A Gitest-Based Analysis and Evaluation in Language Testing

—A Case of a Final Exam

Gan Li[1]

[1] School of Foreign Languages, Guangdong Polytechnic College, Guangzhou, China

Correspondence: Gan Li, School of Foreign Languages, Guangdong Polytechnic College, Guangzhou, China. Tel: 86-133-1615-7898. E-mail: ganlickd@126.com

Received: May 31, 2013 Accepted: July 1, 2013 Online Published: August 15, 2013

doi:10.5539/ells.v3n3p88 URL: http://dx.doi.org/10.5539/ells.v3n3p88

Abstract

This paper analyses and evaluates an English final exam paper through GitestIII, which was developed and popularized by Guangdong University of Foreign Studies. The analysis includes detailed item response analysis for objectives, obtaining such information as reliability, validity, difficulty, discrimination index, etc. With this case study done, language teachers would be encouraged to adopt the application of modern statistics such as Gitest and SPSS in similar exams to get more useful information about the exams, thus guiding the teaching and testing more efficiently and scientifically.

Keywords: Gitest, analysis and evaluation, language testing

1. Introduction

Nowadays, technology becomes more and more developed, which makes testing and teaching more and more scientific, efficient and effective and *"Gitest" and SPSS* are among the most basic tools in language testing. With the help of them, teachers could get detailed information about the tests so as to guide the teaching and testing in an objective way. And this paper shows the practice of analysis and evaluation of an examination paper.

This final examination is an achievement test given to the 47 freshmen of Grade 2009 of three-year-course Business English majors in Guangdong Polytechnic College (a higher vocational college in Guangzhou) at the end of the term. It is based on the course syllabus or the textbook—*Integrated Skills of English (Book 1)* to see what the students have learnt within that term. The simple objective of the analysis and evaluation of the test is to help us language teachers know what the test-takers have learned about and write better tests. Test construction is essentially a matter of problem solving, with every teaching situation setting a different testing problem (Alderson et al., 2000). It is necessary to understand the principles of testing and how they can be applied in practice including validity, reliability, practicality, and beneficial backwash etc because too often language tests fail to measure accurately whatever it is they are intended to measure (Arthur Hughes, 1989). Scientific methods are needed to standardize all kinds of language tests to get beneficial backwash on teaching and learning.

2. Theoretical Framework

2.1 Reliability

The scores of the timed test are used to test the reliability and validity of the test paper, because reliability and validity are the major criteria to test the scores of a test. "Reliability is concerned with the consistency of examinee performance, and the quantification of that consistency (or inconsistency) is the business of reliability analysis." remarked Robert Wood (1993: 132) Reliability is the extent to which test scores are to be consistent and perfect reliable scores would be accurate or free from errors of measurement. There are many factors that affect performance other than abilities we want to measure on tests and the constitute sources of measurement error. We should try to minimize the effects of factors we don't want so as to maximize the effects of the ability we want to measure and thus to maximize reliability of test scores. (Bachman, 1999: 161)

2.2 Validity

Validity is the extent to which a test measures what it is intended to measure: it relates to the uses made of test scores and the ways in which test scores are interpreted, and is therefore always relative to test purpose. We must ensure that it measures the ability in and of itself, and nothing else. Validity is ultimately more important than reliability of the observations. If these cannot be trusted, then a misleading judgement concerning validity is likely to be reached.

When validity standards were first codified, five types of validity were identified: Content validity; Construct validity; Predictive validity; Concurrent validity; Face validity. In this case study, Construct validity is examined. The determination of construct validity is essentially a search for evidence that will help us understand what the test is really measuring and how the test works across a variety of settings and conditions.

Technically, tests do not measure constructs directly, rather, they measure performance or behavior that reflect constructs. If tests (or items) measure the same constructs, scores on the tests should be correlated; conversely, scores on tests that measure different constructs should have low correlations.

Because of the potential ambiguities involved in interpreting a single correlation between two tests, correlational approaches to construct validation of language tests have typically involved correlations among large numbers of measures.

A commonly used statistical procedure for interpreting a large number of correlations is factor analysis, which analyzes a set of correlation coefficients between measures and identifies the number and nature of the constructs underlying the measures.

These interpretations about reliability and validity echo Lado (1961), Bachman (1990), Weir (1990), Bachman and Palmer (1996), Hugh (1989), etc.

2.3 About Gitest

Gitest is an examination data analysis system, which was developed and popularized by Guangdong University of Foreign Studies (GUFS). It has three main functions as follows: (1) original exam data edition; (2) exam data analysis based on the classical testing theory and the modern testing theory. For example, Item Response Theory (IRT) in Rasch Model, which gained much popularity in many studies on language testing (Zhang, 2006; Myford and Wolfe, 2000; Weir and Milanovic, 2003) and the data analysis of fitness between test papers and examinees. The results include a lot of information such as each examinee's score on each specific item and their total score of the paper, the mean score, standard deviation, and percentile for both a specific item and all the items, proportion correct for each specific item, reliability, discrimination index, correlation matrix for the scores of each subtest, the distribution of all the scores, the distribution of difficulty level and discrimination index for each subtest, etc.; (3) print the results.

Gitest can be used in many situations, especially suited for objective items grading, statistics and analysis (Wang, 2009, Ding 2010, Jin, et al., 2011). Besides it can be applied to analyze a paper with less than 200 items and upto ten thousand examinees. The whole processing can be finished within several minutes or half an hour.

The following is a list of printed data from Gitest.

(1) ITEM ANALYSIS TABLE

(2) TEST TABLE AND SUBTEST TABLE

(3) P-VALUE AND R-BIS. CROSS TABLE

(4) CORRELATION MATRIX AND FACTOR LOADINGS

3. Method and Procedure

The test consists of two parts-subjective and objective. "Gitest" is used to analyze the reliability and validity of the Part I Listening, Part II Vocabulary & Structure, Part III Cloze, Part IV Reading, all the objective parts of the exam which are required to select the correct answers from a choice of ABCD four answers (or MC-multiple-choice questions). Only one of them is correct for each item. And the evaluation of the subjective parts is omitted here and can be analyzed with SPSS. The test paper is designed according to the mainstream of tests nowadays. Part I Listening is made up of only 4 items which has the same form from the text. According to Li: the number of items of section is controlled from 10-20 at minimum. Part IV Reading is composed of only two passages, because the students are from a vocational college, most of whom came from secondary vocational school where they seldom learn English before. So their basic knowledge is poor. There are 60 items in the test, that is less than the 80-100 required in a standard test (Li, 2001:71), but it is fair to the test-takers. This paper covers only 70% of the total scores and the other 30% come from their accumulated results. The following table presents the format of the test.

Table 1. The construction of the test

Part	Content	Rubric	Number	Score (%)	Time
Part I	Listening	Multiple Choice	1-4	8%	
Part II	Vocabulary & Structure	Multiple Choice	5-24	20%	
Part III	Cloze	Multiple Choice	25-39	15%	
Part IV	Reading	Multiple Choice	40-49	20%	
Part Ⅴ	Translation	Translate English into Chinese	50-54	10%	
		Translate Chinese into English	55-59	10%	
Part Ⅵ	Writing	Retelling	60	17%	
Total			60	100%	120mins

4. Results and Analysis

4.1 Reliability

Table 2 shows R11—reliability is 0.84, which is a little below the ideal point of 0.9 (Li, 2001: 100), and aVALUE (an alpha coefficient) is 0.72, a little less than the minimum standard of 0.8 (Li, 2001: 100), which is within Lado's (1961) acceptable ranging from 0.70 to 0.79 at minimum. The test has a mean score of 29.98, having 61.18% of the answers correct, indicating it is a medium-level test, the standard deviation is 7.71 against the expected standard deviation 7.16, having a range of 29, indicating the scores are widely spread. It has the fair discrimination index Rbis 0.59. The skew -0.09 and kurt -0.78 shows that they are beyond the normal distribution range of 1 ~ -1(Li, 2001: 97), so it is a negative Skewed Distribution, which means that most students get high scores. On the whole, the distribution of the scores is normal. The test is reliable, the standard error of measurement is +- 3.04. See table 2.

Table 2. Final test table

Total No. of items: 49 Total No. of subjects: 4

Date of test: 2005 Date of analysis: 01-16-2006

Mean	SD	Varn	p+	pd	R11	Rbis	Value	Skew	Kurt
29.98	7.71	59.46	0.61	11.86	0.84	0.59	0.72	-0.09	-0.78

4.2 Validity

In the following parts, we will test the validity of the objective items through correlation, factor analysis and item analysis.

4.2.1 Construct Validity

A test, part of a test, or a testing technique is said to have construct validity if it can be demonstrated that it measures just the ability which it is supposed to measure. The 'construct' refers to any underlining ability which is hypothesised in a theory of language ability (Hughes, 2000:26). The content of test needs to be in accordance to, or to be pertinent to and representative of the content of a teaching syllabus and testing syllabus. Table 3 is the correlation matrix of the different parts within the test and correlation between each part and the total scores.

Table 3. Final correlation matrix

No. of candidates: 47

Correlation	0	I	II	III	IV
0	1.00	0.55	0.86	0.82	0.78
1		1.00	0.35	0.49	0.32
2			1.00	0.51	0.52
3				1.00	0.59
4					1.00

Mean	29.98	3.00	12.17	9.15	5.66
S D	7.71	0.96	3.85	2.66	2.11

0=total 1=lis 2=voc 3=cloze 4=read

All the four sections are correlated with the total at the 0.01 level, indicating that all of them contribute to general language proficiency. The part of Vocabulary has the highest correlation coefficient (.86) with the total. This indicates that this part best reflects the students' English performance. Cloze has the second highest correlation coefficient (.82) with the total, followed by Reading (.78) and listening (.55).

As shown in Table 3, there is a 0.35 correlation between Section 1 and Section 2; 0.49 between Section 1 and Section 3; 0.32 between Section 1 and Section 4, 0.51 between Section 2 and Section 3, 0.52 between Section 2 and Section 4, 0.59 between Section 3 and Section 4, indicating that what they each test has significantly correlated.

Additional information which helps in the interpretation of construct validity is provided by the data of different sections, which were factor analyzed, using principle component analysis to extract only 2 underlying factor which accounted for 64% of the total variance (see Table 4).

Table 4. Factor loadings

Have been varimax rotated

	Factor1	Factor2	Factor3
1	0.222	-0.634	-0.135
2	0.501	-0.260	-0.390
3	0.586	-0.500	-0.104
4	0.716	-0.202	-0.153

P.C% 28.94 18.99 5.11

Loading P.C% 53.03

1=lis 2=voc 3=cloze 4=read

Factor analysis of the test shows that it accounts for 53.03% of the knowledge and skills of the examinees. Table 4 shows the correlation of the 4 subtests with hypothetical Factor 1 to Factor 3. Factor 3 is not significant for "a single section's loading under 10% is pointless (Li, 2001: 103). Factor 1 seems to be a knowledge factor for Section 2 Vocabulary, with a correlation of 0.501, Section 3 Cloze, with a correlation of 0.586 and Section 4 Reading, with a correlation of 0.716 with it. This means that vocabulary, cloze and reading test something similar. They belong to one factor. Listening belongs to another factor.

4.3 Item Analysis on Objective Items

Table 5. P-value and R-BIS. cross table

R / P	0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0	Total
<0.1	0	0	0	0	0	1	0	0	0	0	1
0.1-0.2	0	0	0	0	1	0	0	2	0	0	3
0.2-0.3	0	0	0	0	1	0	1	0	1	0	3
0.3-0.4	0	0	0	0	1	0	0	0	0	0	1
0.4-0.5	0	0	1	2	0	2	3	0	1	0	9
0.5-0.6	0	0	0	0	1	4	1	2	2	0	10
0.6-0.7	0	1	0	0	1	3	1	1	3	0	10
0.7-0.8	0	0	0	0	1	1	2	1	3	0	8
0.8-0.9	0	0	0	0	0	2	0	1	0	0	4
Total	0	1	2	2	6	13	8	7	10	0	49

VD=5% D =5% I=60% E= 15% VE= 5%

Item analysis shows that there are 0% very easy items, 34.69% easy items, 59.18% items of medium level, 6.12% difficult items, 0% very difficult items. On the whole, the test has good discrimination power. There are 7 items that do not meet the requirement.

The table below explains the criteria of each item indicators.

Table 6. Criteria for item analysis

		Name	Symbol	Numerical range	Ideal value	Reference controlled value
The item index		proportion correct	P or Pt	0~1	.5	70% of the number of items fall within .3~.7; 15% of the number of items under .3; 15% of the number of items above .7.
		Difficulty level	Δ or Pd	1~25	13	68% of the number of items fall within 9~17.
		Discrimination index	r_{bis}	-1~1	>.3, the larger, the better	95% of the number of items are larger than .3; Items with a value smaller than .2 cannot be adopted.
The option index	the correct answer index	the number of examinees	n or p	0~100% of the number of examinees	50% of the number of examinees	Try to make it fall within 20~80%.
		Mean score	m or m_s	1~25	>14, the larger, the better	Keep it above 13, or consider revising items.
		Discrimination index	r	-1~1	>.3, the larger, the better	Keep it above .3, or consider revising items with a value of or under .2
	distractor index	the number of examinees	n or p	0~100% of the number of examinees	10~25% of the number of examinees	fall within 5~30%; those smaller than 5% are invalid; those above 30% are too strong.
		mean score	m or m_s	1~25	<12, the smaller, the better	Keep it under 13, or consider revising items.
		Discrimination index	r	-1~1	>.1, the larger, the better	Keep it under 13, or consider revising items.

4.4 Item Analysis of Listening, Vocabulary, Cloze and Reading

Table 7. Final subtest table (listening)

No. of items: 4 (from 1 to 4)

Mean	SD	Varn	p+	pd	R11	Rbis	Skew	Kurt
3.00	0.96	0.91	0.75	10.29	0.35	0.81	-0.91	0.68

Difficulty total No.(<.3) items

Difficulty	Total	No.(<0.3)
VD	0	0
D	0	0
I	1	0
E	3	0
VE	0	0

Final subtest table (vocabulary)

No. of Items: 20 (from 5 to 24)

Mean	SD	Varn	p+	pd	R11	Rbis	Skew	Kurt
12.17	3.85	14.80	0.61	11.90	0.76	0.57	-0.22	-0.48

Difficulty	Total	No.(<0.3)	Items
VD	0	0	
D	1	0	
I	13	1	16
E	6	1	21
VE	0	0	

Final subtest table (close)

No. of items: 15 (from 25 to 39)

Mean	SD	Varn	p+	pd	R11	Rbis	Skew	Kurt
9.15	2.66	7.09	0.61	11.88	0.56	0.50	-0.19	-0.88

Difficulty	Total	No.(<0.3)	Items
VD	0	0	
D	0	0	
I	10	3	36/37/39
E	5	1	31
VE	0	0	

Final subtest table (reading)

No. of items: 10 (from 40 to 49)

Mean	SD	Varn	p+	pd	R11	Rbis	Skew	Kurt
5.66	2.11	4.45	0.57	12.34	0.58	0.62	-0.62	-0.54

Difficulty	Total	No.(<0.3)	Items
VD	0	0	
D	2	0	
I	5	0	
E	3	1	40
VE	0	0	

Analysis of each component shows that

Listening has a mean score of 3, having 75% of answers correct, indicating it is a fairly easy component. On the whole, in this component listening, there are 0 item(s) that do not meet the requirement. It has fair DI (0.81≥0.3) on the whole, and low reliability of only 0.35 because of only 4 items of listening in this exam.

Vocabulary has a mean score of 12.17, having 60.85% of answers correct, indicating it is a medium-level component. On the whole, in this component vocabulary, there are 2 items 16 of I and 21 of E that do not meet the requirement.

Cloze has a mean score of 9.15, having 60.99% of answers correct, indicating it is a medium-level component. On the whole, in this component cloze, there are 4 items (36, 37, 39, and 31) that do not meet the requirement.

Reading has a mean score of 5.66, having 56.6% of answers correct, indicating it is a medium-level component.On the whole, in this component reading, there is 1 item (40) that does not meet the requirement.

Table 8. Final item analysis

Date of test: 2005 Date of analysis: 02-16-2006

Test code	Item No.	Pt	Pi	Pd	P	MA	MB	MC	MD	MO
gana	Voc 16	11.86	11.90	12.89	0.51	13.24	13.26	10.40	12.62	0.34
No. of Can	Key	Ar	Br	Cr	Dr	A	B	C	D	O
47	B	-0.05	0.08	0.33	0.06	15	24	3	5	0
Test code	Item No.	Pt	Pi	Pd	P	MA	MB	MC	MD	MO
gana	Voc21	11.86	11.90	8.02	0.89	0.34	13.19	12.30	10.74	0.34
No. of Can	Key	Ar	Br	Cr	Dr	A	B	C	D	0
47	B	0.00	0.24	0.08	0.28	0	42	2	3	0

Notes: Pt=difficult index of the whole test; Pi=difficulty index of the subtest of listening; Pd=difficulty index of the item; P=facility value; Ar=discrimination index (DI) for Option A in this subtest, (reason out the rest by analogy); A=number of test takers choosing Option A, (reason out the rest by analogy).

As shown in Table 8, in Item 16, B is the key to the choice. 24 out of 47 test-takers chose the right option, B and 15 out of the 47 test-takers chose A, that means the distractions of C, D are too weak while A is too strong. Db is 0.06 ≤0.3, its distraction A bears negative Ar -0.05, implying bottom learners have done better than top learners. Therefore, this item needs to be moderated.

Item 21 is an easy item with low Ar 0.00, Cr 0.08 ≤0.3 and low Pd 8.02≤13, because 42 out of test-takers choose the right option B. That means other distractions are too weak, especially C, which no one chose. Therefore, this item also needs to be moderated.

5. Conclusion

On the whole, the distribution of the scores is normal. The test is reliable, the standard error of measurement is +- 3.04. There are only 60 items in the test, which is less than the 80~100 required in a standard test (Li, 2001:71) and should be adjusted nest time. Gitest is a very good and scientific means of analyzing test papers, which teachers should use in teaching. With the help of Gitest and SPSS, teachers could get detailed information about the tests in order to guide the teaching and testing more efficiently and objectively.

References

Alderson, J. C., Clapham, C., & Wall, D. (2000). *Language Test Construction and Evaluation*. Beijing: Foreign Language Teaching and Research Press. Cambridge: Cambridge University Press.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford University Press.

Bachman, L. F., & Cohen, A. D. (2002). *Interfaces Between Second Language Acquisition and Language Testing Research*. Beijing: Foreign Language Teaching and Research Press.

Ding, X. (2010). Study of reliability and validity in performance testing using Gitest – A case of College English. *Journal of Petroleum Educational Institute of Xinjiang, 11,* 296-298.

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Hughes, A. (2000). *Testing for Language Teachers*. Beijing: Foreign Language Teaching and Research Press. People's Education Press. Cambridge: Cambridge University Press.

Henning, G. (2001). *A Guide to Language Testing: Development, Evaluation and Research.* Beijing: Foreign Language Teaching and Research Press.

Jin, X., & Zhou, J. (2011). A Gitest-based item analysis in of Multiple Choice – A case of a final exam paper for Advance Career English. *China Electric Power Education, 16,* 216-218.

Lado, R. (1961). *Language Testing – the Construction and Use of Foreign Language Tests*. London: Longman.

Li, X. (2001). *The Science and Art of Language Testing* (New Ed.). Changsha: Hunan Educational Press.

Liu, Runqing, & Han, Baocheng. (1991). *Language Testing and Its Methods*. Beijing: Foreign Language Teaching and Research Press.

Myford, C. M., & Wolfe, E. W. (2000). Monitoring sources of variability within the test of spoken English assessment system. *TOEFL Research Report No. 65*. Princeton, NJ: Educational Testing Service.

Wang, B. (2010). Gitest III Item Analysis of a Listening Test. *Journal of Hunan University of Science and Engineering, 5,* 256-258.

Weir, C. J. (1990). *Communicative Language Testing*. London: Prentice Hall.

Weir, C. J. (2005). *Language Testing and Validation: an Evidence Based Approach.* Basingstoke: Palgrave Macmillan.

Weir, C. J., & Milanovic, M. (Eds.). (2003). *Continuity and Innovation: The History of the CPE* 1913-2002. Cambridge: Cambridge University Press.

Weir, C. J., & Wu, J. (2006). Establishing test form and individual task comparability: A case study of a semi-direct speaking test. *Language Testing, 23,* 167-197. http://dx.doi.org/10.1191/0265532206lt326oa

Wood, R. (2001). *Assessment and Testing: A Survey of Research*. Beijing: Foreign Language Teaching and Research Press.

Wood, R. (2001). *Assessment and Testing: A Survey of Research.* Beijing: Foreign Language Teaching and Research Press. People's Education Press. Cambridge: Cambridge University Press.

Zhang, J. (2006). Study of sources of score variability in performance testing using Many-facet Rasch Model – A case of CET-SET (College English Test-Spoken English Test). Unpublished MA dissertation, Zhejiang University, China.

Copyrights