Quality Analysis of an English Test Designed against the Framework of *China' s Standards of English Language Ability*

Haizhen Zhao

School of Foreign Languages, Guilin University of Electronic Technology, Guilin, China
**Corresponding author. Email: 7629332@qq.com*

ABSTRACT

Taking a final college English test for non-English majors as an example, this study describes the design and content of the test against the framework of China' s Standards of English Language Ability. 363 test takers are selected as samples and SPSS software is used to analyze the quality of the test from perspectives of reliability, validity, difficulty and discrimination. The results show that the overall quality of the test is good, but the type and difficulty of some test items need to be improved and adjusted. The paper also provides a positive exploration and experiment for the application of China' s Standards of English Language Ability in college English assessment. Moreover, the paper provides guidance for the the alignment of the China' s Standards of English Language Ability with the school-based English test of the author' s university.

Keywords: *China' s Standards of English Language Ability, English test, Reliability, Validity, Difficulty, Discrimination.*

1. INTRODUCTION

As a measure tool, language tests play an important role in language teaching and learning. Language tests can evaluate learners' language ability, diagnose their learning and be used for screening and selection for various purposes. In addition, as an important part of the teaching and learning process, tests can provide rich feedback information for teachers and promote teaching accordingly. If designed scientifically, language tests can provide accurate and objective descriptions of current teaching and learning, provide sufficient feedback for teaching, and help teachers improve their teaching [1]. Therefore, the analysis and evaluation of tests is very crucial. Scientific analysis and interpretation of test results are indispensable to properly evaluate a test and make decisions accordingly.

2. CHINA' S STANDARDS OF ENGLISH LANGUAGE ABILITY

The design and development of tests cannot be achieved without scientific standards, and the release of China ' s Standards of English Language Ability provides an important reference for language assessment. In 2018, China's Standards of English Language Ability (hereinafter referred to as "the Standards"), the country's first English proficiency standard for English learners, was jointly issued by the Ministry of Education and the State Language Commission, aiming to build a bridge between English learning, teaching and assessment at all levels of education, and to achieve the same standard and balance in English teaching and assessment. The standards provides specific guidance for the teaching and learning of English at university level. In terms of content, the Standards gives a comprehensive and detailed description of learners' English proficiency, taking into account the unique characteristics of the English learning environment and English learners. Language abilities are divided into different aspects such as listening, speaking, reading, writing, translating and learning strategy, each of which is divided into nine levels, and each kind of ability is described by several sub-skill scales. The release of the Standards builds a bridge between language learning, teaching and assessment in the Chinese context, and provides a unified scale of language competence [2].

Many universities now implement school-based English proficiency tests, i.e., English proficiency tests designed to suit the level of their students. If school-based English proficiency tests can be carried in alignment with the Standards, they can play a more effective role in guiding teaching and learning. In college English learning, teaching and assessment, the Standards serves as a bridge, giving English learning, teaching and assessment a common reference and achieving comparability between them [3].

3. INTRODUCTION TO THE SCHOOL-BASED ENGLISH TEST

According to Alderson, test design involves setting test standards, publishing syllabus, determining the content and structure of the test and its form, and determining the scoring system [4]. This process involves determining the language ability to be measured and how they are to be tested. A sound design of questions is therefore a prerequisite for scientific testing and an important guarantee.

According to the classification of test purposes by Runqing Liu and Baocheng Han [5], the English final examination of the author ’ s university is an achievement test, which is mainly used to evaluate students' mastery of English during a semester.The test in this paper is the English final examination for non-English majors at the end of their third semester. The test is designed on the Syllabus for College English and the ability description in the Standards. First of all, the test structure is defined, then the corresponding levels in the Standards were marked against each of the teaching requirements in the syllabus (listening comprehension, reading comprehension, writing and translating), and finally the corresponding test items were designed according to the proficiency descriptions in the levels. The test structure and the corresponding scores for the language skills examined are shown in Table 1.

Table 1. The language abilities evaluated in the test

Test section	Language ability					
	Understanding	Expressing	Pragmatic competence	Language Strategy	Translating	Total
Listening comprehension	10		5	5		20
Reading comprehension	20		10	5		35
Vocabulary and Structure	10		3	2		15
translating					15	15
writing		13		2		15
Total	40	13	18	14	15	100

As shown in the table, the test covers five sections. Each section is designed to test certain language abilities like understanding, expressing, translating, etc.

Specifically, the test items are selected and marked in terms of discourse difficulty, discourse type and cognitive ability based on the description of sub-skill levels in the Standards to ensure that the test is scientific. In this paper, the section of reading comprehension is used as an example and the test information is shown in Table 2.

Table 2. Statistics on the reading materials, discourse types, types of abilities assessed and scores in the section of reading comprehension

Passages	Textual theme	Discourse type	Language abilities tested and item numbers
Passage 1	Man and Society (football fans)	exposition	Comprehension of details (2) Inferring (2) Word guessing (1)
Passage 2	Man and Society (Punishment of Criminals)	argumentation	Inferring (4) Comprehension of details (1)
Passage 3	Man and Nature (energy use)	exposition	Inferring (2) Comprehension of details (1) summarizing (2)

It is shown from the table, that passages of different discourse type are well selected and each passage is aimed at testing at least two cognitive abilities.

4. ANALYSIS OF TEST QUALITY

4.1 Research Method

This paper intends to analyze and evaluate the scores of 363 test takers from the School of Business and the School of Information and Communication at Guilin University of Electronic Technology from the perspective of language testing. By analyzing the reliability, validity, difficulty and differentiation of this test, the paper will evaluate the quality of the test, identify the problems in the alignment of the school-based test and the Standards, and provide a scientific basis for better application of the Standards to improve the assessment of English.

4.2. Test Analysis

Once a paper has been designed and implemented, how can you tell how good it is? Does it meet the requirements of a language test? This requires an assessment and examination of its quality. The main criteria for assessment are generally its validity, reliability, difficulty, discrimination, practicality and back-wash effect. In this paper, the first four items are analyzed as follows.

4.2.1. Validity

Validity is the extent to which the scores from a measure represent the variable they are intended to or simply the accuracy and usefulness of a test. As a measure of a test, validity is an important indicator of test quality. There are various types of validity such as content validity, construct validity and criterion validity. In this paper, criterion validity is used to analyse the test. Criterion validity indicates how well the scores or responses of a test converge with criterion variables with which the test is supposed to converge [6]. The measurement of validity is to correlate the results of a test paper with the validity criteria, with a higher correlation coefficient indicating a better correlation between the test and the validity criteria and vice versa. A correlation coefficient of 0.4 to 0.8 is generally considered desirable. The strength of correlation is usually judged by the following range of values: a correlation coefficient of 0.8-1.0 indicates a very strong correlation, 0.6-0.8 indicates a strong correlation, 0.4-0.6 indicates a moderately strong correlation, 0.2-0.4 indicates a weak correlation and 0-0.2 shows a very weak or no correlation [7].

Analysis with SPSS software shows that the test has a criterion validity of 0.754, indicating a strong correlation. In other words, the test has a high validity. The validity correlation is shown in Table 3:

Table 3. Validity correlation

		Grades	Total score
Reading	Pearson correlation	1	.754
	Significance (two tails)		.000
	N	363	363
Total score	Pearson correlation	.754	1
	Significance (two tails)	.000	
	N	363	363

4.2.2. Reliability

Reliability refers to how dependably or consistently a test measures a characteristic. If a person takes the test again, will he or she get a similar test score, or a much different score? A test that yields similar scores for a person who repeats the test is said to measure a characteristic reliably. Reliability is usually expressed in terms of internal consistency. The higher the reliability coefficient, the more consistent, stable and reliable the test results are. A coefficient of 0.60 is generally used as the threshold value for reliability, below which the test is considered to be of low reliability and of no practical use. As shown in the table below, the reliability coefficient of this test is 0.611, which indicates that the test has good stability and reliability. The figure is shown in Table 4.

Table 4. Reliability statistics

Cronbach's alpha	Number of items
.611	5

4.2.3. Difficulty

Reliability and validity are measured for the whole test, while difficulty is for each test item. Therefore, in addition to the two important indicators of reliability and validity, it is important to analyse the quality of a paper by examining the difficulty index of the questions in a test. The difficulty of a question is usually expressed in terms of P. The P value actually refers to the percentage of questions answered correctly. In calculations, the ratio of the mean score on the question to the full mark of the question is usually used to derive the P value. The P value for the test in this study is shown in Table 5.

Table 5. Difficulty

Section	Full mark	Mean score	Scoring percentage	Coefficient of difficulty
Listening comprehension	20	15.47	77.35	0.77
Reading comprehension	35	22.38	63.94	0.64
Vocabulary and Structure	15	5.74	38.27	0.38
Translating	15	10.096	67.31	0.67
Writing	15	10.58	70.53	0.71

There is a range of values to indicate the difficulty of a question in terms of P value. In general, a difficulty coefficient in the range of 0.5 has a good degree of difficulty, with questions greater than 0.8 being too easy and questions less than 0.3 being too difficult. The P value of the test shows that the difficulty coefficient of the questions in this test are basically concentrated between 0.6 and 0.77, which is a moderate level of difficulty. However, it is worth noting that the difficulty level for vocabulary and structure was only 0.38, indicating that the questions were on the difficult side. However, on the whole, the test is moderately difficult, scientific and reasonable.

4.2.4. Degree of discrimination

In addition to difficulty, another measure of the quality of a test is the degree of discrimination. Discrimination refers to the extent to which a question distinguishes a test taker's ability. When measuring discrimination, the total test score is usually used as the actual level of ability of the test taker, and the correlation coefficient between the test taker's score on a question and the total score is used as the discrimination of that question. In this study, the correlation analysis was conducted using Spearman's scale for objective questions and Pearson's scale for subjective questions. The results of the analyses are shown in Tables 6 and 7.

Table 6. Discrimination of objective questions

			Listening	Reading	Vocabulary	Total score
Spearman rank correlation coefficient	Listening	The correlation coefficient	1.000	.267**	.188**	531.**
		Significance (two tails)	.	.000	.000	.000
		N	363	363	363	363
	Reading	The correlation coefficient	.267**	1.000	.252**	.756**
		Significance (two tails)	.000	.	.000	.000
		N	363	363	363	363
	Vocabulary	The correlation coefficient	.188**	.252**	1.000	.487**
		Significance (two tails)	.000	.000	.	.000
		N	363	363	363	363
	Total score	The correlation coefficient	.531**	.756**	.487**	1.000
		Significance (two tails)	.000	.000	.000	.
		N	363	363	363	363

Table 7. Discrimination of subjective questions

The correlation

		Translating	writing	Total score
Translating	Pearson Correlation coefficient	1	.350**	.669**
	Significance (two tails)		.000	.000
	N	363	363	363
writing	Pearson Correlation coefficient	.350**	1	.607**
	Significance (two tails)	.000		.000
	N	363	363	363
Total score	Pearson correlation	.669**	.607**	1
	Significance (two tails)	.000	.000	
	N	363	363	363

**The correlation was significant at 0.01 confidence (two tails)

The degree of discrimination for the objective and subjective questions were 0.531 for listening comprehension, 0.756 for reading comprehension, 0.487 for vocabulary and structure, 0.669 for translating and 0.607 for writing. The correlation coefficient between the other sections and the overall score were all in the range of 0.5-0.7, which is relatively close to the classical measurement theory that requires a correlation coefficient of 0.7 between the sections and the overall score. It is generally considered desirable to have a discrimination level greater than 0.3, which shows a reasonable degree of discrimination for each section.

5. CONCLUSIONS

In this study, SPSS software was used to analyse the quality of an English test in terms of validity, reliability, difficulty and degree of discrimination. In general, the test is reliable and scientific. However, the test also has a number of shortcomings, such as the difficulty of the section of vocabulary and the small proportion of subjective questions. Designed and developed against the framework of China ’ s Standards of English Language Ability, the test provides a positive exploration and experiment for the application of the Standards to English assessment in the author ’ s university, and is a guide for the alignment of the Standards with the school-based English proficiency examinations in the future.

ACKNOWLEDGMENTS

This work is supported by Guangxi Social Science Think Tank Key Project (zkzdktw01), the Teaching Reform Project of Guangxi Higher Education (Grant No. 2020JGZ117, 2021JGA178, 2021JGA188), the Teaching Reform Project of Guilin University of Electronic Technology (Grant No. JGB202015, JGB202116).

REFERENCES

[1] Yang. HZ, valid testing, effective teaching, and valid test use, in: Journal of Foreign Languages, 2015, pp.2-26. (In Chinese)

[2] Liu. JD, Han. BC, theoretical considerations for developing use-oriented China’s standards of English, in: Modern Foreign Languages, 2018, pp.78-90. (In Chinese)

[3] Wang, H, college English teaching and assessment in the context of China’s standards of English language ability, in: Contemporary Foreign Language Studies, 2018, pp 57-61. (In Chinese)

[4] Alderson. J C, Clapham, C, Wall. D, Language Test Construction and Evaluation, Cambridge University Press, 1995.

[5] Liu. RQ, Han. BC, Language Testing and its Methods, Foreign Language Teaching and Research Press, 2000. (In Chinese)

[6] Tim. McNamara, Language Testing, Shanghai Foreign Language Education Press, 2003.

[7] Yang. DH, A Complete list of Examples of Language Research Applications of SPSS Software, China Social Sciences Press, 2004. (In Chinese)