data can be stored physically or electronically; often there is little distinction between the two. The advantages of storing data electronically are immense, but with those advantages come challenges in maintaining the security of those data. Data must be protected physically, electronically, and administratively. Physical security involves the actual location and security of the physical media on which the data are stored, or in the case of physical data, the data themselves. Electronic security includes data encryption and methods for limiting access. Administrative security includes the policies that control the access to and sharing of data. Researchers wishing to use data must comply with the security standards set forth by the appropriate governing agencies. For example, the following legislation addresses levels of security, encryption procedures, secure-role-based access, and auditing and transaction logging in the United States:

- Family Educational Rights and Privacy Act of 1974, as amended (FERPA, 2006, 34 CFR § 99.31)
- Health Insurance Portability and Accountability Act of 1996 (HIPAA, US NIH, n.d.-a), where applicable

25.2.1 Physical Security

The physical security of data is often overlooked. If data are stored on a desktop computer, for instance, is that computer in an office that is locked when someone is not present? If it is in a locked office, who has access to that office in addition to the office resident(s)? Managers? Directors? Graduate students? Custodians? Security personnel?

When data are stored on portable media, such as notebook computers, USB drives, compact discs, etc., physical security becomes much more difficult. Notebook computers are frequently stolen and sometimes contain sensitive data. There have been many high-profile cases of notebook computers being stolen from researchers' vehicles and homes (Associated Press, 2006a, 2006b; Greenemeier, 2008; Sullivan, 2006; Tyson & Lee, 2006; US NIH, 2008; Yen, 2006). Cases such as these emphasize the need for the other two types of security: electronic and administrative.

In either case, it is also important to ask what is done with these items when they are no longer needed. The answer seems simple in cases involving portable media since those are often just thrown out. Disposing of or reallocating a desktop computer, however, is often a more complicated issue. These issues are discussed further in Sect. 25.3.

25.2.2 Electronic Security

A basic form of electronic security is encryption. Encryption is a process by which data are scrambled using an algorithm and a key (much like a password). The data can be unscrambled only by using the key that was used to scramble them. Data that are encrypted are much less susceptible to breach than data that are not encrypted. Encryption is not completely foolproof, but current encryption techniques are extremely secure.

Data that are stored on a desktop computer or a server often can be controlled by Access Control Lists (ACLs). ACLs allow one to specify precisely who is and is not allowed to access the data. Depending on the system being used, rights can be assigned to individual users and/or groups of users. It is often possible also to control who can only look at the data (often called read access) and who can actually change it (often called write access). ACLs on a server are usually maintained by the server administrator.

Another issue when data are stored electronically and need to be shared is that of transmission. There are many ways to share data electronically, but the security of the various methods varies substantially. Depending on the size of the file(s) containing the data, a researcher may be tempted to share the data in an email or as an attachment to an email. Email is an extremely insecure method of sharing data, so any attached files, at a minimum, should be encrypted. For larger files, many researchers use a protocol called File Transfer Protocol (FTP). Unfortunately, FTP in its native form is inherently insecure. The contents of the file being transmitted, as well as the user name and password used to log in to the server, are sent across the network in plain text. Fortunately, improvements have been made to allow researchers to use more secure FTP protocols, including FTPS (FTP using Secure Sockets Layer) and SFTP (FTP using Secure Shell). These protocols encrypt the information that is sent over the network to greatly reduce the risk of a breach. The details of these technologies are beyond the scope of this chapter, but interested readers are referred to Barrett, Byrnes, and Silverman (2005) and Rescorla (2000) for more information.

The *Standards for Educational and Psychological Testing* (the Standards; Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, 1999) address the importance of secure transmission. Standard 5.13 states, “Transmission of individually identified test scores to authorized individuals or institutions should be done in a manner that protects the confidential nature of the scores” (p. 66). Although this standard refers specifically to test scores, the same advice should be heeded for transmission of any data (i.e., numeric, text, audio, video, etc.) by any means (i.e., FTP, email, hard copy, etc.).

25.2.3 Administrative Security

Physical and electronic security must be accompanied by administrative security. Policies must be in place to clarify what physical and electronic security must be enforced. Data involving unique government identifier numbers (e.g., Social Security or Social Insurance numbers) or individual medical information, for instance, must have stricter policies than data of nonidentifiable individuals.

In many research institutions, these policies are set by the IRB. The Code of Federal Regulations (2007) states, “When appropriate, there are adequate provisions to protect the privacy of subjects and to maintain the confidentiality of data” (45 CFR § 46.111 (a) (7)). The Standards also address administrative security in Standard 5.16, which states, “Organizations that maintain test scores on individuals in data files or in an individual’s records should develop a clear set of policy guidelines on the duration of retention of an individual’s records, and on the availability, and use over time, of such data” (Joint Committee, 1999, p. 66). In those cases where there are not existing policies, or where the policies are not stringent, individual researchers should implement their own policies to maintain the security of the data.

The three areas of data security—physical, electronic, administrative—do not exist independently of each other. It is critical to develop policies that address the physical and electronic security of data in each research situation. A layered security system improves overall security through the built-in redundancy of the layers.

25.3 Sanitizing Media

True understanding of how files are stored on a computer and subsequently removed is often outside the realm of a researcher’s expertise. Many people believe that when a file is deleted it has been removed from the computer. Others know that that is not the case but believe by emptying the recycle or trash bin that the files actually have been removed. Unfortunately, both groups of people would be wrong. It is actually very difficult to remove data completely from electronic media, such as hard drives. On most systems, deleting a file and emptying the recycle bin simply removes the pointer to that file. The actual file remains in the slack (the space between the end of a file and the end of the disk cluster it is stored in) of the hard drive and is recoverable with special software until that portion of the hard drive happens to be overwritten with new information. Sometimes, with advanced forensic techniques, it is even possible to recover files that have been overwritten. To truly remove data, they must be sanitized.

There are three methods for sanitizing data, listed in decreasing order of effectiveness. First, the media on which the data are stored can be physically destroyed. Second, a method called degaussing can be used to render the media useless and completely scramble the electronic signature. Third, the data can be overwritten with new information.

25.3.1 Physical Destruction

Physical destruction is by far the most secure method of sanitizing data. Completely pulverizing, burning, or melting the media on which data are stored (i.e., paper, microfiche, floppy disks, hard drives, USB drives, etc.) can render it impossible to recover the data. In many cases, it is impossible or impractical to destroy physically the media containing the data. Luckily, there are good, less-expensive alternatives to physical destruction.

25.3.2 Degaussing

Degaussing electronic media, such as a hard drive, involves using a strong magnetic field to randomize the magnetic alignment of the individual bits in the media. Doing this sufficiently can render data completely unrecoverable. Degaussing is effective but can render a device unusable. A device used for degaussing can be quite expensive. Information technology (IT) departments at large organizations might have such a unit, and there are companies that will rent units and/or provide degaussing services.

25.3.3 Overwriting

Overwriting data to destroy them is the least secure method of sanitization but is nonetheless a good alternative. Special software exists that will wipe a disk by overwriting using various algorithms. These algorithms usually include overwriting many times, using both random strings and specific patterns. When performed to high standards, this software can render data virtually irrecoverable. Specific recommendations of software titles used for wiping data may be available from a researcher's IT department.

25.3.4 Backups

Backups of data can be overlooked easily even when great care is being taken to sanitize media. Some backups are automated, so it is important for researchers to work with their IT department to determine if there are any backups of sensitive data and to make sure the backups are sanitized in addition to the original media.

25.4 Sharing Data

It is often necessary to share data within an organization or amongst people at several organizations. When sharing data, especially with others outside your organization, data security and issues relating to disclosure, confidentiality, and ownership must be considered.

It is the responsibility of the data owner(s) to ensure that all documentation, electronic files, and data are developed, used, and maintained in a secure manner for authorized purposes, protecting the confidentiality of all materials, records, and files. The data owner must ensure that all data collected and presented to end users are regulated by the restrictions on data sharing, for US researchers as outlined by FERPA.

Researchers must prohibit the disclosure of personally identifiable information to any person unless such person is authorized to disclose this information. For example, in education settings, school district personnel at the appropriate administrative level or officials from the state department of education may have the authority to release such personally identifiable information. However, this authorization must be part of a public policy that is available to the individuals affected by this disclosure and approved by the IRB.

25.4.1 Within an Organization

Researchers within an organization often can share data by using shared storage space on a server. Access to this space usually can be reasonably secured (see Sect. 25.2.2) while still allowing authorized persons easy access.

It is critical for researchers within an organization to understand and/or develop administrative policies regarding access to the data and procedures to retain the integrity of the data. Written policies and rules on appropriate access and use of data are essential for communicating the expectations to all researchers within an organization, outlining the processes to be followed, and identifying the potential consequences of violating the policy. Thus, written policies and rules need to be distributed to all researchers or staff members associated with using the data. It would be appropriate for organizations to obtain assurances from researchers that they have read the policies and rules and that they agree to abide by them. A comprehensive policy would cover approved activities associated with the use of data, dissemination of the results, and how best to store the data.

25.4.2 Between Organizations

When sharing data between different organizations, it is especially important to have administrative policies in place that address ownership, eligibility, and access. These policies must address who, among those with access to the data, has the right to share those data with others. Procedures regarding access privileges, the purposes for which access can be granted, and the duration of possession should be developed for policy implementation purposes. Consequences of not abiding by the policy should also be articulated in the policy.

Organizational policy specifically may address appropriate eligibility for access to data. For example, prior to being granted access to data, researchers may be asked to (a) sign nondisclosure agreements, (b) sign acceptable use and security agreements, or (c) submit to background checks. Investigations into the backgrounds of researchers who potentially have access to such datasets may be required prior to authorization to disclose. Pending compliance with such eligibility requirements, researchers then may be provided access. Researchers should note that these procedures may require a substantial amount of time prior to receiving access authorization and details of approved uses. A related important issue is how to share that data securely (see Sect. 25.2.2).

25.5 Educational Data for Research

As mentioned previously, research data use is guided by professional standards and ethics, government legislation, and organizations, such as IRBs. These sources provide guidelines to promote research practices that protect the confidentiality of research participant information. Data confidentiality must be maintained when the data can be linked to an individual. The issue of data security can be categorized into two areas: confidentiality of data and releasing data to others. These areas will be discussed within the context of typical resources relied on by educational researchers for conducting appropriate, ethical, and quality research.

25.5.1 Professional Guidelines

The ethical conduct of researchers is guided by ethical codes, IRBs, and legislation, such as FERPA. This chapter provides several specific examples of practices aimed at promoting ethical research practices. Anthony and colleagues (see Chap. 24) provide an extended discussion on the background and critical issues of research ethics.

To guide educational researchers, the AERA published the *Ethical Standards of the American Educational Research Association* (AERA, 2000; referred to as the Code of Ethics). This document provides standards of researcher responsibility to the field, to research subjects, and for research dissemination. Within the Code of Ethics are the *Guiding Standards: Research Populations, Educational Institutions, and the Public*, which provide the context for educational research.

The Code of Ethics is not prescriptive but does provide standards that describe what it means to respect and that protect the rights of human subjects during educational research. Participants have the right to confidentiality and researchers are responsible for protecting that confidentiality. It is expected that researchers will adhere to the Code of Ethics. AERA does not monitor adherence to the Standards, nor does it investigate possible violations of the Code.

The American Psychological Association (APA) published the *Statement of the Disclosure of Test Data* as a reference for psychologists with regard to the disclosure of test data. Their statement is intended to be consistent with the *Ethical Principles of Psychologists and Code of Conduct* (APA, 2002) and the *Standards for Educational and Psychological Testing* (Joint Committee, 1999). As was true for the Code of Ethics, the APA does not provide oversight to monitor compliance. However, there are organizations that do monitor compliance.

The following organizations and legislation in the United States monitor research, with the power to stop or penalize research that does not maintain sufficient quality control over confidentiality: IRBs (see Anthony et al., Chap. 24), FERPA, and the Institute of Education Sciences (IES).

25.5.2 Institutional Review Boards

An IRB is a committee that has been formally assembled to approve, monitor, and review research conducted on human subjects. Its mandate is to protect the rights and welfare of research subjects. In the United States, the Food and Drug Administration (FDA) and the Department of Health and Human Services (HHS) empower IRBs to approve, require modifications (to secure approval), or reject research designs. An IRB provides critical oversight throughout the study to ensure that research on human subjects is scientific, ethical, and regulated.

IRB is a generic term used by the FDA and HHS; each institution that establishes an IRB chooses its own name. Other names that are synonymous include Independent Ethics Committee (IEC), Ethical Review Board (ERB), or research

Ethics Committee (REC). The Research Act of 1974 (CFR, 2005, § 46.111) defined the purpose of an IRB and required all research receiving funding from what is now HHS to have IRB approval. The IRBs are regulated in turn by the Office for Human Research Protections (OHRP) within HHS.

Exemptions to IRB approval can be obtained. HHS provides examples of situations where IRB approval may not be required (CFR, 2005, § 46.101(b)). The OHRP provides graphic aids to be used by IRBs, investigators, and others in deciding whether a study is considered to be research involving human subjects and, therefore, must be reviewed by an IRB according to HHS regulation 45 CFR Part 46. However, the guidelines provided by OHRP may not be specific enough for a particular situation. It is, therefore, recommended that an application for an exemption be sought through the researcher's IRB.

25.5.3 Family Education Rights and Privacy Act of 1974 (FERPA)

The US Department of Education (US ED, n.d.) provides its own guidelines for use of student information through legislation known as FERPA (CFR, 2006). FERPA deals specifically with the education records of students. Any student who is 18 or has attended a postsecondary institution has the right to inspect and review his or her own education records, request amendments to those records, and have some control over the disclosure of personally identifiable information from these records. There are exceptions to the requirement of student consent.

> Generally, schools must have written permission from the parent or eligible student in order to release any information from a student's education record. However, FERPA (2006, 34 CFR § 99.31) allows schools to disclose those records, without consent, to the following parties or under the following conditions:
>
> - School officials with legitimate educational interest
> - Other schools to which a student is transferring
> - Specified officials for audit or evaluation purposes
> - Appropriate parties in connection with financial aid to a student
> - Organizations conducting certain studies for or on behalf of the school
> - Accrediting organizations
> - To comply with a judicial order or lawfully issued subpoena
> - Appropriate officials in cases of health and safety emergencies
> - State and local authorities, within a juvenile justice system, pursuant to specific State law
>
> Schools may disclose, without consent, "directory" information such as a student's name, address, telephone number, date and place of birth, honors and awards, and dates of attendance. However, schools must tell parents and eligible students about directory information and allow parents and eligible students a reasonable amount of time to request that the school not disclose directory information about them. Schools must notify parents and eligible students annually of their rights under FERPA. The actual means of notification (special letter, inclusion in a PTA bulletin, student handbook, or newspaper article) is left to the discretion of each school (US ED, n.d., para. 5-6).

Students' rights under FERPA are essentially the same while attending an institution and after they leave.

25.5.4 Institute of Education Sciences, US Department of Education

Data may not always be collected through the individuals themselves but through already existing data sources. Issues of confidentiality, IRB oversight, and FERPA regulations are still relevant. The intent is to prevent the release of individual or institutional identifying information. The Institute of Education Sciences (IES), a part of the US ED and the overarching structure for several centers including the National Center of Education Statistics (NCES), differentiates between public-use data and restricted-use data. Public-use data are defined as data in which individually identifiable information is coded or deleted to protect the confidentiality of the participant. Public-use data are available to the general public.

Restricted-use data are data that contain individually identifiable information. The data collected by IES are protected by numerous laws including the Privacy Act of 1974, the Computer Security Act of 1987, the E-Government Act of 2002, the US Patriot Act of 2001, and the Education Sciences Reform Act of 2002. Unlawful disclosure of IES restricted-use data can result in a substantial fine and/or a prison sentence. Researchers can access restricted-use data provided that they obtain a license through IES, which has the right to monitor compliance of research using their data. IES provides security procedure requirements to ensure that restricted-use data are secure at all times.

25.6 Data Integrity

Data integrity is a term used to describe the accurateness, completeness, and validity of the data collected and subsequently analyzed. In the process of collecting and modifying data, researchers must take steps to ensure that the integrity of the original data is preserved.

25.6.1 Data Cleaning

Data cleaning refers to inspecting the data for accurateness and completeness. The initial step in data cleaning is to inspect the data. Each variable should be examined to ensure that the range of values for each variable is within an expected range. For example, if a study collected the birthdates of a group of students in Grade 4, it is necessary to validate that student ages are around 9–11 years, a typical age range for American fourth graders. If the researcher or data analyst discovers errors in the data, such as a student reporting the current year as their birth year and a correct birth year cannot be obtained for the subject, the variable should be coded as missing. *Missing data* refers to places in the dataset where no information is available for that subject on that variable.

The completeness of the data is evaluated by determining where the missing data occur and their prevalence in the dataset. In any dataset there may be numerous missing data due to incorrect or omitted responses. It is reasonable to change an invalid piece of data to missing, but it is difficult to change a missing piece of data without some outside piece of validation. No guessing or estimating missing responses should occur during data cleaning. Readers interested in data imputation, finding plausible values for missing data, should refer to Rubin (1987) and Schafer (1997).

25.6.2 Matching Data

Many studies collect data from multiple sources or at multiple times and must match individual records across datasets. This can be immensely time-consuming, depending on the number of subjects and the quality of the matching. Before any matching is done, researchers should have an idea of the number of matches possible across datasets. To promote a Gold Standard of research, researchers should publish their match rates, as they would for other indicators of population and sample size. A match rate would be analogous to a response rate, which is a mainstay of published survey research. The problem with calculating a match rate (or a response rate) is to determine the denominator accurately. Unfortunately, there are no guidelines or rules to follow. The goal always would be to strive for clarity when reporting study results (actual matches divided by possible matches times 100). The following two examples illustrate this idea.

Example 1: A study is carried out where Grade 4 students in District A (n = 50) are tested in the fall. The following spring, District A again tests its Grade 4 students (n = 52). Of the original 50 students tested in the fall, 48 were tested again in the spring. The match rate was 48/50 = 96% of the maximum possible matches.

Example 2: District A tested all Grade 4 students in the fall (n = 52) and again in the spring (n = 50). Out of the maximum possible matches (n = 50), 48 students were tested twice resulting in a match rate of 96%.

Explanations of how the data-matching process is carried out should be provided with the match rates. Most initial matching is done programmatically. Often the first step is to try to match all variables that are consistent between datasets, for example, name, age, and gender. Although an important first step, the use of a computer to make exact matches often will result in an unnecessarily low match rate. A manual inspection of computer matches is highly recommended. The following example illustrates how computer and manual matching would produce the highest quality, matched dataset.

Dataset A has 12 students and Dataset B has 14 students. The variables used to match records across datasets are provided in Table 25.1. The initial computer match used student name, birth month, birth year, and gender. Exact matches were made between eight records (computer matches). The others could not be directly matched because of missing data or inexact name matches. A visual inspection of the unmatched data quickly revealed that an additional three matches could be made (manual matches). In reporting the match rate, it is helpful to report that of the 12 records in Dataset A, 75% (8 out of 12) were exact matches on name, birth date, and gender. An additional three matches were made by a visual inspection of the data, which corrected for nicknames and missing genders, resulting in a final match rate of 92% (11 out of 12).

The purpose of providing a brief description of the matching process is to be consistent with the intent of the AERA's *Standards for Reporting on Empirical Social Science Research in AERA Publications* (2006). These standards were based on two principles: (a) that adequate evidence should be provided to justify the results and conclusions, and (b) that reports of empirical research should make explicit the methods used throughout the study. Appropriate and sufficient information should be provided, such that another researcher is able to either replicate or reproduce the methods used.

Table 25.1 Dataset matching example

	Dataset A						Dataset B				
	Last name	First name	Birth month	Birth year	Gender	Match	Last name	First name	Birth month	Birth year	Gender
Computer matches											
1	Bepn	Ben	9	98	M	√	Bepn	Ben	9	98	M
2	Cals	Chris	10	97	M	√	Cals	Chris	10	97	M
3	Delx	Dan	2	98	M	√	Delx	Dan	2	98	M
4	Ephr	Eva	11	97	F	√	Ephr	Eva	11	97	F
5	Gent	Greg	4	98	M	√	Gent	Greg	4	98	M
6	Higt	Hal	11	97	M	√	Higt	Hal	11	97	M
7	Illo	Ian	5	98	M	√	Illo	Ian	5	98	M
8	Klen	Kip	12	97	M	√	Klen	Kip	12	97	M
Manual matches											
9	Avel	Ann	8	98	.	√	Avel	Ann	8	98	F
10	Foxl	Francis	12	96	F	√	Foxl	Fran	12	96	F
11	Joen	Jill	1	98	F	√	Joen	Jillian	1	98	F
Unmatched records											
							Fiin	Gus	4	98	M
12	Lose	Lisa	7	98	F						
							Mapp	Mina	6	98	F

25.7 Closing Remarks

Although issues related to data sharing and security may be complicated and limiting, researchers can avoid serious problems by addressing these issues in a systematic way as early as possible in their work. Researchers should recognize that each project may pose specific opportunities or challenges with respect to data sharing and security. However, researchers are responsible for gathering information related to the relevant data-sharing policies and for the documentation of this information in their research plans. They must be knowledgeable about these policies and work to ensure that all individuals working with the data on their behalf adhere to these policies. Researchers are encouraged to consult with experts in data-sharing methodology and procedures in the early stages of their plans (Rodgers & Nolte, 2006).

Researchers should draft their own data-sharing plan for the dissemination of their results. Dissemination plans should be approved by the owners of the data and should be consistent with the policies that govern those particular data. If data sharing is limited and constrains the dissemination of the results, researchers should be aware of this prior to the collection of new data or the use of existing data. These types of limitations should be part of the plan prepared by the researcher.

The field of data sharing is changing quickly. Data sharing goes well beyond numerical data. These policies are expected to apply to all data sharing including those currently shared as text files and video files. As technology progresses, access to combined video files, the use of video analysis systems, and the secondary reanalyses of discourse, conversation, and performance will become more commonplace (see Rossman & Yore, Chap. 26). These expanding opportunities will continue to challenge researchers with respect to data sharing and maintaining confidentiality. And, as new techniques for de-identifying data become available, so will new techniques for re-identifying the data (Zarate & Zayatz, 2006).

Researchers are encouraged to follow current best practice and to follow changes in the methodologies as well. Universities and government agencies will continue to provide professional development opportunities to educate researchers on evolving policies. It is the responsibility of the researcher to identify and take advantage of these opportunities. Researchers are encouraged to incorporate continuing education of this type in the appropriate courses for graduate students.

References

American Educational Research Association. (2000). *Ethical standards of the American Educational Research Association*. Retrieved July 9, 2008, from http://www.aera.net/AboutAERA/Default.aspx?menu_id = 90&id = 222

American Educational Research Association. (2006). *Standards for reporting on empirical social science research in AERA publications*. Retrieved June 23, 2008, from https://www.aera.net/uploadedFiles/Opportunities/StandardsforReportingEmpiricalSocialScience_PDF.pdf

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct*. Retrieved May 31, 2008, from http://www.apa.org/ethics/code2002.html

Associated Press. (2006a, June 21). *Hacker breaks into USDA computer system; Personal data on 26,000 workers may be at risk*. Retrieved July 9, 2008, from http://www.msnbc.msn.com/id/13470744/

Associated Press. (2006b, July 1). *Red Cross laptop with donor information stolen; Social Security numbers on computer, but officials say data is encrypted*. Retrieved July 9, 2008, from http://www.msnbc.msn.com/id/13657607/

Barrett, D. J., Byrnes, R. G., & Silverman, R. E. (2005). *SSH, the secure shell* (2nd edn.). Sebastopol, CA: O'Reilly & Associates, Inc.

Criteria for IRB approval of research, 45 CFR Part 46, § 46.111. (2007).

Family educational rights and privacy act regulations, 34 CFR Part 99, § 1232g. (2006).

Greenemeier, L. (2008, March 25). Security breach: Feds lose laptop containing sensitive data – again, *Scientific American*. Retrieved from http://www.sciam.com/article.cfm?id = security-breach-lost-laptop.

Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd edn.). Washington, DC: American Educational Research Association.

O'Rourke, J. M., Roehrig, S., Heeringa, S. G., Reed, B. G., Birdsall, W. C., Overcashier, M., et al. (2006). Solving problems of disclosure risk while retaining key analytic uses of publicly released microdata. *Journal of Empirical Research on Human Research Ethics, 1*(3), 63–84.

Protection of human subjects, 45 CFR 46. (2005).

Rescorla, E. (2000). *SSL and TLS: Designing and building secure systems*. Harlow, UK: Addison-Wesley Professional.

Rodgers, W., & Nolte, M. (2006). Solving problems of disclosure risk in an academic setting: Using a combination of restricted data and restricted access methods. *Journal of Empirical Research on Human Research Ethics*, *1*(3), 85–98.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Sieber, J. E. (2006). Introduction: Data sharing and disclosure limitation techniques. *Journal of Empirical Research on Human Research Ethics*, *1*(3), 47–50.

Sullivan, B. (2006, May 22). *All veterans at risk of ID theft after data heist; Burglar reportedly took Veterans Affairs disk containing personal info*. Retrieved July 9, 2008, from http://www.msnbc.msn.com/id/12916803/

Tyson, A. S., & Lee, C. (2006, June 7). Data theft affected most in military: National security concerns raised. *The Washington Post*, p. A01. Retrieved July 8, 2008, from http://www.washingtonpost.com/wp-dyn/content/article/2006/06/06/AR2006060601332.html

United States Department of Education. (n.d.). *Family educational rights and privacy act (FERPA)*. Retrieved July 21, 2008, from http://www.ed.gov/policy/gen/guid/fpco/ferpa/index.html

United States National Institutes of Health. (2008, March 24). *Statement from Elizabeth G. Nabel, M.D., director of the National Heart, Lung, and Blood Institute (NHLBI), on a stolen NHLBI laptop computer* [Press release]. Retrieved from http://public.nhlbi.nih.gov/newsroom/home/GetPressRelease.aspx?id = 2559

United States National Institutes of Health. (n.d.-a). *Health Insurance Portability and Accountability Act (HIPAA) privacy rule: Information for researchers*. Retrieved July 9, 2008, from http://privacyruleandresearch.nih.gov/

United States National Institutes of Health. (n.d.-b). *NIH data sharing policy*. Retrieved July 9, 2008, from http://grants2.nih.gov/grants/policy/data_sharing/

United States National Research Council. (2004). *Advancing scientific research in education*. Committee on Research in Education. L. Towne, L. L. Wise, & T. M. Winters (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Yen, H. (2006, June 22). *Feds hit by rash of data breaches; FTC said Thursday that it had lost laptops containing sensitive data*. Retrieved July 8, 2008, from http://www.msnbc.msn.com/id/13492060/

Zarate, A. O., & Zayatz, L. (2006). Essentials of the disclosure review process: A federal perspective. *Journal of Empirical Research on Human Research Ethics*, *1*(3), 51–62.

Chapter 26
Stitching the Pieces Together to Reveal the Generalized Patterns: Systematic Research Reviews, Secondary Reanalyses, Case-to-case Comparisons, and Metasyntheses of Qualitative Research Studies

Gretchen B. Rossman and Larry D. Yore

Literacy, language, and science education research is much like quilting, in which small pieces of fabric are stitched together into repeated units (blocks) to produce a functional bedcovering or artistic wallhanging of a predetermined size and shape. The repeated units—blocks—are normally prescribed and uniform squares of fixed dimensions. Each block contains a whole or partial design that is a fractional part of the final dimensions of the finished quilt. There are prescribed procedures for making quilts that, when followed rigorously, result in a generalized pattern of beauty and practicality. Quilting parties bring together several quilters, each working independently of the others but in their company (a community of practice). Each follows the prescribed pattern producing individual blocks that are finally stitched together by the lead quilters to yield the synthesis—an artistic or geometric pattern (for the interested reader, see http://www.houseofquilts.com).

One variation is a crazy quilt, which is created with leftover bits and pieces of variably sized, variably shaped, and variably colored fabric pieces. Crazy quilting suggests unrestricted creativity for the individual quilter in using a variety of shapes, colors, and textures. Much creativity is possible in crazy quilting; but the quilter is constrained to ensure that the individual blocks yield a shape or size dictated by the intended purpose (bedcovering, wallhanging, baby quilt), available fabric (cotton, linen), and desired function (comfort, aesthetics). The unit of design is not predetermined and may not be visible until the quilt is completed, if then, when the individual contributions are stitched together (for more information, see http://www.nmia.com/ mgdesign/qor/styles/crazy/crzayqlt.htm).

This chapter attempts to address the recommendations of the 2nd Island Conference regarding more effective use of quantitative databases and qualitative information stores and also the production of generalizations across isolated

G.B. Rossman
University of Massachusetts, Amherst

L.D. Yore
University of Victoria

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

research studies within a specific problem space. These recommendations and the resulting solutions are meant to address politicians', policy makers', and decision makers' needs for compelling arguments and claims based on persuasive collections of evidence that are generalizable to their problems, situations, and constituents. Such solutions are reasonably well established, but evolving, in the quantitative research community; however, the processes, techniques, and procedures are not as developed in the qualitative research community.

We provide a brief historical perspective and lessons learned from meta-analysis and secondary reanalysis of quantitative data, followed by an overview of a balanced perspective applied to qualitative findings and the embedded logics, and then a discussion of four promising qualitative techniques from the health care, medical, and social sciences research communities: research review, secondary reanalysis, case-to-case comparison, and metasynthesis. Each approach has potential in science and literacy education research and has had some uptake in these communities. We believe secondary analysis and synthesis will help address the concerns of politicians and bureaucrats that have led to the privileged position of randomized controlled trials (RCTs) as the only Gold Standard for research.

26.1 Quantitative Research Syntheses: Meta-analysis and Secondary Reanalysis

Quantitative research can be viewed as analogous to traditional quilting because it stipulates a predetermined hypothesis, method, data collection, and statistical analysis; these serve as the repeated unit of design. Quantitative inquiries involve formalistic and mechanistic worldviews concerned with forms, characteristics, and their causal relationships, indirect influences or correlation associations, and the belief in correspondence between the observed and the ideal following deterministic logical rules (Roberts, 1982). If done correctly, such procedures should yield results that are generalizable and thus applicable to a broader array of problem settings, similar to those represented by the samples studied. Generalizability is dependent on how well the samples investigated represent the larger population. And strictly speaking, findings can only be (probabilistically) generalized from the sample to the population from which it was drawn.

However, the ideal of random sampling in which all members of a target population have equal probability of being selected is difficult to fully achieve in practice. Protection of human subjects and research ethics requirements, which demand informed choice and voluntary participation and also call for the avoidance of undue power-over research subjects, increase the difficulties in achieving truly random samples to serve as experimental and control groups in literacy, language, and science education research based in actual schools and classrooms. These difficulties have led to the use of nonrandom and convenience comparison groups or to using schools or classrooms as the sampling units and units of analysis. The Gold Standard recommendation of RCTs recognizes these practicalities. However, when stitching the pieces together to reveal the generalized patterns without rigorous application of random sampling and methods, generalization becomes problematic. New approaches that respect the challenges of achieving this ideal are called for.

These issues are not new, and much can be learned from previous considerations of strategies for generalizing across and synthesizing independently conducted research studies. Concerns during the 1970s in education and psychology research identified the need for systematic, unbiased, and trustworthy means of integrating quantitative research results. The call was for strategies to produce generalizations that neither overestimated the value of low-quality studies with weak controls nor underestimated the value of high-quality studies with strong controls (Glass, 2000). A term first coined by Glass (1976), *meta-analysis* is "analysis of analyses … [or] the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (p. 3). Just 10 years after Glass described this process, Bangert-Drowns (1986) noted that meta-analysis "belongs to the fourth class of [research] review, the integrative review" (p. 388). Meta-analysis was introduced to and utilized in science education in an attempt to broaden research approaches and to construct generalizations from the wealth of studies on common reform topics (Anderson, 1983; Anderson, Kahl, Glass, & Smith, 1983). Today, meta-analyses are common in the education, medical, nursing, and psychology research communities.

However, some researchers confuse meta-analysis with systematic reviews or other synthesis studies. Too often, explications of meta-analyses do not focus on the specific statistical process used to combine quantitative data, standardized differences in gain scores, or effect sizes that lead to summary results across numerous studies with similar focus, methods, and outcome and treatment variables. The meta-analysis process was an attempt to find the common strength of relationships (generalizations of sorts, integrations of sorts) across the increasing number of independently conducted, experimental or quasi-experimental research studies about the same or similar popular topics (Bushman & Wang, 1999; Cooper & Hedges, 1994; Hunter & Schmidt, 2004). These included studies of science curriculum and instruction (Shymansky, Kyle, & Alport, 1983; Willett, Yamashita, & Anderson, 1983; Wise & Okey, 1983); factors influencing learning (Wang, Haertel, & Walberg, 1993/94); instructional resources and technologies in writing (Bangert-Drowns, 1993; Ellington, 2003; Goldberg, Russell, & Cook, 2003); reading comprehension (Sencibaugh, 2007); self-beliefs (Ma, 1999; Valentine, DuBois, & Cooper, 2004); writing-to-learn interventions (Bangert-Drowns, Hurley, & Wilkinson, 2004; Graham, 2006; Graham & Perin, 2007); and many other topics.

Meta-analyses draw on numerous independent studies that have generated statistical results regarding effect size on the research problem. While the number of studies included has a wide range (from as small as 4 to over 25), the demand is that the studies are strictly comparable (Cohen, 1988). Each result becomes a unit of analysis that is weighted or unweighted by the sample size in the study to produce a calculation called a summary effect size (H. Cooper, 2003; Hedges, 1994). Although there are no stipulated ranges for a target number, meta-analysis is only possible when reasonable numbers of high-quality and homogeneous studies are available.

Therefore, location and retrieval of research results are important, but selection criteria and quality control are essential. Some advocates of meta-analysis assume that a full range of studies should be included in the database or that quality is not as important since any collection of studies involves indeterminate errors in the results are, most likely, randomly distributed. The inclusion of such errors (+/−) would cancel one another (Glass, 2000). On the other hand, some advocates stress the need for critical selection and identification of quality results as the basic input into any meaningful meta-analysis (Shymansky, Hedges, & Woodworth, 1990). The basic concern here is focused on quality, rigorous, published, and unpublished research studies to overcome the tendency of journals to accept studies with significant results, thereby biasing any collection of studies based only on publication status and leaving many, quality, nonsignificant results in researchers' file cabinets.

Selection criteria for meta-analysis and other forms of research synthesis need to flow from the theoretical foundations of the target problem and research questions and from the standards for high-quality research. The criteria should move beyond limited characteristics, such as the results reach a predetermined level of significance, are published in peer-reviewed journals, or are not graduate theses or dissertations. Studies selected by fair (not prejudiced or biased), consistent (not whimsical), and rigorous (critical and thoughtful) criteria must contain the original information, raw data, or results (means; standard deviations; variance within, between, and residual; or beta values) necessary to calculate composite effect sizes (Hedges & Vevea, 1998; Lipsey & Wilson, 2001; Valentine et al., 2004).

Inference, prediction, deduction, and generalization are the *holy grails* of research. But, H. Cooper (2003) and Glass (2000) cautioned that statistical inference in meta-analysis continues to be a controversial issue. Glass stated, "[T]he chances are remote that the persons or subjects within studies were drawn from defined populations with anything even remotely resembling probabilistic techniques. Hence, probabilistic calculations advanced as if subjects have been randomly selected would be dubious" (p. 10). Glass cautioned the meta-analyst to be sure that the conclusions drawn across the studies are appropriate, given the likely vagaries of sampling. As noted above, randomization permits probabilistic inference; if subjects were sampled through nonprobabilistic methods, the inferences rest on more shaky ground.

Modern technologies have improved the efficiency and potential quality of meta-analysis and other research syntheses in that literature searches and retrievals and follow-up interrogations of authors and researchers are much less laborious than before the advent of the Internet. But the selection procedure continues to be as demanding as ever, and "those who accumulate and integrate other people's data ought to be held to similar standards of methodological rigor as the researchers whose evidence forms the bases of their [synthesis]" (H. Cooper, 2003, p. 3). However, meta-analysis may not be the preferred method of choice if the goal "is to critically appraise a research literature (study by study) or to identify particular studies central to a field[, … where] conceptual and methodological approaches to research on a topic have changed" (pp. 3–4) during the period of consideration, and when targeted studies have used decidedly different methods.

Furthermore, the results of meta-analyses should be applied judiciously and with care to respect the quality (strengths and weaknesses) and limitations of the original studies selected and used to calculate the summary effect sizes. Generalizing beyond the sample of studies must be cautiously undertaken, and high-risk speculations should be discouraged. However, meta-analyses can outline promising agendas to be investigated with further research by providing strength relationships and ideas to help articulate more focused and probing research questions and hypotheses within the problem space. Caution needs to be expressed to organizations and policy makers who attempt to justify, for example, best teaching practices and most effective instructional materials based solely on meta-analysis results.

Smaller clusters of research results—too small in number to justify meta-analysis that are similar to, or replications of, one another and provide access to the original data—afford opportunities for different types of statistical integration. Such a situation becomes a basic problem of data integration and secondary analysis or reanalysis of the collective or unified dataset. For example, Gunel, Hand, and Prain (2007) integrated six studies from an ongoing research program about writing-to-learn science, all with the same basic research design, focus, outcome, and treatment variables using an ANOVA of the collective dataset. These pretest–posttest studies assessed differences in students' science understanding for pairs of treatment and comparison groups. The tests consisted of multiple-choice (recall) and extended-response (conceptual understanding) questions constructed jointly by the teachers and research team. The difference across the studies was that the treatment groups engaged in diverse writing tasks along the writing-to-learn for authentic audiences' continuum while the comparison groups engaged in writing tasks found in most traditional science instruction. Each study attempted to enact reasonable quality controls; that is, attention was paid to the amount of instructional time on a particular topic, and teachers did not teach to the test.

The availability of original data for similar achievement results within a defined problem space makes it possible to conduct a secondary reanalysis by standardizing and combining these datasets into a single dataset representing a reasonably large convenience sample for a more powerful case study. This approach increases the sample size, reduces standard error, avoids accumulation of Type I errors, and provides more efficient, stable, and precise estimates of effect (Hinkle, Wiersma, & Jurs, 2003; Lipsey & Wilson, 2001). Researchers can discover much more information from regenerating the fundamental statistics with the combined dataset than they could with a meta-analysis of the means and standard deviations of the individual studies. The general statistical assumptions involved in this secondary ANOVA (normality, linearity, homogeneity) were addressed using a simple graphical method and normal probability plots of model residuals, plotting standardized residual values against the predicted values and Levene's test for equal variances, respectively. Satisfaction that the data from the separate studies met these assumptions permitted combining the separate datasets into an integrated dataset. ANOVA or *t*-test findings of the unified pretest results across the collective treatment and comparison groups indicated whether an ANOVA, *t*-test, or an ANCOVA should be the chosen statistical method to test the posttest differences to produce a summary effect size for the multiple studies.

While this method of analysis on combined datasets is not common in secondary analyses within educational research, it is used in medical research (Murali et al., 2004; Revicki, Zodet, Joshua-Gotlib, Levine, & Crawley, 2003). Furthermore, as researchers share datasets more frequently—as in the Human Genome Project and other DNA databanks and as recommended for educational research in the US National Research Council (US NRC, 2004) report on advancing scientific research in education—variations and derivatives of this approach will become more common in educational research communities.

26.1.1 The Context: A Need for Balance

Calls for better understanding of available datasets and research results are currently heard in a variety of political, professional, and academic communities. Much of the momentum behind the *Gold Standard for Educational Research* in the United States (US Department of Education, 2003) is about the need for compelling, well-supported generalizations and syntheses—integrations of the findings from a collection of studies—that policy makers can use as foundations for public policy, shaping decisions about public education, educational spending, and future directions. Unfortunately, the Gold Standard privileges quantitative evidence and the results of meta-analyses such as those outlined above to the exclusion of the wealth of high-quality, interpretive, research evidence.

We believe such oversight does not fully recognize education and educational research as a social science that grows both by normal hierarchical development and by the insertion of new theoretical discourses alongside existing ones (Yore & Lerman, 2008). Mathematics, literacy, and science education have benefited from both quantitative and qualitative approaches to knowledge building over the last 30 years. The question is not an either/or issue but one of rigorous and appropriate consideration of multiple approaches that reflect the research question, development of the problem space, and associated research techniques, procedures, and technologies.

Jonathan Osborne (2007), Past President of the National Association for Research in Science Teaching, called for “a bit more armchair science education research” (p. 10), claiming that 50 years of research, curriculum development, and implementation has not presented consistent and compelling patterns of outcomes. His quick inspection of three leading science education journals and Google™ Scholar citations suggested that not enough research synthesis articles have been produced, even when such contributions are highly valued by the science education community. The call for cross-study syntheses, especially those that use qualitative approaches, applies equally well to mathematics and literacy education as to science education (August & Shanahan, 2006b; Firestone, 1993; Yore, 2003).

Similar calls for and examples of such qualitative metasyntheses are found in the health science research communities (Bowman, 2007; Thorne, Jensen, Kearney, Noblit, & Sandelowski, 2004; Zimmer, 2006), but few are found in educational research communities. Sadly, some of the most popular and most recent books on qualitative research used in mathematics and science education do not mention metasynthesis and only briefly consider the general issues of generalizability, if at all, holding to the purists' interpretation of strict contextual restrictions to qualitative research. This is unfortunate in that high-quality, rigorous, naturalistic inquiries are having very limited effect on policy makers and decision makers, who tend to view each study as an isolated *info-bit* anchored strictly to a unique context or educational setting that cannot be applied widely to their target concerns or constituents. Therefore, the very strength of qualitative approaches is considered to be an overwhelming weakness.

We believe this need not be the case. There are several useful approaches to achieve integration, secondary analysis, and synthesis of qualitative research results: research reviews; secondary analyses; case-to-case syntheses of studies with common focus, data sources, and methods, also referred to as meta-ethnographies; and metasyntheses. Fox (2005) suggested that systematic reviews of qualitative research, secondary analyses, and metasyntheses can be useful for increasing interest among policy makers and others in deciding critical issues, policy coverage, and intervention effectiveness in the health sciences. We argue to just such an audience that qualitative research syntheses in education are appropriate and valuable.

26.2 Qualitative Research Syntheses

We return to our metaphor, noting that qualitative research is much like crazy quilting: no matter how expert the sewing and crafting, each unit of design is unique. Application beyond the original situation may not be readily apparent. Qualitative inquiries involve contextualist and organicist worldviews concerned with *events in situ* and "integrated wholeness … making the pieces fit together into an organic whole" (Roberts, 1982, p. 279). Thus, any generalized pattern or application beyond the original context of high-quality studies is typically left to the reader. However, with increasing demands for systematic, insightful research within a problem space, qualitative researchers, we argue, should move beyond a kind of parochialism—a radically local contextualism—to engage more directly in the pressing education policy issues facing society. Entering into that conversation can only be accomplished through the articulation of strategies and procedures for generalizing and synthesizing across the richness of qualitative studies.

26.2.1 The Logics of Generalizing and Synthesizing

Before describing strategies for generating general knowledge across qualitative research studies, it may be useful to distinguish *synthesizing* from *generalizing*—because the processes are related. Generalizing entails applying conclusions (general statements or findings) drawn from one set of circumstances to another set of circumstances. There is a strong predictive element to it; that is, conclusions derived from one study or setting are argued to be predictive of outcomes in other circumstances. Eisner (1991) noted that such general statements allow us to "see our past experiences in a new light" (p. 205).

The notion of generalization, however, has become impoverished in social science discourse, largely because of the hegemonic claims to its definition implied by the Gold Standard criteria for research. The concept has become unnecessarily restricted, "associated with notions of random selection and statistical significance" (Donmoyer, 1990, p. 176), thereby excluding its much more rich, evocative meanings. In its restricted sense, generalizing occurs within specified limits of confidence to the population from which a randomly selected sample was drawn; that is, the results of the inquiry can be applied to the larger population, given identified limits. Most often, however, research report consumers generalize the results far beyond the original population, relying on a more elaborate concept of generalization.

As an example, imagine that we identify the population of interest for our study as middle school students in out-of-school learning programs. We randomly select a treatment sample and a control group from this population and then conduct some experiment. However, because we do not have the resources to draw our sample from across the entire country, we limit the population to middle school students in a local metropolitan area. We conduct the experiment impeccably, draw conclusions, and then want to generalize them. However, we can only probabilistically generalize the findings to the population of middle school students in the host city.

After we publish our results, a science educator in another part of the same country is interested in learning from our research. Can the findings be of interest to that person? Yes. Can they be useful in designing new programmatic initiatives? Surely. But are the findings from our study strictly generalizable to comparable urban populations in this different part of the same country? Not according to the logic of statistical inference. But the logics of analogy and of comparison and contrast allow the potential user to determine if the results of our study will be useful to his or her particular interests. And the writer of the experimental research report can identify those domains to which her or his findings can be fruitfully applied. Thinking about how research results illuminate other, similar circumstances is a softer, more humble, yet richer concept of generalization than the restrictive notion. As Eisner (1991) noted, "whether produced through statistical studies or through case studies, [generalizations in education] need to be treated as tentative guides, as ideas to be considered, not as prescriptions to follow" (p. 209).

From the above example, it becomes clear that the notion of generalizing has at least two definitions of interest here; even in statistically driven studies, it involves two decision spans (Cornfield & Tukey, 1956). One applies findings from the sample on which the study was conducted to the population from which that sample was drawn (assuming randomization and within specified confidence limits): the logic of probabilities. The other logic—that of analogy—applies those findings to another population or set of circumstances "believed or assumed to be sufficiently similar to the study sample that findings apply there as well" (Kennedy, 1979, p. 665). Also described as assertorial logic, this form of argumentation asserts or affirms that something is so and draws on supportive evidence to convince the reader that conditions in the new circumstances are sufficiently similar to the original research conditions for generalization to be appropriate.

In contrast, synthesizing is a process of putting together parts into a whole, the formation of something complex from simpler elements. A synthesis is complete unto itself. The concept of synthesis suggests that the result of the synthesizing process is different from and more complex than a mere aggregation of component parts. In chemistry, it means the creation of a complex compound by combining simpler elements; thus, the process results in the creation of something new. As Strike and Posner (1983) described it, synthesis "involves some degree of conceptual innovation, or employment of concepts not found in the characterization of the parts as means of creating the whole" (p. 346).

These processes entail working from textual material as the writer integrates the disparate cases under consideration into a new understanding of the subject. Related to qualitative data analysis and research review development, syntheses identify general patterns, themes, metaphors, and images across the cases through the processes of comparison and contrast. Patton (1990) described syntheses of disparate qualitative studies as "a form of cross-case analysis … [but notes that these should be] much more than a literature review" (p. 425). Similarly, in one of the definitive works on synthesizing cases, Noblit and Hare (1998) noted the link between syntheses and literature reviews but claimed that the latter are all too often "the study-by-study presentation of questions, methods, limitations, findings, and conclusions [that] lack[s] some way to make sense of what the *collection* of studies is saying" (pp. 14–15).

If we examine the literature on literature reviews, however, we find important parallels to syntheses across cases. H. Cooper (1988) provided a taxonomy of literature reviews, defining two goals of integrative reviews as "synthesizing knowledge from different lines of research [and] inferring generalizations from a set of studies [or] formulating general statements from multiple specific instances" (p. 108, citing Strike & Posner, 1983). While distinctions are made between generalizing and synthesizing, they are clearly related processes, which entail identifying general themes, patterns, metaphors, or "lessons learned" (Patton, 2002, p. 220) from the disparate cases and creating a new framework for understanding the subject.

More closely related to inferring and drawing conclusions than to generalizing, synthesis does not have the explicit predictive meaning that generalizing carries. Having said this, however, it is important to acknowledge that synthesizing also connotes the fuller definition of generalizing outlined above. That is, having developed general statements that synthesize the salient elements, conditions, and qualitative causal models (explanations) of a set of cases, future application to other circumstances is often presumed; and such applicability is one criterion of the value of the synthesis, especially in evaluation work (Guba & Lincoln, 1989; Patton, 1990). The logical processes of syntheses are inductive (inferring more general statements from disparate cases), analogic (distinguishing the cases through comparison and contrast), and interpretive (creating new meaning that integrates the cases into a new whole).

The remainder of this chapter invokes our earlier metaphor of crazy quilting. We offer four strategies for stitching together the pieces of qualitative research to reveal generalized patterns that can inform policy making, programmatic design decisions, and practice within schools and classrooms: research reviews, secondary reanalyses, case-to-case comparisons, and metasyntheses. These strategies can be used to develop generalizations and syntheses across qualitative studies that focus on similar issues and use similar or common methodologies to more fully document, map, describe, and address the problem space. Note that all such approaches rely on the logic of comparison and contrast, drawing from independently conducted studies to detect similarities and differences and to verify the criticality of detected attributes. They also rely on analogic reasoning where multiple sources of evidence are used to support preliminary knowledge claims or working understandings within the situated and conditional limits of the contextualist and organicist worldviews.

Each strategy discussed in this section differs in emphasis and methodology, but all have the overarching purpose of building knowledge across a set of qualitative studies. And each offers promise to add value to existing scholarship, clarify knowledge claims and understandings, identify promising research agendas and areas of inquiry missing in the extant literature, and suggest generalized assertions and applications across wider contexts. We begin by discussing research reviews, followed by secondary reanalyses, case-to-case comparisons, and finally metasyntheses.

26.2.1.1 Research Review

Research reviews are critical summaries and interpretations of the available research literature on a specific topic. Available in journals specifically dedicated to reviews (e.g., *Psychological Bulletin* and *Review of Educational Research*), such critical summaries are wanted and frequently cited by other researchers to capture the background of specific issues and to map the territory of inquiry. These reviews provide in-depth and readily accessible references to readers (Osborne, 2007) to ascertain the current state of knowledge within a field. While there are many typologies of research reviews (see, e.g., H. Cooper, 1984, 1988; Kennedy, 2007), these can be categorized into four overall types:

> The first type of review identifies and discusses new developments in a field. The second uses empirical evidence to highlight, illustrate, or assess a particular theory or to tentatively propose new theoretical frameworks. Third, a reviewer can organize knowledge from divergent lines of research. (Bangert-Drowns, 1986, p. 388)

Bangert-Drowns goes on to identify statistical meta-analyses (discussed above) as belonging to “the fourth class of review, the integrative review” (p. 388). In addition, research reviews can focus on theory, methodology, or findings, or some combination.

Somewhat simplifying the development of review typologies, Bowman (2007) pointed out that there are two types of qualitative reviews: nonsystematic and systematic. The nonsystematic review provides a broad stroke to the background that touches all the bases, much like the traditional background chapters in graduate theses and dissertations. At worst, these reviews are loosely connected summaries clustered under major headings; they frequently provide little added value, serving more as annotated bibliographies than as critical reviews that provide new insight. At best, such reviews reconceptualize the knowledge produced about a field, setting directions for future research as well as providing a Google™ Earth-quality mapping of the terrain.

However, Kennedy (2007) noted that the adjective *systematic* has been appropriated recently, given the pressures of the Gold Standard, to stipulate a review that focuses on a narrowly specified research question, often relying on RCT-type studies. She provided a critique of the term *nonsystematic*, noting that the term "implies deficiency" (p. 139). She argued for a more inclusive conceptualization, showing how the *Review of Educational Research* (the coin of the realm for review articles in education) lists "integrative reviews, theoretical reviews, methodological reviews, and historical reviews" (p. 139) as appropriate for that journal.

As an example of a systematic review of the more inclusive kind, Yore, Bisanz, and Hand (2003) reviewed 25 years of language arts in science education research to celebrate the 25th anniversary of the *International Journal of Science Education* and to honor its contributions in sustaining this area of research. The historical review incorporated parallel analyses by a team of established researchers of oral discourse, reading, and writing in science education that captured both qualitative and quantitative studies emphasizing the contributions of the host journal. The selected studies were systematically segregated into the early and late years of the 1978–2003 period in an attempt to detect the influences of changing theories of learning and models of reading and writing. Without such consideration, the research review would have integrated the results across 25 years, thereby missing current trends and conceptualizations within the historical noise of the early years.

Specifically describing reviews of qualitative research, Bowman (2007) argued that "[s]ystematic reviews are a form of research" (p. 171) that integrates and synthesizes a selective body of qualitative research. Such reviews require thoughtful deliberation, critical analysis, and narrative descriptions to identify the central issues and draw overall conclusions from the primary sources. The synthesis process typically involves five recursive and dynamic stages (Bowman; H. Cooper, 2003): (a) formulation of problem focus; (b) source identification, selection, and collection; (c) information extraction and evaluation; (d) analysis and interpretation of these data; and (e) summary and presentation of results. The focus is central to any synthesis; therefore, it must be clearly articulated and shared within the community of discourse. Source identification, selection, and collection entails mapping the available research literature and then relying on selection criteria to identify and categorize qualitative studies with common or similar focus, data sources, data collection, data interpretation, and outcomes. Information extraction involves a continuous consideration of the quality of the work and its potential value to achieve the purpose of the review. The extracted summaries of each study (the unit of analysis) become the data that will be warranted as the evidence for any assertion, knowledge claim, or generalization. The analyses or critical interpretations must be presented as a clear, logical, compelling argument (presentation of results) that is persuasive and soundly based on evidence (Yore, 2003). These processes do not proceed in a linear fashion; in fact, they are recursive, cycling and recycling back through data, interpretations, arguments, and warrants. As Bowman stated, "[s]ynthesists are free to start, stop, backtrack, adjust the methodology, and retrieve data as needed for a thorough examination of the literature" (p. 172).

Thoughtful and systematic research reviews demand a clear explication of their purpose and focus. Does the author intend to critically summarize results? Compare theoretical frameworks? Contrast methods of data collection or analysis? H. Cooper (2003) identified three general purposes for such reviews: (a) offer an integrative discussion that builds generalizations, resolves conflicting perspectives, or builds connections across ideas or concepts; (b) critique existing research reports; and (c) identify central issues or questions (see H. Cooper, 2003, Table 2, p. 7, for conceptual guidelines). He also noted that focus is salient; a review can focus on research results, methods, theories, or applications. Getting clear about both purpose and focus, we argue, is key to a well-conducted research review.

Coverage of the literature surveyed, selection criteria, and selection process, as stated earlier for meta-analysis, is critical and essential in any systematic research review. The criteria must reflect the underlying theoretical constructs being reviewed and standards for high-quality interpretive research. These established and explicit criteria must be applied in a fair, consistent, and rigorous manner to the selection of research results included, excluded, emphasized, and ignored. Again, information communication technologies have improved the efficiency in locating and retrieving research results and clarifying and verifying ideas and assertions with the original authors and researchers, but this might increase the cognitive demands on selection. H. Cooper (2003) suggested that systematic reviews have great potential toward informing practitioners, policy makers, and the general public and that, as such, effective communication with the target audience will require explicit clarity about focus, goals, coverage, and review methods, and less technical terminology and detail, while "paying greater attention to the implications" (p. 5).

26.2.1.2 Secondary Reanalysis

Researchers with access to original data generated from a similar research focus or agenda and data-collection methods across unique settings, informants, or contexts can conduct a secondary analysis, or reanalysis, of the data using a refined or improved lens or interpretive framework. Again, data sharing is becoming more common in scientific communities and has been recommended as a method to improve the quality of educational research (US NRC, 2004).

Anticipating the need for such secondary analyses, McDermott and Hand (2008) reinterpreted the original transcripts from six independent studies of the Science

Writing Heuristic (SWH) using a consistent, improved, interpretive framework afforded them after a lengthy research program into writing-to-learn science, which they applied to the common anchor interview responses, test items, writing samples, and other artifacts. These markers allowed them to trace SWH results across several years of their research agenda, to cluster studies for further examination, and to consolidate the information across several small samples to produce a rather large and sensitive sample size. The secondary reanalysis of the qualitative results relied on a constant-comparison approach of the word documents or text files, which were used to establish common assertions across the group of studies. Their analyses revealed common and consistent results across the studies, much like the results generated through a meta-analysis of the quantitative data (Gunel et al., 2007). We argue that the consolidated results based on a reanalysis of original data from studies with similar research focus can afford greater discovery power than a meta-analysis and will have a higher probability of convincing and persuading stakeholders about the efficacy and effectiveness of this writing-to-learn science approach.

Secondary reanalysis of the combined original data has great potential to present stronger assertions and explanations from qualitative research that will influence policy and decision makers and increase public awareness about evidence-based learning, teaching, curriculum, and data sharing. Some journals require authors to provide their raw data and computer programs, syntax, and coding for quantitative studies and the functional equivalents for qualitative studies with identities and names of informants masked. Disclosure risks related to confidentiality and security issues have presented significant ethical and technical challenges that have limited the attempts at data sharing, which retains their value for secondary reanalysis (M. Cooper, 2007; Sieber, 2006).

We believe that as the ethical and technical challenges are resolved the increased access to combined text files and use of discourse analysis software (e.g., Atlas TI™, Nudist 6™, Nvivo 7™, XSight™), access to combined video files and use of video analysis systems (e.g., StudioCode™, Transanna™, Videograph™), secondary reanalyses of discourse, conversation, and performance will become commonplace. This does not reduce the importance and procedural demands of developing and rigorously applying valid interpretive frameworks to identify coding procedures, classes, and trends from which to build assertions and identify supportive evidence, responses, and performances. The interpretive frameworks should draw from established theoretical foundations to construct analytic frameworks that encourage generalizations and explanations. If the reanalyses of studies across contexts are done well, then qualitative research approaches will produce more robust knowledge claims, have greater impact on educational policies and decisions, and be viewed as evidence-based findings.

26.2.1.3 Case-to-case Comparison

The Gold Standard for education research and program evaluation in the United States is based on stage 3 of a medical drug trial model. It does not recognize the need for studies of individuals or small-sample-size case studies, which are analogous to stages 1 and 2 of drug trials. Single-subject and small case studies avoid unreasonable costs and manage risk in the early development of new drugs or treatments. They provide substantial insight about feasibility and effectiveness before *going to scale*. To contribute to policy dialogues and programmatic decisions, qualitative case-study researchers should employ strategies that build knowledge across the cases, contributing to a broader and deeper understanding of the problem studied.

In education, case studies have recognized the unique sociocultural, sociocognitive, and contextual features of learning, teaching, and assessment. Such studies emphasize uniqueness and context-specificity and do not set out to generate probabilistic generalizations. This is viewed as an asset to qualitative research, providing in-depth portraits or narratives that depict educational processes in action. The underlying epistemological assumptions are quite different from those of the statistically driven generalizations flowing from random sampling, hypothetico-deductive reasoning, and control-experimental studies. However, the challenge remains to build knowledge across such case studies while recognizing their respect for the uniqueness of context.

Several approaches to case-to-case comparison can be found in the literature. Here we discuss two: analytic generalization and case-to-case synthesis. Analytic generalization focuses on the theoretical models shaping qualitative case studies. This approach maps quite neatly onto H. Cooper's (1988) focus on theory for research reviews. Firestone (1993) argued that analytical generalizations across qualitative case studies can be achieved through consideration of the theoretical models and common features across the individual studies. Analytical generalization involves critical reflection about the theoretical framework shaping a case study. In contrast with secondary analyses, it does not focus on determining comparability of samples or groups of learners. Here, theory-based or model-driven predictions are deductively made from the theoretical foundations; these predictions can be tested—supported or rejected based on the results of the individual cases. As Firestone stated, "[a]nalytical generalization attempts to show that a theory holds broadly across a wide variety of circumstances … that is, the conditions under which it applies" or does not apply (p. 17). Analogous to the constant-comparative method in grounded theory (see Charmaz, 2000, 2005; Glaser & Strauss, 1965) in which researchers "build explanatory frameworks that specify relationships among concepts" (Charmaz, 2000, p. 510), this approach is particularly fruitful when seeking generalized conclusions across a set of case studies that, while focusing on a common topic, relied on differing sample sizes and specific methods to generate data.

An example of this approach can be found in the National Science Foundation's *Academies for Young Scientists* initiative. This initiative has funded 16 programs across the United States to build student interest in science, technology, engineering, and mathematics (STEM) fields. K-12 students are provided out-of-school programs (called informal learning opportunities) to "deepen their interest in, understanding of, and career awareness with regard to STEM disciplines" (Center for Informal Learning and Schools, n.d.). These programs vary widely in specific out-of-school activities and target populations. Yet the National Science Foundation is deeply interested in systematically developed conclusions that respond to the working hypothesis of this initiative: if provided with rich, inquiry-oriented learning experiences, students will build interest in pursuing careers in STEM fields. The overall program evaluation focuses, among other assessments, on the analytic constructs and underlying theoretical principles about informal learning to build explanations across the somewhat disparate cases.

Case-to-case synthesis involves the consideration of independent cases with a common focus, method, or outcomes as individual cases in a multicase study (Florence & Yore, 2004; Rossman, 1993; Yin, 2003). The synthesis is intended to build integrative understanding of the problem space taken up in the independent case studies. Stake (1995) suggested that researchers can explore several situations in which a common or similar phenomenon, event, or population occurs and can consider the combined cases as the collective case. An example comes from evaluation interests of philanthropic organizations where funding initiatives focus on a variety of interest areas, rely on differing implementation strategies with differing populations, and have outcomes specific to the focus. Yet, the problem space identified by the theoretical foundation is the evaluation question: Are our funding streams effective in achieving our goals? In this instance, the cases could be differing programmatic initiatives: out-of-school science experiences for middle school children and intensive summer professional development for mathematics teachers. The funding agencies seek conclusions about effectiveness across these disparate cases—their various initiatives around STEM. They seek a synthesis across the cases.

Building on Turner's theory of social explanation, Noblit and Hare (1998) proposed a form of synthesis in which the central metaphors of cases are systematically compared with one another. Described as a process of translation, their approach relies on interpretation and reasoning by analogy. Idiomatic translations, rather than literal ones, are compared. Thus, rather than focusing on empirical observations of social practice (literal renditions), the synthesis "conveys the sense of things" (p. 31). The synthesis is achieved when the central metaphors of various cases map fully onto one another.

Because the process is fundamentally interpretive, different researchers will focus on different aspects of the case, reflect on and integrate those accounts into their own differing experiences, and render different syntheses. This relativistic aspect of the synthesizing process is not unlike what we would expect from two different integrative research reviews of the same corpus of studies. Because researchers bring different experiences and conceptual lenses to the task, two reviews of the same body of research would likely be organized differently, emphasize different elements of the texts, and draw different conclusions. In fact, this interpretation is what makes research reviews (and syntheses of case studies) interesting. It validates and celebrates the authorship of the text and raises the resultant work above the mere recitation of previous studies so soundly critiqued by Patton (1990) and Noblit and Hare (1998).

Miles and Huberman (1994) described two central strategies for case-to-case comparisons—case-oriented approaches and variable-oriented approaches—as well as a mixed approach. In the case-oriented approach, one case is analyzed and a grounded theory or working explanation is crafted. This working explanation is then applied to subsequent cases to test out the robustness of the explanation. In the variable-oriented approach, particular themes are identified and compared across cases. In this latter approach, the complexity of specific cases is "bypassed or underplayed" (p. 175) in favor of theme analysis. This disadvantage can be overcome, Miles and Huberman argue, by relying on mixed approaches where some balance is struck between the full analysis of comparative cases and the discrete, more focused analysis of variables or themes.

Dillon, O'Brien, Moje, and Stewart (1994) concluded that research about the problem space dealing with language and literacy in science education had, to date, considered questioning techniques, patterns of verbal interaction, quality of texts, the nature of readers, and how students used reading to learn in science classrooms. However, they noted that research had not addressed how teachers' beliefs about teaching students and science content influenced their use of literacy events in secondary science classrooms and how they selected and structured these events to achieve their content goals. Based on this assessment of the problem space and its development, Dillon and colleagues decided to utilize symbolic interactionism as a theoretical framework and ethnography as a methodology to explore case studies of three secondary science teachers' beliefs, instructional decisions, and implementation of literacy events in science classrooms. Their purpose, focus, foundation, design, and procedures reflected the early developmental status of the problem space, established knowledge about literacy events in secondary science classrooms, and indicated a desire to produce findings that were applicable across more than a single setting.

Dillon and colleagues (1994) conducted separate, 1-year case studies of three teachers, their science classroom and students, and other related school community members. They focused on how teachers' philosophies about teaching students and science content shaped their literacy events in secondary science and how literacy was structured and manifested in science lessons. They collected information about beliefs, events, and actions utilizing field notes, video- and audiotaped lessons, interviews, and instructional artifacts (student work samples, study guides, laboratory sheets, lesson plans). Data from these sources were analyzed as each case study progressed, using constant comparison to detect emerging patterns and categories that were confirmed or negated as additional information was collected and interpreted over the year. Results for each case study were reported for the common trends that developed across the three cases: teacher's philosophies and uses of literacy (as foundation and as facilitator). The case-to-case comparison "consisted of looking for patterns that were similar and different across the three teachers with respect to their teaching philosophies and their literacy practices" (p. 350). Similarities and differences were detected by compare–contrast techniques for philosophies, use of literacy as foundation, and use of literacy as facilitator:

> All three teachers have philosophies of teaching that lead them to create classroom climates in which students are valued. The three teachers care deeply about whether students learn, and they strive to provide a classroom climate in which students can learn. … Although the

> three teachers created structures that are designed to support students, they did so in ways undergirded by markedly different philosophical positions on science and science teaching. These different philosophical positions have a significant effect on how learning is organized, how lessons are framed, and ultimately, how literacy is defined. (p. 358)

Under this generalization, variations in literacy events selected by teachers and utilized in science classrooms across the cases were linked to teachers' beliefs about science.

26.2.1.4 Metasynthesis

Thorne and colleagues (2004) suggested that the pressure for evidence-based health care, which parallels the pressures in education for evidence-based instructional strategies and materials, has promoted scholarly activity called metasynthesis of qualitative research that is distinct from conventional literature reviews, secondary analyses, and other endeavors to deconstruct research studies and construct shared patterns across common treatments. They stated:

> We understand that product to be fundamentally different from the original parts, capable of substantiating a more convincing argument about the major theoretical elements with the phenomenon of interest and positioned to advance the science in that particular substantive field more forcefully. (p. 1343)

Metasynthesis provides an umbrella "mechanism for thinking about qualitative integrations" that brings together, breaks down, and combines findings (not raw data) into transformed results (Finfgeld, 2003, p. 897). The goal of metasynthesis is to:

> produce new and integrative interpretation of findings that is more substantive than those resulting from individual investigations. This methodology allows for the clarification of concepts and patterns, and results in refinement of existing states of knowledge and emergent operational models and theories. (p. 894)

Metasyntheses are reasonably well accepted in medical and health care research, integrating anywhere from 3 to 292 individual research reports (see Table 2 in Finfgeld, p. 896); but similar popularity in literacy, language, and science education research has not been found. Early advocacy for (Yager, 1982) and concerns about (Orpwood, 1983) qualitative synthesis in science education were related to methods of strategic planning and deliberative visioning to establish frameworks, set priorities, and outline future research and development agendas. The National Science Teachers Association's Project Synthesis (Harms & Yager, 1981) and the Science Council of Canada's Deliberative Inquiry (Orpwood & Souque, 1984) provided procedural insights into the use of collaborative teams and focus group validation for synthesis. But they focused more on establishing an assessment of desired state, actual state, and needed improvements in science education curriculum than seeking generalizations across research studies. Therefore, we have relied mostly on health care and nursing researchers for the following insights into metasynthesis of qualitative research results.

Metasynthesis focused on theory building utilizes grounded formal theory and the standard techniques or metastudy of data, methods, and theories that investigate quality, epistemic, philosophical, cognitive, and theoretical issues. This is followed by a synthesis of the results to build general theories across collections of independent studies of the target phenomena (Finfgeld, 2003). Theory explication involves deconstructing, reconstructing, and synthesizing findings across studies focused on a specific theoretical construct. Descriptive metasynthesis addresses broader phenomena by translating results across studies.

Again, procedural steps similar to the other integrative approaches described above apply to metasynthesis: focus, sources, sample size, analysis, and integrity of findings (Finfgeld, 2003). Recognition that a central focus might exist across several independent qualitative studies is an essential first step in metasynthesis.

> This supports the notion that seasoned qualitative researchers recognize metasynthesis as an alternative strategy for moving their work forward rather than continuing to conduct serialized investigations. … Ergo, experienced qualitative researchers are urged to identify studies related to their research interest areas that can be used to push … knowledge forward. (p. 898)

The focus for a metasynthesis needs to be sufficiently defined and delimited to produce meaningful results but broad enough to fully capture the target phenomenon and the surrounding problem space. In education, this would mean that similar studies from a variety of contexts, content areas, or grade levels or studies of similar constructs (such as critical thinking, metacognition, reflective practice) would be included in the problem space and in the associated search of the research literature.

Identifying and selecting relevant qualitative research studies for metasynthesis involves the same concerns expressed earlier for quantitative meta-analyses and research reviews. The identification and selection processes require criteria flowing from standards for qualitative research and argumentation (Finfgeld, 2003) and from the theoretical foundations for the target problem and research questions under consideration. The number of studies (sample size) for a metasynthesis depends on the specific goal of the synthesis: well-defined and limited collections for building grounded, formal theories and larger, more comprehensive collections for metastudies (secondary synthesis of a metadata analysis, metamethod synthesis, and metatheory synthesis of the same collection of qualitative studies to create new theoretical interpretations). Sampling should include high-quality studies from various content domains and demographics to allow generalizability and clarification of constructs. Finfgeld suggested that expert and experienced researchers familiar with and active in the problem space under investigation might require smaller samples to draw valid consolidated claims.

Analysis considers epistemological issues, deconstruction and decontextualization, and relationships amongst findings (Finfgeld, 2003). She stated, “[S]ome researchers object to interpreting findings resulting from different epistemological perspectives because of their variant foci and theoretical structures … [while other] investigators have found this restriction unnecessary, and in fact, they embrace the opportunity to synthesize studies from differing epistemological perspectives”

(p. 900). Recall the earlier description by Bangert-Drowns (1986) that reviews "can organize knowledge from *divergent lines of research* [italics added]" (p. 388).

Analysis in metasynthesis varies across the spectrum of typical strategies for qualitative analysis and interpretation building. Some researchers apply grounded analysis to recontextualize the research findings by moving toward new trends, codes, or assertions flowing from the findings while others apply predetermined codes derived from the theoretical frames to reinterpret each set of findings in a stepwise, recursive fashion (see Rossman & Rallis, 2003, for a discussion of open-ended or prefigured coding practices). Still others immediately move toward synthesis, consolidation, and unification of the findings from the metaphors identified. Data analysis ascertains the degree of support or refutation amongst findings under consideration. A collection of independent findings that split along supportive and oppositional lines will require distinctively different analysis than collections that are either overwhelmingly supportive or refutational.

Integrity of findings can be improved by utilizing research teams, focus groups and open deliberations, triangulation, supporting evidence, audit trails, and assessing truth value (Finfgeld, 2003). Metasyntheses are labor-intensive and demand diverse expertise across a variety of research methodologies and theoretical constructs related to the target areas. A research team composed of diverse and distributed expertise could address these demands. Sharing preliminary metaresults with informed critics as a focus group or researchers of the selected studies to deliberate, verify, and check the consolidated results does much to ensure integrity (Orpwood & Souque, 1984; Yager, 1982). Integrity also flows from the argument provided in the metasynthesis where knowledge claims are supported by original data results or respondent quotations from the selected studies. Explicit descriptions of the procedures and criteria for identifying, selecting, and analyzing research studies and their associated findings are essential to integrity. Brief summaries of the selected studies in an appendix, if space allows, or a searchable database at a journal or personal Web site allow readers to assess integrity for themselves.

> Knowledge development is iterative in nature; thus, the process of verifying metasynthesis findings will undoubtedly follow this pattern. As findings are published and cautiously scrutinized, applied, and tested, their ultimate truth value will be affirmed or dispelled. When the latter occurs, additional primary qualitative studies may be called for, or ongoing metasyntheses may be conducted using different interpretive lenses. (Finfgeld, p. 902)

We could find few examples in education. However, one comes from Bair's (1999) synthesis of 118 qualitative inquiries completed between 1970–1998 regarding doctoral student attrition and persistence. She relied on meta-ethnographic synthesis techniques (Noblit & Hare, 1998) to design and guide the articulation of selection criteria, identification, and translation of "each study selected into each other study" (p. 8). Inductive integration was used to analyze the findings recursively. Bair summarized each study selected and verified by external referees, assessed how each study was related in a matrix of key findings, and established analogous connections between studies "juxtaposed, cross-compared, and integrated [to reveal] common findings, similarities and contradictory findings" (pp. 13–14).

Emergent themes and overarching constructs emerged as columns and cells converged and were consolidated.

A second, more extended example in education comes from literacy studies. The National Literacy Panel on Language-Minority Children and Youth (composed of distinguished scholars from Canada and the United States) utilized meta-analysis, secondary analysis, and systematic interpretation of quantitative and qualitative research results to address the development of literacy amongst learners whose home language (L1) was not the language of majority and instruction (L2), mainly English (August & Shanahan, 2006b). This project attempted "to identify, assess and synthesize research on the education of language-minority children and youth with respect to their attainment of literacy" (August & Shanahan, 2006a, p. 1). The resulting report and searchable database were notable because they illustrated many of the recommendations of the 2nd Island Conference: clarity, procedural rigor, shared database, effective use of existing data and information, and the production of generalizations across a problem space and related research studies. The report explicitly outlined the general research questions for the panel and the specific research foci for each of the five working subcommittees, the theoretical framework and procedures for the review (definitions of the variables, information sources, selection criteria, search procedures, studies identified, coding rubrics, external verification, and analyses), and the generalizations asserted. The findings identified the need to develop precursor oral and print skills, the importance of L1 proficiency and individual attributes, and the surprising outcomes involving assessment practices, teacher judgments, and sociocultural influences.

The transparency of purpose, focus, procedures, and outcomes, as outlined in this chapter, are essential to allow open and full evaluation of the results. Grant, Wong, and Osterling (2007) provided such a review; they criticize the sociocognitive interpretive framework and traditional definition of literacy, summarizes the findings, provides an alternative framework, and outlines implications from a critical literacy perspective. Such reactions, rebuttals, and counterclaims are expected and encouraged by secondary analysis and synthesis—in fact, by all research—because it is within such critical discourse problem spaces that knowledge is expanded.

The methodologies used across the five subcommittees involved a variety of synthesis techniques resulting in six general findings (Grant et al., 2007):

- Instruction focused on phonemic awareness, phonics, fluency, vocabulary, and text comprehension was beneficial to the target students.
- Print-focused instruction was necessary, but oral proficiency was also important.
- Oral proficiency in the students' L1 can facilitate L2 learning.
- Individual differences produce significant effects on English language development.
- Many assessments generally do not provide useful insights into individuals' language resources and needs.
- Sociocultural factors revealed little effect on English language learning.

These generalizations do not match the L2-only approach of some jurisdictions and the social justice agenda of some critical literacy researchers. Grant and colleagues' review of this report provided an explicit context for their rebuttal and alternative heteroglossic, sociocultural, and multidimensional framework. This, in turn, may influence the selection of studies, synthesis techniques, interpretation of the included studies and the results, and counterclaims worthy of consideration. Their consideration of the heteroglossic nature of biliteracy can be informative to science literacy research focused on moving learners from L1 to L2 and onto L3 (language of science) in the three-language problem of being a science language learner (Yore, Chinn, & Hand, 2008; Yore & Treagust, 2006). Grant and colleagues stated:

> Understanding the nature and extent of cross-language effects in the acquisition of literacy is critical. … In contrast to monolingual English-speaking students, language-minority students bring an additional set of resources or abilities and face an additional set of challenges when learning to read and write in English as a second language [and scientific English as a third language]. (p. 601)

26.3 Closing Remarks

There are many similarities among medical, nursing, health care, literacy, language arts, and science education research in terms of pressures for evidence-based practices and external-driven questions about the quality, utility, and practicality of the research evidence flowing from these communities. Furthermore, high-quality qualitative research results are having little impact on policy and program decision makers since findings are viewed as isolated info-bits applying only to unique contexts and not applicable to these stakeholders' situations. Each of these research communities operates within discourse fields that valorize RCTs and devalue qualitative studies. Specifically, each operates under the externally driven belief in the hierarchical quality of findings flowing from random field or clinical trials and measurements, the internally imposed exclusion of qualitative research findings from considerations of best practices, and the qualitative research purists' beliefs that situational and contextualized inquiry results cannot and should not be integrated (Sandelowski, 2004). Compounding this, the sometimes unique and creative representations (dramas, plays, poems, stories, etc.) used by qualitative researchers to describe relationships make potential synthesis with more traditional representational modes difficult or impossible (Annells, 2005). However, researchers who wish to increase the potential impact of their findings need to anticipate synthesis and provide common markers or reasonable connections to other research studies for such integration to occur.

"[U]nlike folklorists, … researchers are obliged to make the utility of stories explicit" and the messages, arguments, and claims clear (Sandelowski, 2004, p. 1377). Sandelowski stated:

> [Qualitative integration] presents dilemmas that researchers have yet fully to recognize, address, and resolve. Most notable among these challenges are (a) distinguishing qualitative studies from other species of research, (b) distinguishing qualitative metasynthesis for other species of synthesis or narrative reviews of the literature, (c) locating relevant qualitative studies for inclusion in bibliographic samples, (d) understanding research reports written in diverse discipline-specific styles, (e) locating the findings in these reports, (f) classifying these findings, (g) determining which findings are about the same target phenomenon or event, (h) determining which findings merit inclusion, (i) deciding which methods and techniques to use to combine different kinds of findings, (j) determining what form the product of analysis should take, and (k) determining how best to present this product to showcase its relevance for a target audience. (p. 1379)

She then cautioned that:

> Increasing publication of reports of studies designated as qualitative metasynthesis that are little more than conventional literature reviews is generating new concerns that qualitative metasynthesis is becoming the latest methodological fad to attract would-be researchers eager for an easy entrée into research and qualitative research, in particular. (p. 1379)

We have outlined a few strategies for such integration and provided some examples from educational and health care research of how to integrate qualitative research results, but there are likely other types of cross-study integrations and metasyntheses that we have not mentioned. Furthermore, there are no firm guidelines for many of these approaches. Some groups of health care researchers are maintaining web-based projects to provide a forum for qualitative synthesis and for interested researchers to share ideas and resolve common concerns, issues, and problems (see http://www.joannabriggs.edu.au/cqrmg/index.html and http://www.unc.edu/~msandelo/handbook for two examples).

The critical demand for qualitative integration at this time is to recognize the limited impact of high-quality qualitative inquiries and the foolishness of some researchers who turn out numerous replications of a given inquiry that do not appear to move the collective understanding and knowledge forward. We sense that the next consideration will need to be more closely articulated strategies for systematic integration of a full range of quantitative, qualitative, and mixed-methods studies to fully capture the evidence about specific issues and problems. The space limitations for journals and the required elaborations needed for research integrations can be partially addressed by journal or personal Web sites to store searchable databases, appendices, and elaborated information about the selection criteria, studies considered, and procedures used.

Lopes and colleagues (2008) conducted such an innovative, secondary analysis/synthesis of a mixture of qualitative and quantitative studies that illustrates the evolving use of techniques to find common patterns and potential generalizations across independent studies of similar research questions within a common problem space. They located a corpus of studies dealing with science teaching and learning across a variety of topics, teachers, and grade levels published during 2000 and 2001 in the three leading science education research journals (*International Journal of Science Education*, *Journal of Research in Science Teaching*, and *Science Education*). The selection criteria (practical relevance, curriculum design, and formative situations) were formulated from an analysis of the literature and research findings on science teaching from the European tradition of didactics. These three dimensions were further disaggregated into 23 variables for analysis. The researchers used these criteria to identify 35 studies. The selection process focused on keywords generated from the literature review and was multilayered, involving cross-verification amongst the researchers. The analytical frame was developed by crafting a series of critical questions that could be addressed with a binary response: yes (1) or no (0). This framework was validated by multiple considerations of a reference set of studies involving pairs of the six researchers. The analytical frame was applied to the selected studies resulting in a 35 × 23 matrix of results. These data were cluster-analyzed using a software program producing linked variables that were more like those included in the cluster than those not included in the cluster. This meta-interpretative synthesis revealed that global practical relevance, curriculum design, and formative situations formed transversal traits common to several independent studies and across the complexity of science teaching and learning. These researchers were rigorous and justified the criteria within established knowledge stores, explored stability of results with multiple analyses of subsets of the studies, shared the listing of studies involved, and expressed appropriate tentativeness with hedges regarding their knowledge claims. The transparent approach and shared data sources allow readers to assess the validity of the results.

We echo the call from Estabrooks, Field, and Morse (1994) over a decade ago to move beyond "one-shot [research studies towards inquiry agendas that address the] incremental business of accumulating knowledge" (p. 510). Our scholarly communities can no longer endorse or avoid rejecting the senseless repetition of *cookie-cutter* inquiries that do not appear to benefit from the inquiries that have preceded them—those who are not aware of the prior research, history, and canonical wisdom that precede an event are destined to repeat the mistakes that occurred earlier. Much qualitative research in health sciences and education is infrequently consulted and has little influence on policies and decisions (Sherwood, 1999). Sandelowski (2004) cautioned researchers that many metasyntheses of qualitative studies add little to extant knowledge and are little more than literature reviews. We believe that qualitative integration has much to offer in producing meaningful generalizations, presenting insightful syntheses, outlining necessary future inquiries, identifying generative theories, and—most importantly—getting policy and decision makers to take qualitative results seriously as evidence on which to base future educational policies and programmatic decisions.

References

Anderson, R. D. (1983). A consolidation and appraisal of science meta-analyses. *Journal of Research in Science Teaching, 20*(5), 497–509.

Anderson, R. D., Kahl, S. R., Glass, G. V., & Smith, M. L. (1983). Science education: A meta-analysis of major questions. *Journal of Research in Science Teaching, 20*(5), 379–385.

Annells, M. (2005). A qualitative quandary: Alternative representations and meta-synthesis [Guest editorial]. *Journal of Clinical Nursing, 14*(5), 535–536.

August, D., & Shanahan, T. (2006a). Introduction and methodology. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the national literacy panel on language-minority children and youth* (pp. 1–42). Mahwah, NJ: Lawrence Erlbaum.

August, D., & Shanahan, T. (Eds.). (2006b). *Developing literacy in second-language learners: Report of the national literacy panel on language-minority children and youth*. Mahwah, NJ: Lawrence Erlbaum.

Bair, C. R. (1999). *Meta-synthesis*. (ERIC Document Reproduction Service ED437866).

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin*, *99*(3), 388–399.

Bangert-Drowns, R. L. (1993). The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. *Review of Educational Research*, *63*(1), 69–93.

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research*, *74*(1), 29–58.

Bowman, K. G. (2007). A research synthesis overview. *Nursing Science Quarterly*, *20*(2), 171–176.

Bushman, B. J., & Wang, M. C. (1999). *Integrating results through meta-analytic review using SAS software*. Cary, NC: SAS Institute.

Center for Informal Learning and Schools. (n.d.). *National Science Foundation academies for young scientists (NSFAYS)*. Retrieved June 5, 2008, from http://cils.exploratorium.edu/ays/

Charmaz, K. (2000). Grounded theory: Objectivist and constructivist methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative inquiry* (2nd edn., pp. 509–535). Thousand Oaks, CA: Sage.

Charmaz, K. (2005). Grounded theory in the 21st century: Applications for advancing social justice studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative inquiry* (3rd edn., pp. 507–535). Thousand Oaks, CA: Sage.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edn.). Hillsdale, NJ: Lawrence Erlbaum.

Cooper, H. (1984). *The integrative research review*. Beverly Hills, CA: Sage.

Cooper, H. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society*, *1*(1), 104–126.

Cooper, H. (2003). [Editorial]. *Psychological Bulletin*, *129*(1), 3–9.

Cooper, H., & Hedges, L. V. (Eds.) (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Cooper, M. (2007). Sharing data and results in ethnographic research: Why this should not be an ethical imperative. *Journal of Empirical Research on Human Research Ethics, 2*(1), 3–19.

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *The Annals of Mathematical Statistics*, *27*(4), 907–949.

Dillon, D. R., O'Brien, D. G., Moje, E. B., & Stewart, R. A. (1994). Literacy learning in secondary school science classrooms: A cross-case analysis of three qualitative studies. *Journal of Research in Science Teaching*, *31*(4), 345–362.

Donmoyer, R. (1990). Generalizability and the single-case study. In E. W. Eisner & A. Peshkin (Eds.), *Qualitative inquiry in education: The continuing debate* (pp. 175–200). New York: Teachers College Press.

Eisner, E. W. (1991). *The enlightened eye: Qualitative inquiry and the enhancement of educational practice*. New York: Macmillan.

Ellington, A. J. (2003). A meta-analysis of the effects of calculators on students' achievement and attitude levels in precollege mathematics classes. *Journal for Research in Mathematics Education*, *34*(5), 433–463.

Estabrooks, C. A., Field, P. A., & Morse, J. M. (1994). Aggregating qualitative findings: An approach to theory development. *Qualitative Health Research*, *4*(4), 503–511.

Finfgeld, D. L. (2003). Metasynthesis: The state of the art—so far. *Qualitative Health Research, 13*(7), 893–904.

Firestone, W. A. (1993). Alternative arguments for generalizing from data as applied to qualitative research. *Educational Researcher*, *22*(4), 16–23.

Florence, M. K., & Yore, L. D. (2004). Learning to write like a scientist: Coauthoring as an enculturation task. *Journal of Research in Science Teaching*, *41*(6), 637–668.

Fox, D. M. (2005). Evidence of evidence-based health policy: The politics of systematic reviews in coverage decisions. *Health Affairs*, *24*(1), 114–122.

Glaser, B. G., & Strauss, A. L. (1965). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8.

Glass, G. V. (2000). *Meta-analysis at 25*. Retrieved September 19, 2007, from http://glass.ed.asu.edu/gene/papers/meta25.html

Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A meta-analysis of studies from 1992 to 2002. *Journal of Technology, Learning, and Assessment, 2*(1). Retrieved from http://escholarship.bc.edu/jtla/vol2/1/

Graham, S. (2006). Strategy instruction and the teaching of writing: A meta-analysis. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 187–207). New York: Guilford.

Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools — A report to Carnegie Corporation of New York*. Washington, DC: Alliance for Excellent Education. Available from http://www.carnegie.org/literacy/pdf/writingnext.pdf

Grant, R. A., Wong, S. D., & Osterling, J. P. (2007). Developing literacy in second-language learners: Critique from a heteroglossic, sociocultural, and multidimensional framework [Essay book review]. *Reading Research Quarterly*, *42*(4), 598–609.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

Gunel, M., Hand, B., & Prain, V. (2007). Writing for learning in science: A secondary analysis of six studies. *International Journal of Science & Mathematics Education*, *5*(4), 615–637.

Harms, N., & Yager, R. E. (1981). *What research says to the science teacher* (Vol. 3). Washington, DC: National Science Teachers Association.

Hedges, L. V. (1994). Statistical considerations. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 29–38). New York: Russell Sage Foundation.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504.

Hinkle, T., Wiersma, W., & Jurs, S. (2003). *Applied statistics for the behavioral sciences* (5th edn.). New York: Houghton Mifflin.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd edn.). Thousand Oaks, CA: Sage.

Kennedy, M. M. (1979). Generalizing from single case studies. *Evaluation Review, 3*(4), 661–678.

Kennedy, M. M. (2007). Defining a literature. *Educational Researcher, 36*(3), 139–147.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Lopes, J. B., Silva, A. A., Cravino, J. P., Costa, N., Marques, L., & Campos, C. (2008). Transversal traits in science education research relevant for teaching and research: A meta-interpretative study. *Journal of Research in Science Teaching, 45*(5), 574–599.

Ma, X. (1999). A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics. *Journal for Research in Mathematics Education, 30*(5), 520–540.

McDermott, M. A., & Hand, B. (2008, January). *A secondary analysis of writing-to-learn studies in science: Focus on the student voice*. Paper presented at the international meeting of the Association for Science Teacher Education, St. Louis, MO.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd edn.). Thousand Oaks, CA: Sage.

Murali, N. S., Murali, H. R., Auethavekiat, P. R., Erwin, P. J., Mandrekar, J. N., Manek, N. J., et al. (2004). Impact of Futon and Naa bias on visibility of research. *Mayo Clinic Proceedings, 79*(8), 1001–1006.

Noblit, G. W., & Hare, R. D. (1998). *Meta-Ethnography: Synthesizing qualitative studies* (Vol. 11). Thousand Oaks, CA: Sage.

Orpwood, G. W. F. (1983). Comments on "Factors involved with qualitative syntheses: A new focus for research in science education". *Journal of Research in Science Teaching, 20*(4), 369–371.

Orpwood, G. W. F., & Souque, J.-P. (1984). *Science education in Canadian schools: Introduction and curriculum analyses* (Vol. 1). Ottawa, Ontario, Canada: Science Council of Canada.

Osborne, J. (2007). In praise of armchair science education. *E-NARST News, 50*(2). Retrieved from http://www.narst.org/news/e-narstnews_july2007.pdf

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd edn.). Newbury Park, CA: Sage.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd edn.). Thousand Oaks, CA: Sage.

Revicki, D. A., Zodet, M. W., Joshua-Gotlib, S., Levine, D., & Crawley, J. A. (2003). Health-related quality of life improves with treatment-related GERD symptom resolution after adjusting for baseline severity. *Health and Quality of Life Outcomes, 1*(73). Retrieved from http://www.hqlo.com/content/1/1/73. doi:10.1186/1477-7525-1-73

Roberts, D. A. (1982). The place of qualitative research in science education. *Journal of Research in Science Teaching, 19*(4), 277–292.

Rossman, G. B. (1993). *Building explanations across case studies: A framework for synthesis*. (ERIC Document Reproduction Service ED373115).

Rossman, G. B., & Rallis, S. F. (2003). *Learning in the field* (2nd edn.). Thousand Oaks, CA: Sage.

Sandelowski, M. (2004). Using qualitative research. *Qualitative Health Research, 14*(10), 1366–1386.

Sencibaugh, J. M. (2007). Meta-analysis of reading comprehension interventions for students with learning disabilities: Strategies and implications. *Reading Improvement, 44*(1), 6–22.

Sherwood, G. (1999). Meta-synthesis: Merging qualitative studies to develop nursing knowledge. *International Journal for Human Caring, 3*(1), 37–42.

Shymansky, J. A., Hedges, L. V., & Woodworth, G. (1990). A reassessment of the effects of inquiry-based science curricula of the 60's on student performance. *Journal of Research in Science Teaching, 27*(2), 127–144.

Shymansky, J. A., Kyle, W. C., Jr., & Alport, J. M. (1983). The effects of new science curricula on student performance. *Journal of Research in Science Teaching, 20*(5), 387–404.

Sieber, J. E. (2006). Introduction: Data sharing and disclosure limitation techniques. *Journal of Empirical Research on Human Research Ethics, 1*(3), 47–50.

Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage.

Strike, K. A., & Posner, G. J. (1983). Types of synthesis and their criteria. In S. Ward & L. Reed (Eds.), *Knowledge structure and use: Implications for synthesis and interpretation* (pp. 343–362). Philadelphia: Temple University Press.

Thorne, S., Jensen, L., Kearney, M. H., Noblit, G. W., & Sandelowski, M. (2004). Qualitative metasynthesis: Reflections on methodological orientation and ideological agenda. *Qualitative Health Research, 14*(10), 1342–1365.

United States Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Washington, DC: Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Available from http://ies.ed.gov/ncee/pubs/evidence_based/evidence_based.asp

United States National Research Council. (2004). *Advancing scientific research in education*. Committee on Research in Education. L. Towne, L. L. Wise, & T. M. Winters (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist*, *39*(2), 111–133.

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993/1994). Synthesis of research: What helps students learn? *Educational Leadership*, *51*(4), 74–79.

Willett, J. B., Yamashita, J. J. M., & Anderson, R. D. (1983). A meta-analysis of instructional systems applied in science teaching. *Journal of Research in Science Teaching*, *20*(5), 405–417.

Wise, K. C., & Okey, J. R. (1983). A meta-analysis of the effects of various science teaching strategies on achievement. *Journal of Research in Science Teaching*, *20*(5), 419–435.

Yager, R. E. (1982). Factors involved with qualitative syntheses: A new focus for research in science education. *Journal of Research in Science Teaching*, *19*(5), 337–350.

Yin, R. K. (2003). *Case study research: Design and methods* (3rd edn.). Thousand Oaks, CA: Sage.

Yore, L. D. (2003). Quality science and mathematics education research: Considerations of argument, evidence and generalizability [Guest editorial]. *School Science & Mathematics*, *103*(1), 1–7.

Yore, L. D., Bisanz, G. L., & Hand, B. (2003). Examining the literacy component of science literacy: 25 years of language arts and science research. *International Journal of Science Education*, *25*(6), 689–725.

Yore, L. D., Chinn, P. W. U., & Hand, B. (Eds.) (2008). Science literacy for all: Influences of culture, language, and knowledge about nature and naturally occurring events [Special Issue]. *L1—Educational Studies of Language & Literacy*, *8*(1). Retrieved from http://l1.publication-archive.com/public?fn = enter&repository = 1

Yore, L. D., & Lerman, S. (2008). Metasyntheses of qualitative research studies in mathematics and science education [Editorial]. *International Journal of Science & Mathematics Education*, *6*(2), 217–223.

Yore, L. D., & Treagust, D. F. (2006). Current realities and future possibilities: Language and science literacy—empowering research and informing instruction. *International Journal of Science Education*, *28*(2/3), 291–314.

Zimmer, L. (2006). Qualitative meta-synthesis: A question of dialoguing with texts. *Journal of Advanced Nursing*, *53*(3), 311–318.

Chapter 27
The Gold Standard and Knowing What to Do

Stephen P. Norris, Linda M. Phillips, and John S. Macnab

The call for evidence-based educational practice presumes that science is a way to good knowing and often presumes as well that good knowing leads more or less directly to good acting. We will not critique science as a means to good knowing, particularly regarding the effectiveness of educational interventions. Rather, we shall urge educators to pay more attention to the relationship between scientific knowledge and what can be done with that knowledge. Providing an accurate view of this relationship is critically important to how science can serve as a vehicle for change in social practice. “At issue are the potency and value ascribed to certain forms of evidence in supporting propositions that arise in educational practice” (Thomas, 2004, p. 1).

Much of the impetus for the recently revived debate about the role of scientific evidence in education stems from two pieces of legislation passed in the United States. The first is the No Child Left Behind Act of 2001 (NCLB, 2002). However, the second, the Education Sciences Reform Act of 2002 (ESRA, 2002), is more important to our business here. ESRA established four new centers in the US Department of Education (US ED): The Institute of Education Sciences, National Center for Education Research, National Center for Education Statistics, and National Center for Education Evaluation and Regional Assistance. Of these, the President of the United States said at a press conference:

> Today I have signed into law H.R. 3801, an act to provide for improvement of Federal education research, statistics, evaluation, information, and dissemination, and for other purposes. This Act will substantially strengthen the scientific basis for the Department of Education’s continuing efforts to help families, schools, and State and local governments with the education of America’s children. This Act is an important complement to the No Child Left Behind Act enacted earlier this year. (Bush, 2002, para. 1)

It is statements contained in subsequent documents from the Institute of Education Sciences (IES) regarding the use of scientific research in education, to be described presently, that will help to motivate the argument in this chapter,

S.P. Norris, L.M. Phillips, and J.S. Macnab
University of Alberta

M.C. Shelley II et al. (eds.) *Quality Research in Literacy and Science Education,*

which is that no results of research—no matter how well the research is conceived in science or in other forms of inquiry—can by themselves determine what we ought to do in practice.

We have structured the chapter to enrich and extend other chapters in Part V around five sections. In the first section, we provide some additional context and motivation for the problem we wish to address. Section two is devoted to the question of what constitutes good human action—how we know what to do. Section three examines the nature of scientific theories and research-based knowledge in general and explores what *must* be the case for such knowledge to be put into practice. In the fourth section, we look at a particular case of a scientifically based intervention study and demonstrate how problems of implementation arise that have not been contemplated in the documents provided by the IES. Finally, we turn in the fifth section to some conclusions and policy implications.

27.1 Context and Motivation

Science and scientific knowledge frequently are employed to lever change in social practices, such as medical and nursing care, child care and social welfare, and education. In the medical field, the Cochrane Collaboration has for over a decade been dedicated to an increase in scientific evidence-based practice in medicine:

> The Cochrane Collaboration is an international not-for-profit and independent organization, dedicated to making up-to-date, accurate information about the effects of healthcare readily available worldwide. It produces and disseminates systematic reviews of health care interventions and promotes the search for evidence in the form of clinical trials and other studies of interventions. (Cochrane Collaboration, n.d., para. 1)

A quick scan of the Cochrane website (http://www.cochrane.org) reveals that since 1993 the collaboration has produced over 5,000 meta-analyses of medical intervention studies. There has been an annual international colloquium since its inception. Taken together, the meta-analyses and the colloquia have had a staggering effect on medical practice globally.

Although not as organized as the Cochrane Collaboration, there is a similar move toward science in the nursing field. In many jurisdictions, nursing education has moved away from a hospital-based apprenticeship toward a postsecondary institution-based and scientific knowledge-based profession. Obtaining a nursing license typically requires as a minimum a bachelor’s degree, and nurses are taught more and more by individuals who have research-based doctorate degrees in nursing. Currently in our own jurisdiction, for example, there is a grave shortage of nurses. However, this shortage is preceded and exacerbated by another, namely, the shortage of Ph.D. nursing professors to teach nursing in the university context. The clear aim in the medical fields has been to base practice squarely upon science, on the presumption that science is a way of good knowing and that such knowledge leads to good practice. In these fields, there is a widespread call to turn to scientific evidence to find out what to do.

Education's history is somewhat different. The move in education toward scientifically based research started early in the 20th century and reached its peak around the middle of the second half of the century. At about the same time as scientific educational research was reaching its most dominant status, trenchant criticisms of science as an objective basis for social science research—criticisms that had been articulated much earlier in the century (e.g., Rudner, 1953)—began to take hold. The effect of these criticisms was a move toward qualitative forms of inquiry and away from experimental control and statistical probability as criteria for educational research conclusions. Some believe that the pendulum has begun to swing back toward scientifically based educational research although most researchers, including ourselves, do not wish to embrace the naive forms of empiricism that typified some research during the 20th century. A clear indication of a swing change was the formation of the Campbell Collaboration in 1999/2000. Whereas the Cochrane Collaboration is named after a famous British epidemiologist (Archie Cochrane), the Campbell Collaboration is named after Donald T. Campbell, who had enormous influence on the conduct of experimental inquiry in education and other social sciences (see, e.g., Campbell & Stanley, 1963). The purpose of the Campbell Collaboration is "to help people make well-informed decisions about the effects of interventions in the social, behavioral, and educational arenas" (Campbell Collaboration, n.d., para. 1). It provides research reviews in the areas of crime and justice, education, and social welfare. In contrast to the thousands of reviews in the Cochrane Collaboration, at the date of writing this chapter, only 5 education reviews were completed by the Campbell Collaboration with another 15 in progress.

A further indicator of the recent shift back toward scientifically based educational research is the legislation in the United States mentioned in the introduction. Shortly after its formation, IES published a document, *Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide* (US ED, 2003) that helps frame the argument we make in this chapter (see Shelley, Chap. 22). There are several key features of the document that we will highlight at this point so that they can serve as subsequent foci for critique. The first feature is contained in the title, which indicates that the document is not only about identifying educational practices that are supported by educational research but also about implementing those practices. That is, the document is not only about what we know but also about what to do. We note that, of the total 19 pages in the document, fewer than 1.5 are devoted to implementation. The second feature is the decision to narrow the focus of attention on scientifically based research to "randomized controlled trials—research's 'gold standard' for establishing what works" (p. iii). A third feature is the failure to acknowledge and speak explicitly to how educational goals are adopted and justified. For example, embedded in a statement such as *randomized controlled trials are considered the "gold standard" for evaluating an intervention's effectiveness, in fields such as medicine, welfare and employment policy, and psychology* is a claim or presumption about what counts as effective. No empirical inquiry, no matter how golden, can by itself justify a claim about what counts as effective. Claims about effectiveness and goals are normative and require for their justification normative arguments, although they may depend also upon relevant empirical evidence.

A fourth, and the final, feature we will highlight is the failure to draw the relationship between knowledge that has been generalized and abstracted from several contexts and the use of that knowledge in particularized and concrete settings. Indeed, in attempting to construct such a relationship, the document falls into a contradiction. First, there is the recognition that "slight differences [between implementation settings and settings of the studies] could lead to substantially different outcomes" (US ED, 2003, p. 14). The suggested way to determine whether different outcomes are occurring in, for example, an evidence-based reading program, is to compare the results in the particular context to "a comparison group of schools or classrooms, roughly matched in reading skills and demographic characteristics, that is not using the program" (p. 14). However, such information, in failing to come from a randomized controlled trial (RCT), would not reach the Institute's lowest level of acceptable evidence; hence, by their own arguments, it cannot override the evidence from the RCTs upon which the reading program, by hypothesis, is based.

The reason for falling into contradiction is that the relationship between Gold Standard evidence and use of that evidence in particular contexts has been drawn too tightly by the IES document. The failure is to recognize that obtaining the evidence and putting it into use are two different activities, each with its distinctive processes of reasoning and justification. Since the activities are different, it is not at all inconsistent to point to a practice as having the best evidence for effectiveness and to decide that the practice is not the right one to adopt in the circumstance, or to adopt a practice that has poor evidence to support it. We will show how the relationship between Gold Standard evidence and its use can be drawn less tightly and, thus, without contradiction, leave room for maneuver in deciding what to do.

As a consequence of the critiques we shall make, we will conclude that education needs to be clear how the results of scientific educational research are related to educational practice. We shall conclude also that scholars of education need to help educational policy makers understand how research can be related to practice.

27.2 What Is Good Acting?

We focus in this section on questions of what we ought to do and what sort of justifications is required to answer satisfactorily such questions. The first point is that good acting and good knowing are intimately connected. What we do somehow must be connected to what we know. If we are concerned with what we *ought* to do, then the knowing had better be *good* knowing. We find support for this seemingly obvious claim from Dewey (1929/1984): "*the* problem of practice is what do we need to *know*, how shall we obtain that knowledge and how shall we apply it" (p. 30). Code (1987) also drew a connection between what we know and how we ought to behave: "an epistemic community will be strong in intellectual virtue only if good knowing is valued as a condition of human flourishing" (p. 246). Dewey and Code both were speaking of empirical knowledge of which scientific knowledge is the paradigm case. Thus, conclusions based upon the Gold Standard reasonably can be seen as a basis for good acting.

In addition to good empirical knowledge, what we do must also be based upon sound normative principles. We cannot legitimately infer what ought to be done solely from some fact of the matter. As Dewey (1929–30/1984) again said, "laws and facts, even when they are arrived at in genuinely scientific shape, do not yield *rules of practice*" (p. 14). The is/ought gap to which Dewey was alluding is credited to Hume (1739–40/1962) who noticed that one cannot infer without controversy from an *is* to an *ought*, from a description of how things are to a recommendation how they might be. Upon the descriptive statement, *Smoking causes lung cancer*, by itself, one cannot base any of the recommendations:

> The tax on cigarettes should be high enough to discourage smoking.
> Children should not be permitted to purchase cigarettes.
> People should not smoke.

To see how the direct inference fails, simply consider the additional descriptive claim, *Smoking brings pleasure*. It is now apparent that, in order to infer that children or people generally should not smoke or that the tax on cigarettes should be high, it is necessary to show that the negative consequences of smoking outweigh the positive ones. Showing the latter requires normative judgments that no descriptive statements by themselves can settle.

Without introducing some evaluative or normative premise in the form of a sound principle (people should not engage in behavior that causes serious disease; the state should move to discourage people from causing risk to their health; or laws should be enacted that forbid children from engaging in actions that risk grave harm), we cannot infer from what is the case to what we ought to do. Once we suggest such evaluative premises, however, we see that further problems arise in deciding what should be done based upon what we know. Consider the following line of reasoning:

> Smoking causes lung cancer.
> Laws should be enacted that forbid children from engaging in actions that risk grave harm.
> A law should be enacted to prevent children from smoking.

Now, this line of reasoning gets us what we want, that is, the prevention of children smoking. However, the cost of endorsing the particular evaluative premise might be higher than we are prepared to pay. Clearly, we do not wish to enact laws that forbid children from engaging in *all* actions that risk grave harm to them. Else, we would need laws forbidding hockey playing, bicycle riding, roller blading, rope skipping, plugging in the toaster, and perhaps walking down the stairs. The evaluative premise successfully links the descriptive claim that smoking causes lung cancer to the recommendation that children ought not to smoke, but the evaluative claim is too encompassing. Thus, we must seek a narrower claim that still successfully

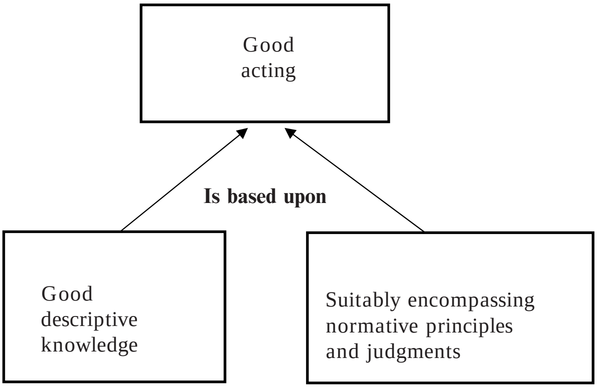

Good acting
Is based upon
Good descriptive knowledge
Suitably encompassing normative principles and judgments

Fig. 27.1 The basis of good acting in good descriptive knowledge and in normative principles and judgments

makes the link between the effects of smoking and our desire to prevent children from doing it but does not so restrict children's lives as to make them unbearable. The evaluative premise might be modified, for instance, to call for laws that forbid risky actions when those actions have no weighty positive outcomes. Such a premise would not rule out bicycle riding; because, even though it carries risks, it also provides enormous benefits in fun, exercise, ease of transportation, and skill and agility development that can last a lifetime.

We thus conclude that good acting requires both good descriptive knowledge—much of which comes from scientific research—and sound evaluative principles, which arise from our imaginations and very broad ethical concepts such as fairness and justice. We have depicted this relationship in Fig. 27.1. The dual basis of good acting is depicted by the dual arrows coming from descriptive knowledge and from normative principles and judgment, both of which serve as part of the basis. To act upon descriptive knowledge alone is a failure in critical reflection, because no amount of knowledge alone can imply what ought to be done. On the other hand, we must be careful of the evaluative premises we choose. If we adopt an evaluative premise to link knowledge to action in one situation, then consistency demands that we apply that premise to all cases that fall under it. If we are not careful in our choice of premises, we can rule out action that we desire highly. Thus, critical reflection is also needed in choosing evaluative premises.

27.3 Nature of Scientific Theories and Research-based Knowledge

In addition to the considerations in the previous section on the link between knowledge and action, the nature of scientific theories also bears upon how they can be put to use. We shall use the term *scientific theory* in our discussion, which seems acceptable given that the Gold Standard is a norm derived from science. Nevertheless, we do not limit our conclusions to science proper. Indeed, the characteristics of scientific theories that we highlight also are characteristics of research-based knowledge generally, including educational research-based knowledge.

27.3.1 The Semantic Conception of Theories

We derive our view of theories from Suppe (1977, 1989), and a fuller description of our derivation can be found elsewhere (Norris, 2000; Norris & Kvernbekk, 1997). According to Suppe, theories are models or abstract systems. Abstract systems are abstract in the following sense: theories are concerned with phenomena only insofar as the phenomena are characterizable by a small number of parameters abstracted from them. A theory cannot characterize a phenomenon in all of its complexity. Abstraction in this sense must occur if theories are to be general.

Given the attention to science paid by those advocating the Gold Standard, it seems appropriate to examine an example from the natural sciences. Consider the Kinetic Theory of Gases. In addition to the parameters of pressure, volume, and temperature, the theory is based on abstracted parameters for the size of molecules, the shape of molecules, the motion of molecules, the density of molecules, the elasticity of the collisions among molecules and between molecules and the walls of the container, and the attractive and repulsive forces through a distance among molecules and between molecules and the container. Typically, as is the case here, theories contain idealizations on some parameters. We might, for example, idealize an interaction to be negligible or zero, as we do if we assume perfectly elastic collisions among molecules. A state of an abstract system is defined by the values on each of its parameters at a given time, and the behavior of an abstract system is its changes in state over time. Changes in state are defined by laws of the theory.

The nature of the relationship between abstract systems and concrete systems is one of *counterfactuality*. Theories do not describe accurately concrete phenomena but describe what the phenomena *would have* been *had* the selected parameters been the only ones exerting any influence and *had* the idealizations been real. The nature of the relationship between theories and concrete systems leads to a number of implications, which we explore in the following three sections.

27.3.2 Impossibility of Direct Application

The abstraction and idealization necessary for the existence of theories must be taken into account when applying them. There never can be a recipe to get from an abstract and general theory to a concrete and particular system. The connection has to be indirect through auxiliary hypotheses that specify the influence on the concrete system of factors not specified in the theory. Auxiliary hypotheses are required for *all* applications of abstract systems to concrete systems. Auxiliary hypotheses specify the effects of outside influences that are identified based upon an “appraisal of the situation” (Norris, 2000, p. 181). The possible ways of applying a theory are in principle unlimited, and the possible number of auxiliary hypotheses useful in mediating between a theory and concrete systems is in principle unlimited. For many cases of application, we do not have all of the requisite auxiliary hypotheses, especially not when the concrete system is characterized by great variability and flux. This can be difficult work—deciding how theoretical knowledge needs mediating for use in particular situations.

Theory application always involves normative considerations about whether and how to apply the theory. These normative considerations might involve questions of economics, aesthetics, ethics, and prudence. Such considerations are not part of the theory but may affect making connections between the theory and concrete systems. When applying a theory, there must be some more or less clear notion of what to try to achieve. For example, if the desired accuracy of prediction is low, it perhaps would be sufficient to take into account through auxiliary hypotheses only some of the most important influences that lie outside the scope of the theory. If the desired accuracy is high, then more influences might need to be taken into account, and taken into account more precisely. This final point leads to a discussion of the variability of choice in application.

27.3.3 Variability in Application

Situational appraisals cannot be made in the abstract. Rather, they must be made in light of the particular theory being applied, the particularities of the situation, and the outcomes desired. Situational appraisals are *judgments* that can be made only by those knowledgeable of the application situation. Theoretical and practical knowledge have equal importance in the application situation.

Let us consider a situation headlined as follows in a recent newspaper article: “Should your daughter get the needle?” (Anderssen & Alphonso, 2007). The article was about the question of whether or not girls 12–14 years of age ought to be given a new vaccine that has been shown to confer immunity against cervical cancer caused by the human papillomavirus (HPV), a sexually transmitted disease (STD). Here are some of the medical facts:

- There are about 150 types of HPV; most clear the body but about 40 can linger and cause various types of cancer, which, in women, primarily is of the cervix.
- Randomized controlled clinical trials have shown that a vaccine affords nearly 100% immunity to infection by four of the most common types of HPV, which cause 70% of all cervical cancer—a disease that kills on average more than one Canadian woman per day and leaves survivors infertile—and 90% of genital warts.
- The trials were conducted on 16–23-year-old females.

- The tests show the immunity to last at least 5 years, but it will take decades to learn how long the immunity lasts.
- The vaccine works only if administered before exposure to the viruses.
- In 2006 in Canada: about 1,350 women were diagnosed with cervical cancer and 390 died, more than 22,000 were diagnosed with breast cancer, and the number who died from cervical cancer is a small fraction of those who died from lung and ovarian cancer.

So, the medical research is very clear. The degree of immunity afforded by the vaccine is outstanding. Should the course of action be obvious? Based upon medical facts, certain recommendations and actions have been taken:

- Medical experts advise giving the vaccine before girls are sexually active.
- 12–14-year-old girls were offered the drug at public expense and with parental consent in four Canadian provinces in the fall of 2007.

What normative considerations led the medical experts to their recommendation? On what normative basis did the four provinces make their decisions? On the one hand, it might be argued that the trials were so positive that it would be unethical not to make the vaccine available. On the other hand, one might wonder why and how different provinces with the same facts reached different conclusions.

These considerations lead us to the question of how the research can be applied in specific cases. What ought particular parents to do: provide or withhold their permission? On what basis should they decide? We will argue that there can be great variability of application, many bases for decision, and that the same theory or knowledge can be applied legitimately in different ways according to the context and the situational appraisal. Considerations of the child's size and physical and emotional maturity, the history of cervical cancer in the child's lineage, and the existence of developmental disorders might all come into play. Consider the following sketches of arguments by a number of parents reported in the newspaper article.

Parent 1: Anything that will protect my daughter from cancer is worth the risk. (Spoken as justification for giving permission for inoculation)

Parent 2: I have fear of side effects, question the motives of the drug company, and feel queasy about dosing girls as young as 10 with protection against an STD. (Spoken as justification for declining permission for inoculation)

Parent 3: At this age, kids are pretty innocent and this is not something they should have to worry about. (Spoken as justification for declining permission)

Parent 4: If a doctor said I can provide a vaccine against cancer of the lung, I wouldn't think twice about it. (Spoken as justification for giving permission)

Parent 5: It's not like vaccinating your kid against polio. There is no epidemic of cervical cancer. (Spoken as justification for declining permission)

Parent 6: I'll wait and see whether more is known in a couple of years' time. (Spoken as justification for declining permission)

Parent 7: The vaccine may promote early sexual behavior or unsafe sex or a belief that it is ok to be sexually active. (Spoken as justification for declining permission)

Parent 8: We can't trust the medical community to know what is best for our children. (Spoken as justification for declining permission)

All of these parents knew that the research evidence by itself was insufficient as a guide to action. Other normative considerations—some political, some religious, some ethical, some prudential, some pragmatic—had to be brought into play before a decision could be made on whether to give or decline permission for their daughters to be vaccinated. Not only is there room for judgment, it is demanded, because the medical evidence by itself does not imply which action to take. Even though different decisions were reached, all of the parents were using the same research-based knowledge. Moreover, we find it rhetorically striking that, even though the medical evidence was overwhelmingly positive about the effectiveness of the vaccine, 75% of the small sample of parents withheld their permission.

In education, the empirical evidence on effectiveness of treatments is never so clear. We must be cautious, however, not to view the link between knowledge and action as weaker than it is. Although theories and knowledge do not prescribe precisely, they can constrain action. Some actions are not in accord with the medical research on the vaccine against cervical cancer. For example, a program of vaccination in senior secondary school would not make sense because of the increased chance of exposure to the viruses through intercourse.

27.3.4 Role of Values and Choice in Application

We saw in the previous section that application involves normative considerations about whether and how to apply a given theory. Ultimately, we must have a clear conception of what we to try to achieve when we are applying knowledge because what is desired can alter the auxiliary hypotheses needed for application. For example, if we want only a rough prediction or want to make only a small modification in the world, then we might take into account through auxiliary hypotheses only some of the more important influences that lie outside the scope of the theory. If a parent wanted above all else to reduce the risk of cervical cancer in a daughter to a minimum (perhaps because of a personal traumatic experience with the cancer), then the parent might not be satisfied with the low risk that exists even without vaccination and opt for the inoculation in order to achieve the lowest risk possible.

If, however, a parent has other beliefs, such as that drugs are inherently dangerous and that the risks of catching the disease (known to be low) are not clearly higher than the unknown risks from drug side effects, the parent could easily justify foregoing the vaccination. We wish to make clear that we do not advocate an anything-goes policy. Take, for instance, the parents who based a decision to permit their daughter's vaccination on the grounds that anything that protects her from cancer is worth the risk. It is unlikely the parents actually believe this justification. For example, one way to protect a child from skin cancer due to sun exposure is never to permit the child to go outdoors. All foods, even organic ones, expose the body to some carcinogenic substances. A way to avoid exposure is not to eat, which is a ridiculous course of action. So, the parent does not mean *anything* that protects against cancer is worth the risk because some things that protect against cancer impose an even more severe risk of other, even more undesirable consequences. All action requires a trade-off between competing values. It is often difficult in such trade-offs to see one value trumping all others.

27.3.5 Summary

We have attempted to portray the main points of this section in Fig. 27.2. First, note that theory is related to some phenomena through the relationship of explanation—the theory explains the phenomena. The relationship between the theory and some application situation might also be one of explanation; but it might also be one of prediction, description, intervention, or perhaps other possibilities. We have used the overarching expression *applies to* to capture all these possibilities. The figure shows that the application emanates from the theory with two supplements indicated by the addition symbols, first, auxiliary hypotheses about the workings of the application situation and, second, normative considerations for and against the application. The point is to show that application is not a direct line from theory to the application situation.

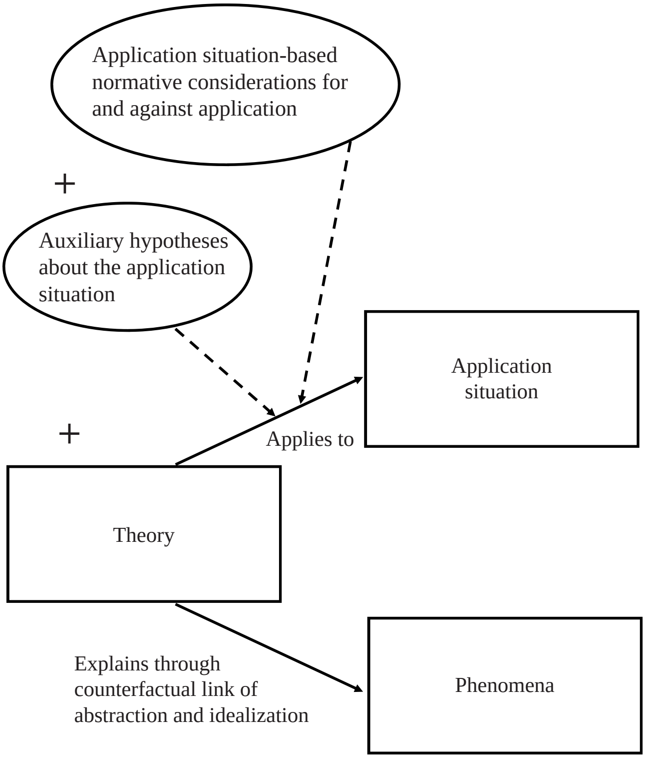

Application situation-based normative considerations for and against application
+
Auxiliary hypotheses about the application situation
Application situation
+
Applies to
Theory
Explains through counterfactual link of abstraction and idealization
Phenomena

Fig. 27.2 The relationship among theory, phenomena, and application situations

27.4 An Intervention Study

Now that we have outlined our theoretical machinery, we turn to a longitudinal, early literacy intervention study that two of us helped to conduct (Phillips, Norris, & Mason, 1996). We introduce this example because it matches to a large degree the type of study for establishing what works that falls under the Gold Standard and because we have full information on the study, including access to the raw data. There were three treatment groups—a school-only treatment, a home-only treatment, and a home and school treatment—and a control group. Treatment children were given extra instruction in early literacy concepts using a series of *Little Books* (McCormick & Mason, 1990) in Kindergarten and the effects were followed until the end of Grade 4. Positive effects were strongest at the end of Grade 2 and in this order: school treatment, home and school treatment, home treatment.

In the school-only treatment, children and teachers read *Little Books* in school in addition to the approved language arts program. In the home-only treatment, children and their parents read the same *Little Books* at home, and the children received the approved language arts program in school. In the home and school treatment, *Little Books* were read in both settings in addition to the approved language arts program. In the control group, *Little Books* were not used at all and the children received the approved language arts program.

The study randomized on classroom and analyzed data by students, thus not keeping constant the unit of analysis. Covariance analysis was used to remove the effects of preexisting differences among the groups. As such, it was not a strictly Gold Standard study but met well the criteria of “an intervention backed by ‘possible’ evidence of effectiveness” (US ED, 2003, p. 11). However, every point made in the subsequent discussion would apply even if the study had met strictly the Gold Standard.

Figure 27.3 presents a scatter plot of the relationship for the school-only treatment children between their pretest scores at the beginning of Kindergarten (Metropolitan Reading Readiness Test [MET 1], Nurss & McGauvran, 1987) and their posttest scores at the end of Grade 2 (National Achievement Test [NAT II], Wick, Fraenkel, Mason, Stewart, & Wallen, 1989). Each point represents a single child’s pair of scores. Scores are in standard deviation units so that scores of zero are average for both measures. The diagonal line represents children whose relative standing on both measures is the same, that is, their scores are above or below the mean by the same number of standard deviation units on each test. To the right and below the diagonal, children had a higher relative standing on the pretest than they did on the posttest. Pick any of those points, and you will see that the child represented by that point has a higher standard deviation score on MET I than on NAT II. To the left and above the diagonal, the relative standing of children was higher on the posttest than on the pretest.

The intersection of the vertical and horizontal lines is the centroid for the control group, that is, the average scores in standard deviation units that the control group children received on both measures. You can see immediately that on average the

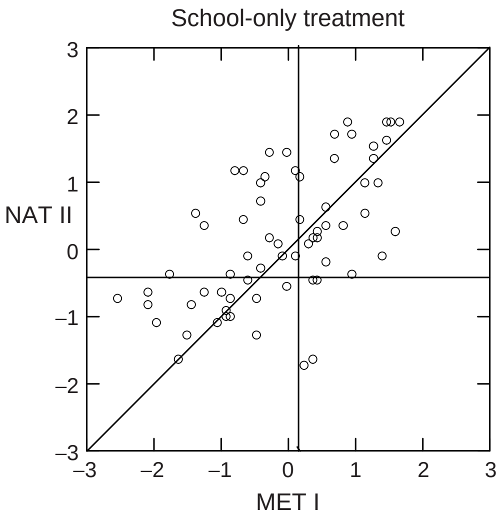


Fig. 27.3 Scatter plot for the school-only treatment children between their pretest scores at the beginning of kindergarten (MET I) and their posttest scores at the end of second grade (NAT II)

control group children performed relatively better on the pretest than they did on the posttest because the centroid is to the right and below the diagonal. Students whose scores fell to the right of the vertical line did better than the control group on the pretest; to the left, they did worse than the control group on the pretest. Children above the horizontal line did better than the control group on the posttest; below the line, they did worse than the control group. We also can combine the information from looking at those falling left and right of the vertical line and those falling above and below the horizontal line: children in the lower-left quadrant did worse on both measures than the control group children; those in the upper-right did better on both measures than the control group; children in the lower-right lost ground compared to the control group because they were above the average of the control on the pretest measure but below the average on the posttest; in the upper-left quadrant, children gained ground because they scored lower than the average for the control children on the pretest but higher than average on the posttest.

Look now to Fig. 27.4, which provides the same information for the home and school treatment children. The data are distributed differently. In particular examining the lower-right and upper-left quadrants, you can see that, compared to the school-only treatment, a greater proportion of the home and school treatment children lost ground with respect to the control group (0.10 versus 0.06) and a smaller proportion gained ground (0.14 versus 0.25). So, the home and school treatment did not work as well as the school-only treatment.

Examine Fig. 27.5 for the home-only treatment children in which the contrasts to Fig. 27.3 are even starker than for Fig. 27.4. Compared to the home and school

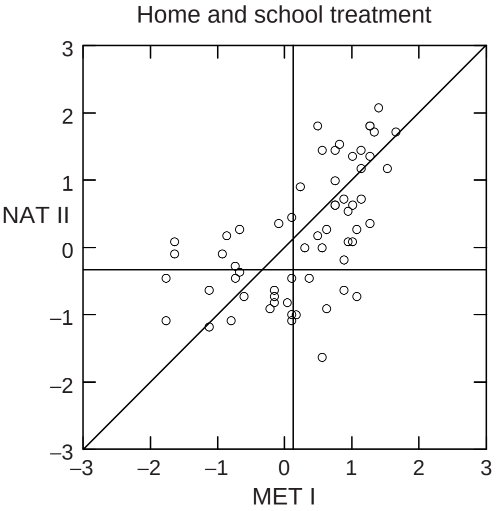


Fig. 27.4 Scatter plot for the home and school treatment children between their pretest scores at the beginning of kindergarten (MET I) and their posttest scores at the end of second grade (NAT II)

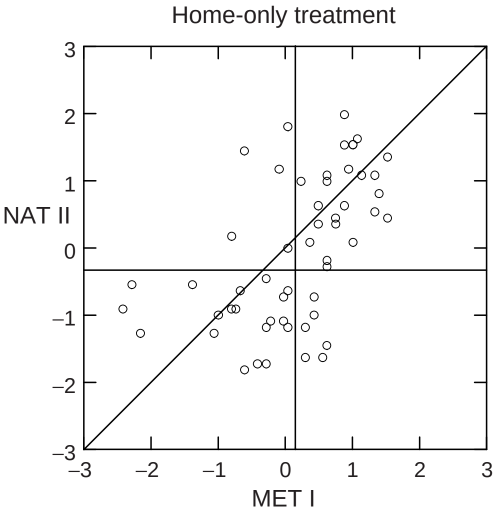


Fig. 27.5 Scatter plot for the home-only treatment children between their pretest scores at the beginning of kindergarten (MET I) and their posttest scores at the end of second grade (NAT II)

treatment, an even greater proportion lost ground with respect to the control group (0.11) and an even lower proportion gained ground (0.08).

We wish now to use this example to motivate a general analysis of the desired conclusion from an intervention study. In general, we wish to infer from a claim of the form '*a* caused *b*' (i.e., what happened in a particular case) to a claim of the form ‘*A*s cause *B*s’ (i.e., what happens in general) or from ‘*a* did not cause *b*’ to ‘*A*s do not cause *B*s’. The first type of claim, specific causal claims, includes past tense singular claims about what has happened. The second type of claim, general causal claims, contains tense-less general claims about standing states or conditions. There is often an implied *usually*, *generally*, *frequently* as part of claims in this latter category. The distinction is very similar to that drawn by Campbell and Stanley (1963, p. 5) between internal validity (“Did in fact the experimental treatments make a difference in this specific instance?”) and external validity (“To what populations, settings, treatment variables, and measurement variables can this effect be generalized?”).

Although it is conceivable, perhaps even likely, that some results may be particular to individual experiments, the point of randomized experimental design is lost if one cannot expect some level of generality from the research. We are always interested in moving from ‘*this intervention* caused an effect of size ε in *this sample*’ to ‘*interventions of this type* cause effects of size ε in *samples from this population*.’ This is the type of information we wish to gather from Gold Standard research.

We have chosen to use explicitly causal language, rather than Campbell and Stanley’s “make a difference” (1963, p. 5). Sometimes there is an attempt to avoid the imputation of causation on the grounds that it implies mechanical or deterministic systems. Most such attempts are unsuccessful (see, e.g., Ennis, 1982). We say unsuccessful because causation is implied by such language as *brought about, led to, succeeded in creating*, *made a difference*. We believe also that the concept of intervention contains a causal implication. Interventions are actions we take with specific intentions to alter the course of events from what they would have been otherwise. Nevertheless, when we look at what happens to individual students, some might experience an effect equal to ε, some experience an effect larger than ε, some experience an effect smaller than ε, some experience a negative effect, and some experience no effect at all. Results like this typically occur. The Gold Standard is not about what happens to individuals but about what happens to the group on average.

So, even with Gold Standard evidence, there is still a decision about what to do on the basis of it: Are the gains by those who gain worth the losses by those who lose? Is there an intervention with more acceptable trade-offs? Is the monetary cost worth the gains that are found? Therefore, based upon this study, what should schools do? First of all, it is not immediately obvious that they should adopt the *Little Books* intervention. There is an effect size of about 2 standard deviations needed to bring the children targeted by the intervention up to the mean of their peer group. The *Little Books* interventions produced an effect equal to about 0.25 of a standard deviation. Not all children profited from the intervention, and some even fell behind. Unfortunately, such is typically the case even with interventions that pass the Gold Standard of effectiveness. In education, there is rarely one approach that works for everyone. Perhaps the schools would like to wait to find an intervention that works for more children. Perhaps they would like to try a combination of interventions. Of course, combining interventions can lead to complications because a positive effect that an intervention has when used in isolation may not be sustained when used in combination with something else. Basically, there is so much left to decide, even when the evidence is in and it points to effectiveness!

27.5 Conclusions and Policy Implications

We hope to have shown that evidence does not determine action, in the sense of leaving open one and only one possible way to proceed. The evidence alone—even from Gold Standard research—cannot tell us what to do. Even with clear conclusions, much is left to individuals with situation-specific knowledge to decide how research is best applied in their contexts. This is the same point that we made earlier working from a theoretical perspective on the nature of theories and research-based knowledge.

Similar points are argued in a very insightful set of chapters in a book edited by Thomas and Pring (2004). For example, Cordingley (2004) made the important and often overlooked point that, even in the context of full evidence (if such can be imagined), there will always be a role for professional judgment in deciding how the evidence is best applied to particular settings and particular students. Eraut (2004) made a similar point to Cordingley's in the context of medicine, namely, that the idea of evidence-based practice seems to presuppose the incorrect view that somehow evidence can determine what ought to be done in practice. Eraut argued that, in addition to research-based evidence, practitioners need to draw upon knowledge derived from their own experience, which he calls practice-based evidence. The main point of Hodkinson and Smith's (2004) chapter is that there is no such thing as safe research, research that points with perfect reliability to a course of action. Above all, they claimed, the relationship between research and practice is imbued with an uneliminable political element.

An important caution arising from our analysis is that policy makers need to be fully aware of the politics involved in the use of educational research. We believe they need to know and grasp the significance of at least the following: that, in using the results of scientific educational research to guide practice, even evidence based upon Gold Standard research cannot by itself determine decisions about what to do; that the use of scientific results involves a mediation between abstract and general scientific knowledge and concrete and specific situational knowledge; that they and other educational practitioners are the mediators; and that mediations are rarely clear-cut, because the same knowledge can be applied in different ways in different contexts and at different times in the same context.

Perhaps education programs, particularly those aimed at educational administrators at the graduate level, could focus upon some key abilities needed to use research-based educational knowledge. The ability to formulate reasonable auxiliary hypotheses to mediate between theories and concrete educational situations is unlikely to be something that comes naturally, even to individuals who realize that such hypotheses are needed. Likewise, the ability to employ normative considerations in conjunction with the best scientific and situational knowledge is not something that currently is given much attention by faculties of education.

How is it best to teach these abilities? At least, we conjecture, through plenty of practice with examples that: demonstrate the pitfalls that arise when it is assumed that scientific knowledge is directly applicable; demonstrate the variety of auxiliary hypotheses and normative principles that must be brought to bear for successful application; encourage the explicit formulation of auxiliary hypotheses and normative principles through situational appraisals; encourage the conjecture, consideration, and evaluation of alternative application routes for a given theory in a given context that depend upon different desired outcomes of the application; and encourage the evaluation of whether applications are consistent or not with the theory being applied.

If we wish scientific educational research to serve the public good by providing part of the basis for many of our educational practices, then scholars of education have a role in showing policy makers how they can use scientific educational research results in their practice and in providing policy makers the opportunity to acquire the knowledge, skills, and dispositions to use science wisely. The IES could do well by including an extensive elaboration in their documentation of the role of evidence in implementing changes in educational practice. Gold Standard research is next to impossible to conduct in authentic educational settings and, where it is possible, provides no absolute guidance on what to do. Finally, this last conclusion is not meant to imply that Gold Standard research is not important. Quite the contrary—it is important to have the most robust evidence possible for making educational policy. The conclusion is meant to reiterate that evidence based upon Gold Standard research just does not have the degree of authority that many advocates proclaim it has.

References

Anderssen, E., & Alphonso, C. (2007). Should your daughter get the needle? *The Globe and Mail, Alberta edition,* A1, A16–17.

Bush, G. W. (2002, November 5). *Statement on signing legislation to provide for improvement of federal education research, statistics, evaluation, information, and dissemination, and for other purposes*. Retrieved February 14, 2008, from http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname = 2002_presidential_documents&docid = pd11no02_txt-21.pdf

Campbell Collaboration. (n.d.). *Education coordinating group Homepage*. Retrieved June 28, 2008, from http://www.campbellcollaboration.org/ECG/index.asp

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), *Handbook of research in science teaching* (pp. 171–246). Chicago: Rand McNally.

Cochrane Collaboration. (n.d.). *Homepage*. Retrieved October 3, 2007, from http://cochrane.org/docs/descrip.htm

Code, L. (1987). *Epistemic responsibility*. Hanover, NH: Brown University Press.

Cordingley, P. (2004). Teachers using evidence: Using what we know about teaching and learning to reconceptualise evidence-based practice. In G. Thomas & R. Pring (Eds.), *Evidence-based practice in education* (pp. 77–87). Maidenhead, Berkshire, UK: Open University Press.

Dewey, J. (1929/1984). The quest for certainty: A study of the relation of knowledge and action. In J. A. Boydston (Ed.), *John Dewey, The Later Works, 1925–1953* (Vol. 4: 1929). Carbondale, IL: Southern Illinois University Press. (Original work published 1929)

Dewey, J. (1929–30/1984). The sources of a science of education. In J. A. Boydston & K. E. Poulos (Eds.), *John Dewey, The Later Works, 1925–1953* (Vol. 5: 1929–1930). Carbondale, IL: Southern Illinois University Press. (Original work published 1929–30)

Ennis, R. H. (1982). Research notes: Abandon causality? *Educational Researcher*, *11*(7), 25–27.

Eraut, M. (2004). Practice-based evidence. In G. Thomas & R. Pring (Eds.), *Evidence-based practice in education* (pp. 91–101). Maidenhead, Berkshire, UK: Open University Press.

Education Sciences Reform Act of 2002. Pub. L. No. 107–279, 116 Stat. 1940. (2002).

Hodkinson, P., & Smith, J. K. (2004). The relationship between research, policy and practice. In G. Thomas & R. Pring (Eds.), *Evidence-based practice in education* (pp. 150–163). Maidenhead, Berkshire, UK: Open University Press.

Hume, D. (1739–40/1962). *A treatise of human nature: Being an attempt to introduce the experimental method of reasoning into moral subjects*. (Edited with an introduction by D. G. C. Macnabb). Cleveland, OH: World Publishing. (Original work published 1739–40)

McCormick, C. E., & Mason, J. M. (1990). *Little books*. Glenview, IL: Scott, Foresman.

No Child Left Behind Act of 2001. Pub. L. No. 107–110, 115 Stat. 1425. (2002).

Norris, S. P. (2000). The pale of consideration when seeking sources of teaching expertise. *American Journal of Education*, *108*(3), 167–195.

Norris, S. P., & Kvernbekk, T. (1997). The application of science education theories. *Journal of Research in Science Teaching*, *34*(10), 977–1005.

Nurss, J. R., & McGauvran, M. E. (1987). *Metropolitan reading readiness test*. San Antonio, TX: The Psychological Corporation.

Phillips, L. M., Norris, S. P., & Mason, J. M. (1996). Longitudinal effects of early literacy concepts on reading achievement: A kindergarten intervention and five-year follow-up. *Journal of Literacy Research*, *28*(1), 173–195.

Rudner, R. (1953). The scientist *qua* scientist makes value judgments. *Philosophy of Science*, *20*(1), 1–6.

Suppe, F. (1977). *The structure of scientific theories* (2nd edn.). Urbana, IL: University of Illinois Press.

Suppe, F. (1989). *The semantic conception of theories and scientific realism*. Urbana, IL: University of Illinois Press.

Thomas, G. (2004). Introduction: Evidence and practice. In G. Thomas & R. Pring (Eds.), *Evidence-based practice in education* (pp. 1–18). Maidenhead, Berkshire, UK: Open University Press.

Thomas, G., & Pring, R. (Eds.). (2004). *Evidence-based practice in education*. Maidenhead, Berkshire, UK: Open University Press.

United States Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Washington, DC: Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Available from http://ies.ed.gov/ncee/pubs/evidence_based/evidence_based.asp

Wick, J. W., Fraenkel, J. R., Mason, J. M., Stewart, J., & Wallen, N. E. (1989). *National achievement test*. Chicago: American Testronics.

Part VI
Epilogue: New Standards, New Directions, and New Realities

Chapter 28
Reflections on Beyond the Gold Standards Era and Ways of Promoting Compelling Arguments about Science Literacy for All

Larry D. Yore, Mack C. Shelley II, and Brian Hand

This book flowed from the deliberations of the 2nd Island Conference (held at Dunsmuir Lodge in Victoria, British Columbia, Canada) and several American Statistical Association and National Association for Research in Science Teaching symposia in the United States that examined the ramifications of the Gold Standard for educational research found in government legislation and promoted by the US Department of Education Institute of Education Sciences. Planning and development of the book expanded the authors and contributions beyond the 2nd Island Conference participants so as to sample international perspectives more broadly related to literacy and science education research quality assurance, practices, and other issues. Some of these issues were anticipated and addressed by the authors while other issues were not anticipated and emerged from the collective insights of several authors as the book evolved.

28.1 The 2nd Island Conference

The 2nd Island Conference (held in 2005), *Gold Standard(s) of Quality Research in Science and Literacy Education*, followed the 1st Island Conference (held in 2002) that explored the interdisciplinary foundations and research of science literacy (Hand, Yore, & Prain, 2006). The 2nd Island Conference shared the collective concerns of the literacy and science education research communities composed of established and emerging researchers and graduate students from Asia, Australia, Europe, New Zealand, and North America during the Gold Standard Era; it addressed moving

L.D. Yore
University of Victoria

M.C. Shelley II
Iowa State University

B. Hand
University of Iowa

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

the agenda forward into a post-Gold Standard Era while being aware of the unproductive history of the *research wars*. Within the supportive, open, but diverse environment of the Vancouver Island retreat, learning, literacy, statistics, measurement, public policy, and science education researchers with the common desire to improve research quality, enhance the public trust in education research, and inform public policy engaged a variety of issues. The honest and sometimes tense deliberations addressed the *one-size-fits-all* standard, alternative quality assurance approaches, and the alignment of research approach, logics, problem space, research questions, and interpretative frames leading to evidence-based decisions about literacy and science education instruction. The wide-ranging discussions and presentations produced a fuller understanding of the central problems and international perspectives on literacy and science education research issues.

Education is part of the socioeconomic and sociocultural system of most nations and is essential to the survival and growth of democracies. Education, therefore, is a legitimate strategic component used in governmental heuristics to achieve societal priorities, social justice, and cultural goals. These ideas are explicit in the US legislation regarding education accountability, research quality, and evidence-based decisions and are implicit in other countries' reports of commissions and inquiries as well as mission-driven funding to address specified, pressing, economic, and social issues. We have tried to take an apolitical stance, and we do not pretend to judge these political agendas—but we recognize their existence as fundamental factors in setting and influencing literacy and science education research policies, funding priorities, and practices.

The design of this book moves from the conference proposal and deliberations to issues of pedagogy, theory, and innovative techniques, to policy making and decision making. In Part II, Setting the Agenda: Science Education and Science-based Research, the authors (a) address this process in an era of political pressures, changing budget priorities, great potentials outlined by numerous reforms and task force reports, sincere need for reconsideration of positions, and advocacy for specific research methods in terms of public mandates internationally and (b) recognize that it is no simple matter. The insightful resolution of these issues will bring common good to the public and academic communities and provide a strong foundation from which to lobby and inform just policies and decisions about education and literacy and science education research. Hayward and Phillips (see Chap. 7) question the appropriateness of the wholesale importation and adoption of an evidence-based practice (EBP) model and accompanying evidence hierarchies to educational practice. If EBP is applied without common sense, it may lead to disregarding the essential studies critical to teaching and learning in literacy and science education. They propose that policy makers and funding agencies revisit the implementation of EBP and address barriers to the uptake of research evidence in educational practice.

In Part III, Curriculum and Pedagogy, the authors address essential needs for contemporary inquiries and programs of study into literacy and science education that are informed by quality research and progressive research agendas. Some promising research approaches and agendas are provided to illustrate productive lines of research that go beyond the randomized controlled trials (RCTs) approach promoted by the Gold Standard to explore the complexity of the interaction and match between language as a learning tool and new technologies with mixed methods. Saul and Hand (see Chap. 12) point out the need for researchers to be more concerned about the question(s) they are exploring than about framing their inquiries on a particular method. Given that the border-crossing or convergence between science and literacy education means that old and new questions arise from the histories and integration of traditions, researchers need to engage the concepts and the methods framed by the research questions. Thus, methods should be seen as located on a continuum where various combinations can be used to address any question. A second critical element of this research is the need for researchers to examine the areas of study such that inquiries are better aligned with the realities of the classroom. This is both a methods issue and a relevance issue; large-effect sizes obtained for a particular treatment in a select context may be too difficult and demanding to apply in a normal classroom or to implement in a school program. An example of such research is the current emerging of multimodal representation where researchers are either clustered around developing *magic-bullet* representations that will greatly assist student learning or around how we can best use learning theory to assist students build their own representations. Each approach will have different outcomes and different qualities of research design; as a consequence, each will have to be examined carefully about their particular relevance for classrooms.

In Part IV, Statistics, Research Methods, and Science Literacy, the authors outline approaches that press the Gold Standard envelope of design, logic, and analysis married to the problem space, research questions, available instrumentation, and research approaches. Clearly, these approaches stress rigorous planning, implementation, and interpretation, and designs appropriate to the development of the problem space, research questions, and established knowledge. Mundfrom (see Chap. 21) addresses the fundamental issues of when do quality considerations start, as he considered the question “Can We Make a Silk Purse from a Sow’s Ear?” He points out the disconnect of educational policy formulated independently from educational research about effective instructional practice, which may be due partly to the political nature of education and its policies and partly to the questionable design and analyses employed in the research related to underlying assumptions, experimental units and units of analysis, randomization, statistical choices, confounding variables, overreliance on software and statistical packages, and valid use of test scores.

In Part V, Public Policy and “Gold Standard(s)” Research, the authors focus specifically on the ways in which global funding patterns reflect governmental priorities, potential hidden preferences, and essential evaluation criteria as well as the role of research ethics boards, rules and expectations for maintaining the security of datasets, synthesis of qualitative research studies to provide broader impacts on public policy making, and what should be done with good research findings to ensure good actions. Norris, Phillips, and Macnab (see Chap. 27) outline the links between evidence and actions and the need for evaluative and normative premises of these links. They point out that informed, fair, and just actions require evidence flowing from quality research and shared values.

In this final single-chapter Part VI, Epilogue: New Standards, New Directions, and New Realities, we address a collection of contemporary issues dealing with ideas flowing from the 2nd Island Conference, conference symposia, individual chapters, and task force reports on learning and improving research. We have attempted to integrate and highlight several issues that emerged from the earlier parts and chapters.

28.2 Summary Comments and Elaborations

We take our lead from the previous parts to address the most pressing issues and those not fully addressed by other authors. These eight anticipated and emerging issues will be addressed somewhat here but are highlighted mainly for emphasis and future consideration:

- Appropriate designs and methods.
- Rigor and research agenda.
- Funding criteria and duration.
- Politics of knowledge and reward systems.
- Academic associations, journal editors, editorial boards, and review panels.
- Policy-making process and policy makers.
- Information communication technologies (ICTs) and modern analysis systems.
- Contract research, commercial research organizations, and researchers-for-hire.

28.2.1 Appropriate Designs and Methods

Issues of quality start with clear understanding of the problem space and its development; this applies to the current status of science literacy (Fensham, 2008; Yore, Pimm, & Tuan, 2007). In the intersection of literacy and science education research, this involves the acceptance that *science literacy for all* targets all learners and encompasses the cognitive symbiosis of literacy in science (cognitive and metacognitive abilities, critical thinking, habits of mind, language, and ICT strategies) and understanding of science (unifying conceptual themes, nature of science, scientific inquiry, etc.), sociocultural considerations, and the global goal to enhance peoples' engagement in the public debate about science, technology, society, and environment issues to reach informed decisions and sustainable actions. These identified and yet-to-be identified components and relationships are at various stages of development. No single research design and method (Gold Standard) can or should be applied to all research questions in the science-literacy-for-all problem space. Therefore, assessing and clarifying the problem space is followed by a dynamic, recursive, evolving process of finding worthwhile, causal questions (including sociocultural issues, educational aims, etc.), posing plausible answers or tentative hypotheses, and developing appropriate research designs and methods so as to investigate the resulting questions and test the hypotheses (Brickhouse, 2006; Lawson, 2005; Simon, 2004).

Clearly, *Gold, Platinum, or Diamond—Good Standards* should be applied to guide these diverse inquiries (experimental, nonexperimental, mixed-methods, philosophical, historical, or other approaches) and to produce high-quality results about worthwhile questions (Brickhouse, 2006; Phillips, 2006). But rigorous enactment of the resulting inquiry and collecting and interpreting data (e.g., observations, measurements, interviews, authentic discourse, constant comparison, correlations, inferential statistics, regression analysis, ANOVA, MANOVA, structural equation modeling [SEM], hierarchical linear modeling [HLM], data mining, etc.) do not always lead to publishable results; such decisions are not based solely on unique or significant findings but rather on the researchers' integrity, personal standards, and self-assessment of what was learned from the study (Simon, 2004; Yore, 2003). Journals need to publish high-quality research reports from various types of inquiries—both significant and nonsignificant results that support and do not support current dogmas—to ensure that the available results fully reflect the landscape of findings.

Several funding agencies recognize the importance of the alignment of design, problem space, questions, established knowledge base, and available instrumentation. They implicitly accept quality standards and anticipate proposals utilizing a variety of research approaches to different questions and even different approaches to the same question. The US National Science Foundation (NSF) and other agencies recognize that science education inquiries build theoretical understandings and develop practical innovations that reflect a cyclic agenda: study (basic research); design (development); implement (contextual applications); evaluate (scaled uses); and synthesize (establish insights, new questions, and future agenda). Furthermore, the calls for proposals specify that appropriate designs and methods will vary across specific phases of the research and development cycle. Even the US Department of Education allows for and funds non-RCT studies, which can include qualitative research with RCT-equivalent rigor, that focus on early explorations in a research program. Please note that some authors and researchers interchange RCTs (random controlled trials, random clinical trials) and RFTs (random field trials), but their interpretations require either random selection of participants or random assignment of treatments, or both.

The one-size-fits-all Gold Standard for literacy and science education research needs to be expanded into guidelines that recognize and stress the reality that the problem space and research questions drive decisions about which specific research designs and research method to employ. The National Research Council (US NRC, 2002) provided guiding principles for high-quality scientific research, which emphasized: (a) posing significant questions and using empirical designs developed on relevant theoretical and methodological understandings; (b) linking individual inquiries to an overarching, theory-driven, conceptual framework; (c) selecting appropriate methods for exploring specific questions; (d) providing a coherent and explicit chain of reasoning and compelling arguments that address limitations, biases, counterclaims, and alternative explanations; (e) reporting results that have been replicated and generalized beyond the narrow settings and populations; and (f) submitting research findings open to evaluation, scrutiny, and critique of wider professional communities.

None of these principles disallows quality research or individual inquiries using experimental, nonexperimental, and mixed-methods approaches. Furthermore, these principles are best achieved and utilized when applied to a program of inquiry (research agenda) that moves from preliminary, exploratory, small studies to experimental, controlled studies to large-scale, random, controlled trials. It is just such research agendas that provide rich descriptions, generate informed hypotheses that reflect reality, and establish that "the causal efficacy of programs or interventions is the main, or most important, purpose of educational research" (Phillips, 2006, p. 22). Citing the cause is not enough; quality research connects the argument's warrants and explanations to the established theoretical backings and new theory developed and provides cause–effect mechanisms for the connections among theories, models, and results. Phillips continued:

> Certainly the RFT, based as it is on J.S. Mill's principles of logic, is an excellent way to establish that X causes Y, and it can be used with profit in many educational research and program evaluation studies. But it is not the *only* way to establish a causal relation, and it is not the *necessary* way. And it is important to remember that establishing X causes Y is not the same thing as establishing *why it does so* (that is, establishing the physical or social mechanism), and it is this latter issue that is often of vital interest in science and in the public policy arena; it also is salutary to remember that the RFT is of little or no value in answering this deeper question about causal mechanisms. (p. 22)

Experimental, nonexperimental, and mixed-methods inquiries do not have a monopoly on developing well-documented and supported claims and explanations. Historical and political approaches "are capable of producing well-crafted works that present vital and sometimes mind-expanding insights about education that are well supported by arguments and warranting considerations that can withstand critical scrutiny" (p. 21).

28.2.2 Rigor and Research Agenda

In the events leading to this book, two general concerns arose about research rigor and the lack of consideration of the program of study with an evolving research agenda to reflect the ever-changing problem space, established knowledge base, and available instrumentation. Rigor is more than procedural consistency and mechanical applications (Brickhouse, 2006). Munby (2003) stated, "research is a human enterprise and … its worth is more than simply its trustworthiness" (p. 153). He continued to consider the fundamental psychometrics in quantitative inquiries and the parallel constructs in qualitative inquiries to focus on rigor of the argument and claims, rigor and ethics, rigor and professionalism, and rigor, persuasion, and rhetoric. Both Brickhouse and Munby believed rigor involved ongoing decisions; but it starts with the belief that important questions were being considered and the results would be worthwhile—if there was a sound rationale for the research focus, design, and methodology. Authors in this book stress that rigor involves clear understanding of underlying statistical assumptions, logics, ethics, plausible reasoning, epistemological beliefs, ontological assumptions, evaluative and normative premises, and personal values subsumed in a research approach. Judgments about quality and actions involve more than evidence—evaluative and normative premises require critical attention on what to believe and what to do.

One of the most surprising outcomes experienced over the development of this book was the number of researchers unwilling or unable to assess their research rigor or to submit their research to open and fulsome public evaluation, scrutiny, and critique. Lawson (2007) evaluated research articles appearing in a leading science education research publication—the *Journal of Research in Science Teaching* (1965, 1975, 1985, 1995, 2005)—using a three-level epistemological framework proposed in the same journal by Smith and Wenk (2006): Level 1 represents inductive inquiries, Level 2 represents hypothetico-deductive inquiries, and Level 3 represents theory-driven deductive inquiries. A search of the word-files revealed that 18.0% (1965), 36.5% (1975), 44.3% (1985), 58.3% (1995), and 86.7% (2005), respectively, of the articles contained the word *theory* associated with Level 3 inquiries. A careful reading of a random sample of the articles from 2005 "revealed that most authors were generating and testing hypotheses and/or theories (presumably guided by Level 2 or Level 3 epistemology) albeit in a largely implicit and sometimes haphazard way" (Lawson, p. 1). Lawson appeared generous in his epistemic classification of these articles since many authors were using scientific metalanguage loosely; for example, the terms *prediction*, *hypothesis*, and *theory* were used as any speculation rather than as a deduction based on an established relationship, as a tentative statement of a relationship between variables, and as an umbrella idea that integrates subordinate ideas and provides explanatory power (cause–effect relationship and mechanism). Discussion of these findings with researchers attending the NARST symposium revealed a disappointing lack of awareness of epistemic level regarding research designs and inquiries and an even greater reluctance to engage in public debate about rigor and inferential power of research, which is characteristic of scientific communities. Although application of this epistemological frame to a single study may be uninformative, its application to a program of studies for several years to a specific problem space may be enlightening.

Many funding agencies ascribe large weightings to the applicants' academic record and performance, with less weighting on the quality of the planned inquiry in the evaluation of proposals. However, funders do not pay much attention to the developmental trajectory of the studies and evolution of the research agenda or program of study. Instead, it appears as if they count articles, assess level of journals, and imply impact measured by citations. Such approaches suggest a belief that past performance—rather than measuring actual impact on policy, professional practice, and learning—can predict quality. This lack of consideration of the research agenda and the counting of articles do not detect cookie-cutter repetitions of research designs and studies that do not reflect growth and insight; rather, it encourages isolated inquiries, multiple publications from limited data, and repeated studies using the same method and research questions uninformed by previous studies and unreflective of the ever-changing knowledge.

The concept of rigor is about argument (Munby, 2003); that is, can the researcher(s) provide a strong argument that moves beyond simple methodological issues to provide a clear and defensible link between their question, theoretical foundations, claims, and evidence. We would argue that success in the persuasion involved with argument requires researchers to establish a program of research wherein the concept of the argument is built over time and across a series of studies. This then enables the quality of the research to be established across time rather than be based on a single point in time.

28.2.3 Funding Criteria and Duration

The evaluation processes and criteria reported in the quality assurance and funding chapters were similar across several countries and funding agencies (see Coll et al., Chap. 6; She et al., Chap. 23). Quality assurance processes were based on institutions' scholarship, academic productivity, and research grant success of applicants from the faculty ranks of the institution. Applicants' records of productivity, established expertise, significance of the proposed study, research approach, resources and time provided by host institutions, and potential national/regional/state/local benefits are the most common factors mentioned in the research proposal evaluations and granting process. The weightings of these factors vary across funding agencies and their different funding programs. Curiosity-driven calls for research proposals stress the expertise, productivity, and commitment of the applicants, the importance of the proposed research, and the rationale for the proposed approach (see She et al.). Mission-driven calls for research proposals add the match of the proposal to the specific goals and targets of the funding envelope. Henig (2008) suggested the funding and evaluation process for education research is vague, murky, and not influenced by stakeholder needs and pending policies. Funding agencies, program officers, and review panels frequently try to share the available funds across the largest number of proposals and applicants that fully meet the criteria. Brewer and Goldhaber (2008) stated:

> [A]dequate funding is crucial to conduct research on a scale large enough and with sufficiently rigorous design to make it useful. Large-scale efficacy studies, for instance, require millions of dollars over a sustained period. Yet the federal government's research effort [in the United States] is split among many agencies (and divided into even smaller pots of money within each agency), and each state acts largely on its own. While there are strong pressures to divvy up funds among many small projects and powerful constituencies, the exponentially greater value of consolidated efforts must be highlighted. (p. 364)

Curiosity-driven grants tend to be smaller, of shorter duration, and not large enough to support the evolution of a program of study or to address and remediate systemic problems. Mission-driven grants tend to be larger and of longer duration, but they seriously underestimate the cost and effort of scaling and capacity-building projects.

Brewer and Goldhaber (2008) believed the quantity of education research is healthy; however: “Unfortunately, the bulk of educational research neither is outcomes-oriented nor uses methods that support strong inferences about causality” (p. 361). They continued:

> [Because academics frequently work in departmental silos,] there is often relatively little interaction and collaboration across these boundaries. This has important implications: it reinforces the production of work that reflects a single disciplinary view of the world; it minimizes the sharing of methods and new developments, and it limits institutionwide scrutiny of the quality of research. (p. 363)

The 1st and 2nd Island Conferences and the existence of this book demonstrate that the conference participants and authors believe it takes interdisciplinary communities and diverse perspectives to make a difference and achieve science literacy for all.

28.2.4 Politics of Knowledge and Reward Systems

The politics of knowledge involving university reward systems that superficially assess research impact—rather than actual impact, practical applications, and influence on public policy for grant evaluations and personnel and salary decisions—appear to encourage short-term inquiries, case studies, and action research that do not lead to generalized knowledge claims and explanations. These naturalistic methods can provide rich insights into practice and context and lead to a more informed, acute understanding of the problem space, worthwhile research questions, and plausible hypotheses; but, in isolation, they normally do not build theory or influence policy. Researchers frequently identify the lack of funding and time as barriers to doing research, especially large-scale RCTs.

> However, other than at a handful of the most highly ranked research institutions [in the United States], funding is not required for promotion and tenure. All of this tends to bias work in favor of less-costly qualitative case studies, which focus on process issues rather than on the efficacy of practices, programs, and policies and which do not yield generablizable findings. (Brewer & Goldhaber, 2008, p. 363)

Henig (2008) stated:

> [E]ducational research has a reputation of being amateurish, unscientific, and generally beside the point. Exacerbating matters are high-profile tussles between prominent researchers publicly disparaging one another’s methods and interpretations. … But the portrayal of the debates in the public arena reinforces cynicism. (p. 357)

The issue of quality is not a binary, qualitative-or-quantitative proposition; rather, it is the serious desire to conduct disciplined inquiries about important problems and researchable questions to make a difference for students, teachers, and other stakeholders. Moving your research agenda to approaches that allow generalized knowledge claims and explanations with associated cause–effect mechanisms is the ultimate goal of quality programs. However, making better use of naturalistic studies, small-scale experiments, and large-scale survey databases could help address some of these problems. Therefore, it is critical that literacy and science education researchers explore and develop data-sharing procedures and new secondary analysis and synthesis techniques to afford more complete descriptions of the problem space, produce generalized knowledge claims, and influence public policy and educational decisions about science literacy.

28.2.5 Academic Associations, Journal Editors, Editorial Boards, and Review Panels

High-quality research is a community responsibility. Supporters, producers, critics, and publishers of literacy and science education research need to stand firm on the requirements of empirical inquiries and of qualitative and quantitative evidence-based claims about important problems and researchable questions. Researchers should not be left alone in the struggle to produce and report high-quality research with clear and compelling arguments. The powerbrokers and gatekeepers behind literacy and science education research and publications (e.g., universities, academic associations, professional organizations, publishing companies, editors, editorial boards, and reviewers) must do their part to facilitate and ensure high-quality research. Universities, associations, and knowledge-based businesses reap significant benefits from the efforts of researchers and the high-quality research reports flowing from or found under their agency. Therefore, these organizations have shared responsibilities in developing supportive research cultures and building research capacities. Universities must provide the research culture and services that support ethical, high-quality inquiries and build capacity with high-quality, research-oriented graduate programs and also recruit, retain, and nurture new researchers. Academic associations in literacy and science education need to expand the exemplar efforts to grow the human talent pool and enhance the abilities and opportunities for new researchers in nonresearch-oriented universities and from developing countries with research summer schools, mentorships, networks, research internships, and postdoctoral fellowships.

Associations, publishers, and journals must ensure that research reports provided under their imprints contain quality arguments and have structured abstracts, sufficient details, and data access to allow readers and end-users to evaluate the veracity of the arguments (US NRC, 2004). Quality arguments start with quality evidence, plausible reasoning, defensible claims, and value-added explanations and implications (Yore, 2003). Reports need to provide clear, explicit, and transparent warrants of data and information as evidence for claims and assertions; clues and reasoning chains must be required in manuscripts submitted for review and apparent in articles published (Zientek, Capraro, & Capraro, 2008).

Associations should develop, publish, and promote guidelines or standards of research ethics and professional conduct that clearly illustrate appropriate recruitment of participants, informed and voluntary participation, respect for minority perspective and intellectual properties, storage and secondary uses of data, coauthorship, and graduate student supervision. Journal editors should require proof of ethics approval for all manuscripts submitted involving human subjects and private information. The document *Standards for Reporting Empirical Social Science Research in AERA Publications* (American Educational Research Association [AERA], 2006) provides such a set of fundamental principles and expectations that could serve as a starting point for literacy and science education associations without such standards. Zientek and colleagues (2008) stated, “the specifications on reporting standards are useful for researchers as they prepare manuscripts, for editors and reviewers as they review manuscripts, and readers as they attempt to build their practice or store of knowledge on the basis of the published works” (p. 208).

These “standards are not intended to define the conduct of empirical research. Although research reporting and research conduct are necessarily related, decisions about how to conduct empirical research are the researcher’s responsibility” (AERA, 2006, p. 1). However, if low-quality research results are not published, it is more likely that researchers will be more vigilant in their decisions about problem spaces, research questions, designs, data collections and interpretations, and arguments and also be more aware of negative outcomes when seeking to report low-quality research studies. Organizations could form consortia to develop (a) quality standards for research and reporting that members would have access to and (b) support services and professional development to move research toward acceptable standards, which would improve the status of literacy and science education research in the eyes of end users.

Submission instructions should (a) state that proof of institutional review board approval of the research ethics is required, (b) specify the form-function for parts of a research report, and (c) ensure that clear, transparent, compelling arguments are central to all reports; reviewers should ensure that authors adhere to these requirements. Effective research reporting requires that writers present worthwhile, valid research results (knowledge claims) in a clear and compelling fashion that considers the audience, patterns of argument, and explicit and implicit language rules of the discourse community.

Journal editors, editorial boards, and reviewers—the gatekeepers of the community—must judge the quality of the research by considering the match amongst the research questions, design, and results and the impact of the argument. Editors’ and reviewers’ recommendations to accept, reject, or revise and resubmit must be supported with equally rigorous and detailed arguments plus specific suggestions for revisions to meet the journal’s quality standards. The collective effects of the specific suggestions serve both as evidence to justify the recommendation and to guide authors’ revisions. Editors and reviewers should make separate judgments about the quality of the writing and the quality of the research since revise–resubmit recommendations must be based on research quality—to do otherwise may mislead authors into believing the underlying research is worthy of publication in that journal. This can be a demanding task because writing issues are difficult to disentangle from the judgments about research quality as the message and medium are intertwined.

Reports of high-quality research should clearly reveal the scientific report genre (organization, form, and function), such as (Yore & Yore, 2007):

- *Structured abstract*: The abstract needs to provide a concise and interesting overview of the problem (and its importance), study, results, and implications.
- *Introduction*: The introduction must set the general context of the problem, its importance, and the essential architecture of the background section to follow.
- *Background*: The background must weave an integrated theoretical framework that justifies the research focus and research approach, establishes the data interpretation framework, and provides the backings to warrant the resulting claims and to rebut counterclaims, leaving no surprises to the readers in the later parts of the manuscript.
- *Design*: The authors need to provide an overview of the research approach used and of the critical assumptions, characteristics, and procedures of this approach so as to inform the readers and enable replications of the study.
- *Results*: Claims and assertions should be clearly identified and stated, followed by the supportive evidence and warrants used to justify the claim or assertion. Claims and assertions should contain an appropriate degree of hedging to convey clearly the authors' level of certainty about their results. Authors need to be parsimonious while being convincing; it is difficult to determine when enough evidence is sufficient and not redundant. Statistical studies should contain the descriptive statistics fundamental to the more complex statistical tests, treatments, and modeling; but space requirements must be considered in terms of the number of tables and data displays provided. Writers need to provide quotes from their informants to justify their assertion, but more is not always better—be strategic about selecting each quote to ensure it provides the readers with breadth and depth, convincing evidence for the claim, and adds value.
- *Discussion and implications*: The discussion needs to clarify and elaborate the justification for the claims and assertions by sharing the authors' thinking and decision-making processes about claims and counterclaims. The discussion should not just restate ideas made earlier but should enrich the readers' understanding of the research reported and provide explanations of the results, where possible with related cause–effect mechanisms. The contemporary importance of the study can be achieved by considering applications of the findings, its implications for policy, and future research possibilities.
- *References*: The listing of references provided at the end of a manuscript contains all the references used to make and justify the authors' argument and does not contain tangential readings or mention the big names in the field.

28.2.6 Policy-making Process and Policy Makers

A consistent motive expressed by participants in the 2nd Island Conference and the authors of this book has been to *make a difference*. Brickhouse (2006) reflected the other participants' views when she pointed out that education researchers need to be afforded academic freedom to pursue curiosity-driven inquiries, but they must be held to "a higher ethical standard [in] that research should have at least some potential to improve the quality of education and the lives of children" (p. 4). Relying on the work of Norris and colleagues (see Chap. 27), she noted that research results do not provide sufficient grounding for changing practice or setting policy and encouraged educational researchers to become aware of the policy-making process and the evaluative (evidence) and normative (values) premises involved in policy decisions.

There appears to be a consistency in the view that the influence of public policy is very complex and that research results are more likely to be used to confirm a position or justify a favorite program than to inform or change a policy maker's beliefs or positions. Rees (2008) stated:

> Unfortunately, politics is among those domains of human activity least beholden to sound academic research. First, politics—indeed, social relations of all kinds—is about power, ambition, social status, and personal prestige. Thus, while politicians will readily adopt research that supports their beliefs, many show little affinity for results that challenge their political survival. … Second, politics is ideological and, like other mythic constructs, a political ideology can be a rather ungainly concoction of fact and values, assumptions and illusions. It often gains credence only after frequent repetition and ritualistic affirmation. (p. 10)

There is very little known about how policy makers use research to inform their actions, beliefs, and proposals; in fact, "policy action is often propelled more by myth than science" (p. 10). Hess (2008) suggested that impatience, the desire for rapid and dramatic changes, and increased polarization "have made it less likely that research—even when it is rigorous and reliable—will influence policy" (p. 354).

The influence of research has had variable impact across the social sciences and professional communities and across researchers. Hess (2008) stated:

> While researchers in both health care and education pursue advances with enormous personal stakes for individuals and for society, the health profession has won enough credibility that a substantial reservoir of support for basic research has developed, even though the benefits may not be visible for decades. However, lacking a similar history of successes, educational research has not earned similar trust or good will. (p. 256)

Researchers and the quality and usefulness of their research and counsel can earn trust and goodwill, but they need to monitor and assess quality and strategic applications to the target policy area consistently. Furthermore, they need to be persistent in their efforts and realistic in their expectations when attempting to influence policy and policy makers. Much damage has been done with short-term and sporadic attempts to influence policy makers with ill-informed advice and when idealistic expectations are promoted. Success relies on well-developed, well-documented, achievable claims, dissemination to the appropriate end users, and continuous support.

28.2.6.1 Development, Dissemination, and Direct Services

Henig (2008) identified ICT, the increased number of low-quality education journals, contract research and privatization, and research funding agencies' priorities and resources as emerging issues in quality assurance and impact on public policy.

Fusarelli (2008) pointed out that school leaders do not use traditional educational research reported in academic journals, but they do use action research results and data-driven decision making. These leaders find the lack of time, lack of expertise, cultural conflicts, lack of relevance, and communications between researchers and end users as barriers to using traditional research results. They need easy access, efficient forms, and trustworthy research reports to improve usage.

Local school boards have much to say about the interpretation and implementation of policies that are not totally prescriptive. Henig (2008) said that these:

> school governance bodies tend to have more bureaucratic insulation, more concrete and pragmatic needs for data and analysis, and less ideological polarization. … Broad structural changes have made the national political environment [in the United States] a highly polarized one, with policy debates and partisan strategies shaped more by ideological purists than by those seeking to find common ground based on the public's generally moderate center of gravity. (p. 360)

He suggested that an analysis of these hothouse politics, in which education policy and funding can be a flash point affecting every politician, has produced two practical results:

> First, researchers have heard the message that they should descend from their ivory towers and engage the world. Second, the old model of 'speaking truth to power' in which the scholar as favored advisor whispers into the ear of elite leaders, also is passé; in the age of mass media and the Internet, discourse about research has been democratized. (p. 360)

Many of the traditions and conventions of the academy can serve "as buffers against ideology and the politicization of the knowledge enterprise. These factors also play a role in maintaining a distinction between research and advocacy, between pursuit of knowledge and pursuit of advantage, between sounding good and being right" (p. 360).

Some people would use these insights to bash governments and political persuasion along the spectrum from ultra-conservative to radical-liberal. Such is not the intention of this book or this chapter; rather, it is our intention to provide insights into a more comprehensive view of literacy and science education research and alternative resolutions to quality and utility of research results. Cohn (2006) suggested that academics occupy a special place in society and "have a duty as citizens to use their knowledge for the public good" (p. 8) by taking an active role in the policy deliberations and development process. Academics are afforded a position of trust and privilege; with this affordance comes responsibility and the belief that "scholars have a moral obligation to use their knowledge to advocate for policy that serves the public good [and] … that advocacy is in itself a continual part of the scholar's responsibility to society" (Cohn, 2007, p. 18). Scholars have obtained their academic freedom to explore curiosity-driven research agendas "by being disengaged from the socio-economic and political power needed to implement the ideas" (2006, p. 9). However, it is difficult to advocate the same isolation and detachment from the real world for members of professional faculties—like literacy and science education professors.

Cohn (2006) argued that "academics have ample and frequent opportunities to influence public policy but that the influence available to them is usually indirect and secured by convincing those with power to advocate for and/or act on their ideas" (p. 10). He suggested that a key bridge between the ultimate decision makers (the so-called first community in knowledge utilization literature—politicians, high-ranking appointees, etc.) and academics (the so-called second community) is third community actors (policy advisers, consultants, research officers, support staff, lobbyists, special interest groups, advocacies, etc.). This third community overlaps in some cases considerably with academia and is highly pervasive within both the public and private sectors. These actors use knowledge and information to produce analyses that are useful to and in the language of decision makers and then disseminate these analyses to influence or advise decision makers. Academics who become involved with the research staffs of government ministries, support staff, and counsel to cabinet committees have improved chances of influencing public policy. However, the usefulness of the academics' advice is reflected in the third community's ability to translate these ideas in the target context (constituents, current political climate, nature of their problem, political priorities, etc.) and the window of opportunity for the pending policy. These contextual and temporal considerations are not prime factors in designing and conducting academic research; therefore, some researchers may be ill-equipped to incorporate them into their thinking and operations.

Literacy and science education needs advocates at the local, state/provincial, national/federal, and international levels—like booster clubs for a sports team or music group—to keep science literacy for all in the public's attention and to promote positive actions. Advocacy involves many things, including building a community of support for values and ideas and the persistent and informed participation in the public debate and support of science literacy. Cohn (2007) cautioned academics that without sound understanding of the policy-making process:

> advice can do more harm than good, especially if it involves highly complex and interrelated set of prescriptions. … If a broad-based community of support for the values and ideas that the scholar wishes to advocate are lacking, … government will simply pick and choose among the recommendations, according to ideological disposition and political needs, with little care for the holistic model developed by the scholar/policy advisor. (p. 18)

Academics set on influencing public policy need to increase the scope of their specific research questions, step back and survey the contextual landscape of the problem space to provide sound evidence-based advice that is applicable to the context, and intake a breath of pragmatics and practical reality. Cohn (2006) stated:

> [The] shape and size of [windows of opportunity] are said to go a long way toward determining the character of the policies that are produced. … It might take so long to find *the* answer that its proponents miss the proverbial boat, with the policy window narrowing appreciably during the course of the highly rational research. (p. 13)

Furthermore, researchers need to speak to different audiences (decision makers, bureaucrats, end users) and provide persuasive arguments, recruit or engage advocacy groups (teacher associations, parent groups, think tanks, etc.), and avoid conflict to inform policy making—not set policy.

> In order for the relationship between scholars and policy-makers to succeed, policy-makers (whether policy advisers or decision-makers) need to know what can reasonably be expected of academic researchers, and academic researchers need to know what can reasonably be expected of policy-makers. (p. 25)

Public advocacy is not without problems and negative outcomes when advice or research findings are only partially adopted or public policy is only partially implemented. Academics need to anticipate partial implementation of recommendations in an ever-changing political climate of power, players, and priorities. Cohn (2006) addressed the risk–benefit of policy involvement and stated:

> The risk facing individual scholars is that their efforts could be construed as too oppositional, jeopardizing their status as reliable sources of information and their relationship with public servants. … In terms of academia in general, the risks are far more substantial. Scholars in industrial democracies such as Canada are already under pressure to produce work that is policy-relevant (in other words, that assists governments in their work) and that is driven by commerce rather than by curiosity. (p. 26)

Furthermore, the frequency of mission-driven funding is becoming much more common in the policy-driven climate. Academic integrity and freedom tend to encourage university professors to speak in a forthright manner rather than the more diplomatic approach used by bureaucrats and civil servants.

28.2.6.2 Knowledge Utilization Process

Knowledge utilization, or knowledge transfer, is used frequently to justify research funding in a grant proposal but receives low priority and effort in the actual enactment of funded proposals. Knowledge utilization involves intense, dynamic, recursive, prolonged interactions among researchers, policy advisers, and policy makers (Landry, Lamari, & Amara, 2003). Lavis, Robertson, Woodside, McLeod, and Abelson (2003) identified five questions for research organizations regarding knowledge transfer to decision makers in medicine and health care: what is the actionable take-away message, who is the target audience, who should be the messenger, how should the message be communicated, and what expectations are there for effective transfer of the message. However, they concluded that many Canadian applied health and economic/social research organizations were not organized or staffed to emphasize and achieve effective knowledge transfer. Lavis and colleagues suggested that the most effective organizations:

> (1) … [developed] messages for … target audiences that [moved beyond specific] research reports or specified possible action; (2) [custom designed] their knowledge-transfer approach to their … target audiences and … dedicated [time and] resources to getting to know their target audiences and … discussing research reports and ideas transcending particular research reports; (3) dedicated [staff] to enhancing their internal capacity for knowledge transfer; (4) engaged their target audiences in … the research process; and (5) made [effective] use of Web sites and newsletters. (p. 240)

There are several steps in the knowledge utilization process that need to be anticipated by academics if they are to successfully negotiate the process and then inform and hopefully influence public policy (Knott & Wildavsky, 1980):

1. Reception involves policy makers receiving relevant research.
2. Cognition involves policy makers reading and understanding the academic research.
3. Discussion involves policy makers engaging in activities to discuss findings.
4. Reference involves policy makers citing the research and findings in their own work.
5. Adoption involves policy makers who advocate the adoption of the reported findings into official policy.
6. Influence involves the research report and findings influencing the policy makers' unit.

These steps represent a funnel with many ideas entered into the reception stage (step 1) but few actually influencing official actions (step 6). A single publication or research report may not fully persuade the various audiences and address the various stages in policy making.

28.2.6.3 Communicating with Public Policy

Persuasion means using appropriate language, stressing cooperation and collaboration rather than conflict, and recognizing extended arguments—evidence, claims, counterclaims, and rebuttals. Academics need to provide informative policy briefs in *plain talk* that clearly describes the knowledge claim, underlying premises, and evidence as well as engaging and rebutting alternative claims.

> [T]hird community activities can lead to a more proactive role for academics if targeted at a wider audience, including those knowledge brokers and leaders … who possess the social and economic power that academics lack. Actions targeting those who shape public opinion, and the public itself, as well as those who shape the policy positions of corporations, associations, interest groups, and political parties, go beyond policy-making into the realm of politics. (Cohn, 2006, p. 18)

Speaking and writing to power requires consideration of language form and function inherent in the purpose of the communications. Advocacy organizations, special interest groups, and lobbyists need to provide information in a genre that is efficient and effective. The Society for Research in Child Development (Society for Research in Child Development, n.d.) provides research-based briefs on social policy topics concerning children and families—Head Start and No Child Left Behind (Ludwig & Phillips, 2007; Porter & Polikoff, 2007), improving early mathematics education (Ginsburg, Lee, & Boyd, 2008), and many other issues. The two-page format, style, and message are concise and informative but not overwhelming as is the case in most research reports. Unfortunately, this form of communications is unique to academics and not highly valued in the academy (Yore, Hand, & Florence, 2004).

28.2.6.4 Royal Inquiries, Task Forces, and Commissions

Royal inquiries, task forces, and commissions are instrumental in policy making and laying the foundation for policies in many countries but are not without political difficulties. Membership in these groups may be based on expertise, representation, or other criteria; but once formed they all involve negotiation, persuasion, controversy, and compromise. Those involved in the language and literacy (International Reading Association & National Council of Teachers of English, 1996) and science education (American Association for the Advancement of Science, 1990; US NRC, 1996) reforms in the United States will attest to these internal and external struggles in producing a document from diverse input and lengthy deliberations. The reports of task forces, inquiries, and commissions are called white papers in some countries, which serve as the foundation and source of guidance for many government decisions, while green papers are usually discussion documents that may lead to white papers. The National Science Council of Taiwan bases many of its funding policies on the 2002 White Paper flowing from a high-level and highly regarded Ministry of Education meeting on science education. The European Union (EU) bases much of its research funding decisions on the Green Paper flowing from the Lisbon Summit on socioeconomic growth. Several recent reports or papers have potential for influencing literacy and science education around the world.

Rocard, Csermely, Jorde, Lenzen, Walberg-Henriksson, and Hemmo (2007) reported on the deliberations of a high-level, science education group charged by the European Commission to provide policy and action recommendations for the EU based on expert opinions, knowledge of science education research, analysis of promising projects, and interviews with ministers responsible for science education and academics leading major science education projects. The report is written in plain language with comforting hints of the research foundation and a few specific references designed not to overwhelm the reader—while encouraging thoughtful engagement and further inquiry into the six recommendations dealing with (a) priority, importance, and changes needed to science teaching; (b) the gender gap; (c) participation in science-oriented careers; (d) teacher education; (e) the availability of extracurricular science and technology resources; and (f) exemplar programs.

Fensham (2008) was charged by UNESCO to survey and analyze science education curricula and instructional practices and to establish recommendations for future directions. The report speaks to various international audiences in developed and developing countries in plain, accessible language providing brief backgrounds, recommendations, and prospects and prerequisites for issues emerging from his study, experience, and expertise. He outlined 11 policy recommendations for science education: (a) clearer goals for science and technology education; (b) more students under the guidance of able science teachers, reduction in the implicit and explicit risk factors for females, and consideration of culturally diverse students' language, beliefs, and values; (c) personal and social relevance of topics and activities; (d) context-based approach to learning and teaching; (e) balance of science as established information and nature of science; (f) science literacy as abilities and knowledge; (g) assessment focused on higher-level learning;

(h) ICT across the school system; (i) authentic assessment techniques; (j) primary and elementary programs that emphasize positive and creative encounters with natural and people-built environments; and (k) high-quality, ongoing, focused professional development for teachers of science.

Osborne and Dillon (2008) reported on two seminars funded by the Nuffield Foundation involving leading European science educators to reflect critically on the status of science education in Europe. They reported that over the last two decades science has been accepted as a compulsory course in schools but little has been done to reform the curricula and instructional approaches to achieve science for all students. They stated, “Our view is that a science education *for all* can only be justified if it offers something of universal value for *all* rather than the *minority* who will become future scientists” (p. 7). Therefore, curriculum and instruction needs to reflect contemporary issues and how science works regarding political and moral dilemmas, risk and uncertainty, economic benefits and values, and strengths and limits. They outlined seven recommendations for science education in Europe flowing from the deliberations: (a) educate students about major explanations of the material world and the ways science works; (b) curriculum innovations and organization of teaching needs to address less-motivated students; (c) improved access to human and physical resources about the range of scientific and technological careers; (d) high-quality teachers for primary and lower secondary schools focused on engaging students with science and scientific phenomena; (e) long-term projects of sustained engaged learning; (f) enhanced foundations for assessment of the competencies expected of scientific literacy; and (g) recruitment, retention, and professional development of high-quality teachers.

The National Literacy Panel on Language-Minority Children and Youth (composed of distinguished scholars from Canada and the United States) utilized meta-analysis, secondary analysis, and systematic interpretation of quantitative and qualitative research results to address the development of literacy amongst learners whose home language was not the language of the majority and instruction (August & Shanahan, 2006b). This project attempted “to identify, assess and synthesize research on the education of language-minority children and youth with respect to their attainment of literacy” (August & Shanahan, 2006a, p. 1). The resulting report and searchable database were notable because they needed to be published outside the normal publication process, shared original information sources, and illustrated some results that did not support current federal and state policies regarding English language learners. The findings identified the need to develop precursor oral and print skills, the importance of home language proficiency and individual attributes, and surprising outcomes involving assessment practices, teacher judgments, and sociocultural influences.

Service on these inquiries, task forces, and commissions requires commitment without expectation of personal reward—other than doing the right thing for learners, teachers, and other stakeholders. History and personal experience reveal that such efforts and reports have varying degrees of impact and uptake, but each contribution normally adds to the collective wisdom and incrementally moves literacy and science education toward science literacy for all. Unfortunately, many universities do not recognize these service contributions sufficiently well to justify the time and energy commitments.

28.2.7 Information Communication Technologies and Modern Analysis Systems

ICT and computer software have been mentioned as positive and negative factors in literacy and science education research and policy development. ICT has changed the gatekeepers of public information and the dissemination process of knowledge. Henig (2008) pointed out that many think tanks and advocacy groups disseminate their reports, studies, and summaries by electronic means. They use a complex network of their Web sites linked to other sites and blogs that provides their message in various forms and from multiple perspectives. The ease of posting information without outside monitoring is the best of worlds and the worst of worlds. He believes researchers are under pressure to get results posted before being scooped by someone else: “When speed becomes critical, normal processes for refining, checking, and simply deliberating about evidence can be short-circuited” (p. 358). Clearly, without peer-review, end users must take a critical stance when interpreting research and assessing quality. However, this also is true for many media sources used to inform opinions and influence actions. “Researchers may acknowledge the limitations of their own data and design, but those caveats are often the first things to be stripped from the message as others take it up” (p. 358) or are reported in the popular media. ICT provides a powerful lobbying strategy that can be directed to various stakeholders and agents in the policy-making process and has significant potential to influence the political process. The ease of reformatting information for different target audiences and the interactive nature of ICT mean that the message and process are much more personalized and can have greater impact.

Overreliance on statistical software without fully understanding the underlying assumptions of the application has led to inappropriate applications, unreasonable certainties, and questionable interpretations of data and results. A common example is the often too-easy use of pull-down statistical software menus to apply data analytic methods (such as least-squares-based multiple linear regression) to a categorical dependent variable from a survey or from institutional records when the more appropriate approach would be multivariate logistic regression or discriminant analysis. Other than nonparametric procedures, most major statistical methods are based on assumptions about normal distributions, constant variance, and independence of observations from random samples that are at best questionable and usually quite unlikely.

Promising and innovative applications involve secondary analysis, secondary reanalysis, data mining, data security, and related ethical and rigor considerations involved in secondary uses of data and data sharing. The desire to find patterns in databases and accumulated results from similar research on a problem space brings new demands for awareness of underlying assumptions, limitations, and procedures.

There is growing popularity in discourse, conversation, and performance analysis without equal consideration of the theoretical and procedural demands. Discourse analysis permits researchers to compare prepared and student-generated texts in terms of propositional structure and function. Conversation analysis allows for temporal and sequential dissection of oral language to produce traces of verbal interactions and actions in small and large group settings (Graesser, Gernsbacher, & Goldman, 2003). Performance analysis of activities and interpersonal interactions in context (physical, cognitive, sociocultural) can afford time-sequence of talk, actions, and gesture and reveal potential relationships about these factors (Xu & Clarke, 2007). Thus, these analyses can be invaluable windows into students' representations of what they are learning and the cognitive roles of sensory experiences, prior knowledge, and language in constructing understanding. Many researchers are choosing to do these analyses without the necessary preparation, linguistic background, classroom experience, understanding of the context, language, and activities under investigation, and realization of time demands required by these techniques. Furthermore, the availability of innovative software for discourse analysis (e.g., Atlas TI™, Nudist 6™, Nvivo 7™, XSight™, etc.) and video analysis systems (e.g., StudioCode™, Transanna™, Videograph™, etc.) has fueled this popularity. Unfortunately, some researchers enter into computer-assisted analysis of text and video files without fully understanding the operator demands or coding requirements or without serving an appropriate internship with an expert user.

The inefficient use of education databases and information sources from international, national, and state/provincial tests to make political claims about national rankings and infer superior education systems need better utilization to justify these expensive, time-demanding resources. An NRC report (US NRC, 2004) recommended making data and information files available to other researchers working in similar problem spaces. This would require collaborative efforts from funding agencies, research institutions, and researchers to provide support, develop methods, and perfect procedures so as to share data in secure, ethical, meaningful, uniform, and useable ways. Structural equation modeling, hierarchical linear modeling, meta-analysis, metasynthesis, and secondary reanalysis of large datasets and collective sets of quantitative and qualitative data should be used to better explore relationships, disentangle confounded results, and produce generalized results. Such reanalyses will require secure access to individual test items and original responses.

Data security is a major dimension of complexity and concern for data sharing and secondary data analysis. Sharing data is worthwhile since it encourages diverse analyses and perspectives, fosters new research, and facilitates applications of new or additional analyses of these resources. The ultimate virtue of data sharing is to replicate and extend previous research and thereby both validate previous work and facilitate new researchers' work through consolidating and extending datasets and by avoiding unnecessary duplication of expensive data collection procedures. However, the dilemmas of value, requirement, procedure, and practice are only starting to be explored (O'Rourke et al., 2006; Rodgers & Nolte, 2006; Sieber, 2006; Zarate & Zayatz, 2006). Ethical issues abound in the realm of data sharing; and human subjects’ research requirements are needed to address issues of anonymity, confidentiality, ownership and proprietary use of data, disclosure, and data security. Procedures have been developed to de-identify data to protect confidentiality while allowing informed secondary analyses; but in the era of rapidly changing technologies, new procedures may be developed to reconstruct identity. Sieber (2007) stated that “it is recognized there is always a remote possibility of re-identification and this is a risk we live with” (p. 97).

The use of ICT should be viewed not only as oriented toward the analysis of data but also as a valuable way to collect, verify, and report data. While there has been work moving over from the gaming world of collecting keystrokes and eye movements, the current advances in videoconferencing technologies allow researchers to utilize this technology for data collection, data interpretation, and coauthoring. An essential component of research is the theoretical and methodological debates that are mandatory for quality outcomes. As such, this technological tool enables much broader opportunities for collaborative ventures from an interdisciplinary aspect and from a multisite perspective. Researchers, teachers, and other participants can interact at distances in real time and thus move research programs forward on a continuous basis. Coauthoring, a basic necessity of interdisciplinary and collaborative literacy-science research projects, can take place in real time and with oral, visual, and print media in the composing, reviewing, and revising process, which leads to constructing understanding and generating valid knowledge claims.

28.2.8 Contract Research, Commercial Research Organizations, and Researchers-for-hire

Universities and formal research institutes do not have a monopoly on literacy and science education research any more. Private foundations and industry consortia have joined research communities in many developed and developing countries to enhance educational opportunities, improve school programs, achieve governmental socioeconomic goals, and address social justice issues as well as their business goals. Funding from these private foundations and consortia and some public agencies tends to be for mission-driven contracts rather than curiosity-driven grants; therefore, this practice has led to advantages for nonuniversity-based research and contract research. Henig (2008) stated:

> In interviews with researchers active in the area of vouchers, charters, and school choice, I found that foundations were about three times as important as the federal government as a source of support for their work. The way foundations and researchers come together differs from the formalized and regulated procedures involved with federal funding; it is more common that they seek each other out based on style, trust, and compatibility of mission. Foundations in general are less concerned with peer review and sophisticated research designs and more concerned with helping to shape and disseminate findings that accord with their organizational missions. (p. 360)

Commercial research companies and contract researchers comprise a group of researchers-for-hire (consultants) that operates outside many of the issues contained in this book. Some of these organizations with long, productive histories, like the Australian Council for Educational Research (n.d.), conduct contract research and development projects and deliver products and services; Horizon Research, Inc. (n.d.) provides technological assistance and conducts contract research and evaluations in the United States. Others companies have less-tested records.

These organizations and researchers work for themselves as entrepreneurs, under employment agreements or as subcontractors, and act as agents outside of the research culture and support system of a university. This brings into question issues of outside professional consulting within academic appointments, the need for ethics approval of the contracted research, and intellectual property issues regarding products of the work. Most universities have specific policies involving outside professional activities, but few have explicit policies or procedures to cover the ethics issues of activities conducted during these extra-university activities and the intellectual properties and value assigned to such work in personnel and salary decisions. Brewer and Goldhaber (2008) stated:

> The incentives for those operating in think tanks or private-sector companies are relatively straightforward, since these firms are dependent on ‘soft money’ that comes in the form of grants or contracts. In this environment, researchers face considerable pressure to raise funds from public and private agencies, as their salary trajectory and job security are directly related to the success of any fund-raising endeavors. The upside of the research supplied by soft-money institutions is that it is likely to be policy relevant, because it is ‘client-driven’ and generally formulated to inform a specific issue for which there is an audience. (p. 362)

University literacy and science education researchers need to be cautious about such extra-to-appointment activities and ensure that issues of outside professional activities, ethics approval, ownership, and ascribed value are clearly understood before getting involved.

28.3 Closing Remarks

We have taken our lead from the authors’ and conference participants’ sincere desire to make a difference. The difference can be to advancements in knowledge about literacy and science education, to definitions of and procedural insights into quality research, to inform effective classroom practices, and to influence policies about learning, teaching, curriculum, and assessment related to science literacy for all. Interspersed throughout this book are statements of or references to powerful knowledge claims about literacy and science education, theoretical insights into the role of language in doing and learning science, and the need to position and operationally define teaching in service of learning. Likewise, interspersed are statements, procedures, and examples of high-quality research standards that elaborate and enrich how quality, rigor, and alignment of problem space, research question, canonical knowledge, available instrumentation, and technology and design interplay to produce meaningful results to important questions, insightful explanations involving cause–effect mechanisms, and effective classroom practices and guidance for future inquiries.

Furthermore, like Malcom (2008) suggested, much of the improvements in teaching and the enhancements in learning of science are engineering design (R&D) issues—not scientific inquiry issues. R&D attempts to modify natural and people-built environments to address people's needs and alleviate problems, focuses on production of useful innovations, and recognizes the different constraints and risks of reality (resources, safety, costs, demands, etc.). Like Saul and Hand (see Chap. 12), Malcom believed that we should, first, anchor our problem-seeking and research design in the needs and realities in the complexities of classrooms, students, teachers, and sociocultural influences and, second, ensure that we make significant attempts at applying any results or findings back to these realities. Starting and ending with the constraints and contextual realities may avoid the theory–practice gap found in the two worlds of our current mode of operations and produce more realistic and useful solutions. This would enhance the probability of informing classroom practice and influence public policies about literacy and science education and science literacy for all.

Literacy and science education research communities can better mobilize their resources to speak truth to power more effectively through both indirect and direct approaches to policy makers and to have an impact on the powers that be in any country and under any system of government. There are indirect ways to have an effect; for example: (a) publishing well-documented results with convincing methodology that address high-impact subjects (let's say, K-12 students' science reading proficiency) in a way that can be understood by an informed lay audience; (b) working with professional associations and possibly also with grassroots groups to extend new knowledge to a broader mass audience and mobilize potential blocs of voters; or (c) engaging with media broadcast and print outlets to provide sufficiently high-visibility appearances (e.g., on CNN, Sky News, or Euronews). The direct methods of impacting the powers that be include providing testimony or written documents for governmental units (e.g., parliamentary committees, executive commissions, task forces, and other bodies consisting of or designed to influence policy makers) or working with interest groups and political parties to help select and elect public officials with the *right* mindset on key issues of science education (e.g., supporting teaching evolution as a key part of the science curriculum) or literacy (e.g., supporting bilingual instruction or the *best* approach to ensure that students read at grade level).

Both the indirect and direct approaches require having something worth saying, the evidence to support such claims, awareness of the political process, and the willingness, communication skills, and effort to do so. Service on task forces, commissions, and inquiry panels may go without adequate recognition—but when done effectively, it does make a difference!

References

American Association for the Advancement of Science. (1990). *Science for all Americans: Project 2061*. New York: Oxford University Press. Retrieved from http://www.project2061.org/publications/sfaa/online/sfaatoc.htm

American Educational Research Association. (2006). *Standards for reporting on empirical social science research in AERA publications*. Retrieved June 23, 2008, from https://www.aera.net/uploadedFiles/Opportunities/StandardsforReportingEmpiricalSocialScience_PDF.pdf

August, D., & Shanahan, T. (2006a). Introduction and methodology. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the national literacy panel on language-minority children and youth* (pp. 1–42). Mahwah, NJ: Lawrence Erlbaum.

August, D., & Shanahan, T. (Eds.). (2006b). *Developing literacy in second-language learners: Report of the national literacy panel on language-minority children and youth*. Mahwah, NJ: Lawrence Erlbaum.

Australian Council for Educational Research. (n.d.). *Homepage*. Retrieved June 19, 2008, from http://www.acer.edu.au/

Brewer, D. J., & Goldhaber, D. D. (2008). Examining the incentives in educational research. *Phi Delta Kappan*, *89*(5), 361–364.

Brickhouse, N. W. (2006). Celebrating 90 years of *Science Education*: Reflections on the gold standard and ways of promoting good research [Editorial]. *Science Education*, *90*(1), 1–7.

Cohn, D. (2006). Jumping into the political fray: Academics and policy-making. *Institute for Research on Public Policy (IRPP) Matters, 7*(3), 8–36. Retrieved from http://www.irpp.org/pm/index.htm

Cohn, D. (2007, October). How can academics influence public policy? *Academic Matters*, 18–19.

Fensham, P. J. (2008). *Science education policy-making: Eleven emerging issues*. Paris: UNESCO. Retrieved from http://unesdoc.unesco.org/images/0015/001567/156700e.pdf

Fusarelli, L. D. (2008). Flying (partially) blind: School leaders' use of research in decision making. *Phi Delta Kappan, 89*(5), 365–358.

Ginsburg, H. P., Lee, J. S., & Boyd, J. S. (2008). Mathematics education for young children: What it is and how to promote it. *Social Policy Report, 22*(1), 3–22.

Graesser, A. C., Gernsbacher, M. A., & Goldman, S. R. (Eds.). (2003). *Handbook of discourse processes*. Mahwah, NJ: Lawrence Erlbaum.

Hand, B., Yore, L. D., & Prain, V. (Eds.). (2006). Natural science, cognitive science and pedagogical influences on science literacy: Empowering research and informing instruction [Special Issue]. *International Journal of Science Education*, *28*(2/3), 99–314.

Henig, J. R. (2008). The evolving relationship between researchers and public policy. *Phi Delta Kappan, 89*(5), 357–360.

Hess, F. M. (2008). The politics of knowledge. *Phi Delta Kappan*, *89*(5), 354–345.

Horizon Research Inc. (n.d.). *Homepage*. Retrieved June 23, 2008, from http://www.horizon-research.com/

International Reading Association & National Council of Teachers of English. (1996). *Standards for English language arts*. Urbana, IL: National Council of Teachers of English. Retrieved from http://www.ncte.org/about/over/standards?source = gs

Knott, J., & Wildavsky, A. (1980). If dissemination is the solution, what is the problem? *Science Communication*, *1*(4), 537–578.

Landry, R., Lamari, M., & Amara, N. (2003). The extent and determinants of the utilization of university research in government agencies. *Public Administration Review*, *63*(2), 192–205.

Lavis, J. N., Robertson, D., Woodside, J. M., McLeod, C. B., & Abelson, J. (2003). How can research organizations more effectively transfer research knowledge to decision makers? *Milbank Quarterly*, *81*(2), 221–248.

Lawson, A. E. (2005). Conducting high quality educational research [Editorial]. *International Journal of Science and Mathematics Education*, *3*(1), 1–5.

Lawson, A. E. (2007, March). *How "scientific" is science education research?* Paper presented at the annual meeting of the National Association for Research in Science Teaching, New Orleans, LA.

Ludwig, J., & Phillips, D. (2007). The benefits and costs of Head Start. *Social Policy Report*, *21*(3), 3–18.

Malcom, S. M. (2008, January). *Producing "high quality" teachers: The science and the education*. Keynote Address at the annual international meeting of the Association for Science Teacher Education, St. Louis, MO.

Munby, H. (2003). Educational research as disciplined inquiry: Examining the facets of rigor in our work [Guest editorial]. *Science Education, 87*(2), 153–160.

O'Rourke, J. M., Roehrig, S., Heeringa, S. G., Reed, B. G., Birdsall, W. C., Overcashier, M., et al. (2006). Solving problems of disclosure risk while retaining key analytic uses of publicly released microdata. *Journal of Empirical Research on Human Research Ethics*, *1*(3), 63–84.

Osborne, J., & Dillon, J. (2008). *Science education in Europe: Critical reflections*. London: Nuffield Foundation. Retrieved from http://www.nuffieldfoundation.org/fileLibrary/pdf/Sci_Ed_in_Europe_Report_Final.pdf

Phillips, D. C. (2006). A guide for the perplexed: Scientific educational research, methodolatry, and the gold versus platinum standards. *Educational Research Review*, *1*(1), 15–26.

Porter, A. C., & Polikoff, M. S. (2007). NCLB: State interpretations, early effects, and suggestions for reauthorization. *Social Policy Report, 21*(4), 3–14.

Rees, W. E. (2008, April-May). Science, cognition and public policy. *Academic Matters*, 9–12.

Rocard, M., Csermely, P., Jorde, D., Lenzen, D., Walberg-Henriksson, H., & Hemmo, V. (2007). *Science education now: A renewed pedagogy for the future of Europe*. Luxembourg, Belgium: European Commission. Retrieved from http://ec.europa.eu/research/science-society/document_library/pdf_06/report-rocard-on-science-education_en.pdf

Rodgers, W., & Nolte, M. (2006). Solving problems of disclosure risk in an academic setting: Using a combination of restricted data and restricted access methods. *Journal of Empirical Research on Human Research Ethics, 1*(3), 85–98.

Sieber, J. E. (2006). Introduction: Data sharing and disclosure limitation techniques. *Journal of Empirical Research on Human Research Ethics*, *1*(3), 47–50.

Sieber, J. E. (2007). Respect for persons and informed consent—A moving target. *Journal of Empirical Research on Human Research Ethics*, *2*(3), 1–2.

Simon, M. A. (2004). Raising issues of quality in mathematics education research. *Journal for Research in Mathematics Education*, *35*(3), 157–163.

Smith, C. L., & Wenk, L. (2006). Relations among three aspects of first-year college students' epistemologies of science. *Journal of Research in Science Teaching*, *43*(8), 747–785.

Society for Research in Child Development. (n.d.). *Homepage*. Retrieved June 23, 2008, from http://www.srcd.org/

United States National Research Council. (1996). *The national science education standards*. Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id = 4962

United States National Research Council. (2002). *Scientific research in education*. Committee on Scientific Principles for Education Research. R. J. Shavelson & L. Towne (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

United States National Research Council. (2004). *Advancing scientific research in education*. Committee on Research in Education. L. Towne, L. L. Wise, & T. M. Winters (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Xu, L. H., & Clarke, D. (2007, April). *Artefacts and distributed cognition: Towards a new perspective on science learning*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, New Orleans, LA.

Yore, L. D. (2003). Quality science and mathematics education research: Considerations of argument, evidence and generalizability [Guest editorial]. *School Science and Mathematics*, *103*(1), 1–7.

Yore, L. D., Hand, B., & Florence, M. K. (2004). Scientists' views of science, models of writing, and science writing practices. *Journal of Research in Science Teaching, 41*(4), 338–369.

Yore, L. D., Pimm, D., & Tuan, H.-L. (2007). The literacy component of mathematical and scientific literacy. *International Journal of Science and Mathematics Education*, *5*(4), 559–589.

Yore, L. D., & Yore, S. A. (2007, January). *Effective reporting of research results in the international, multicultural, education community: Bridging the gap between authors and readers*. Paper presented at the international meeting of the Association for Science Teacher Education, Clearwater, FL.

Zarate, A. O., & Zayatz, L. (2006). Essentials of the disclosure review process: A federal perspective. *Journal of Empirical Research on Human Research Ethics, 1*(3), 51–62.

Zientek, L. R., Capraro, M. M., & Capraro, R. M. (2008). Reporting practices in quantitative teacher education research: One look at the evidence cited in the AERA panel report. *Educational Researcher*, *37*(4), 208–216.

Name Index

Subject Index