

Analysis of English Testing Based on Reliability Theory in Language Testing—Taking the English Midterm Test for the First Grade of a Middle School in Hangzhou as an Example

ZHENG Qing-qing

University of Shanghai for Science and Technology, Shanghai, China

Based on the reliability theory in language testing, this study analyzes the midterm English test of a first grade class in Hangzhou using SPSS software. We will analyze the overall structure and reliability of the test paper, and use cloze as a representative of objective questions to calculate their difficulty and discrimination, in order to explore the quality of the questions and the mastery of the knowledge learned by students. The purpose of this analysis and research is to improve the quality of exam questions and help teachers get useful information from exam results, thereby improving teaching quality.

Keywords: English test, midterm examination, reliability

Introduction

Language testing is a discipline that focuses on using scientific methods to measure language ability. It mainly studies how to design, implement, analyze, and interpret language tests, aiming to scientifically, objectively, and accurately evaluate the language ability of learners. Language testing is an important branch of applied linguistics, which is an interdisciplinary and practice oriented discipline that involves multiple fields such as linguistics, psycholinguistics, education, language acquisition theory, etc. The research on language testing not only focuses on the scientific measurement of language ability, but also on the impact and role of language testing on language teaching. This discipline standardizes and defines large-scale exams, such as the college entrance examination and CET4 and CET6, through the principles of testing, to ensure that these exams can effectively test the abilities of the test takers and provide scientific basis for educational decision-making.

Reliability

Language testing is essential in language teaching and learning. According to Bachman and Palmer (1999), the evaluation criteria for test quality generally include six aspects: reliability, validity, authenticity, interactivity, influence power, and practicality. The two most important factors in testing quality are reliability and validity.

Among them, reliability is the reliability of the test results; it embodies two relationships, scores and test questions, test questions and test takers (Li Xiaojun, 2002). Reliability is mainly influenced by factors such as the

level of the test takers, the quality and quantity of test questions, the evaluation of exam results, and the organization and implementation of the test. Generally speaking, the greater the difference in the level of the test takers, the higher the reliability of the test questions, the more questions there are, and the higher the reliability.

According to Bachman, reliability is divided into Test test reliability, Alternative form reliability, Scorer reliability, and Internal Consistency reliability. The former three types of reliability are difficult to estimate based on the results obtained from one test paper and one test, while Internal Consistency reliability can overcome many limitations and better display the reliability of the test. Therefore, this study takes the mid-term test questions of a middle school in Hangzhou as an example, focusing on analyzing the overall structure of the test papers, and its internal consistency reliability, to explore whether this test can accurately reflect students' competence and provide teachers with correct feedback.

Research Design

Research Method

This study used IBM SPSS 27.0 statistical analysis tools to calculate the reliability of the test paper, and used quantitative analysis methods to analyze the collected scores. Firstly, this article provides an overview and analysis of the overall structure of the test paper, including the distribution of test questions, a histogram of score distribution based on original scores, and the calculation of difficulty and discrimination. And use cloze as a representative of objective items for further analysis. Based on the statistical analysis of its data and the calculation of F.V. and D.I. values, suggestions are proposed for teaching and test design.

Research Subjects

The participants of this test are 44 first year middle school students who have been studying middle school English for nine months. The entire exam consists of six parts: listening, cloze, reading comprehension, grammar, vocabulary test and writing. Although there are various types of exam questions, the main test content is still grammar and vocabulary. The maximum score for this exam is 120 points, with objective questions accounting for 100 points and subjective questions accounting for 20 points. As a mid-term exam, its main purpose is to evaluate students' grammar and vocabulary learning achievements in the past half semester. Therefore, when designing exam questions, it mainly focuses on objective questions to test students' basic knowledge.

Data Analysis and Discussion

We will input the scores of each part of the test into SPSS 27.0 software to obtain relevant tables and histograms such as mean, median, variance, etc. We will use SPSS 27.0 software to construct internal consistency coefficients and use Cronbach's alpha to evaluate and calculate the internal consistency reliability. Based on the analysis results, we will conduct in-depth discussions on the quality of testing and propose suggestions on whether certain projects need to be redesigned to better meet teaching demands.

Descriptive Analysis

According to the paper, objective questions account for 83.33% of the total score, while subjective questions account for 16.67% of the total score. Wu Xiaoshan's (2007) research shows that in terms of scoring methods, due to the unique answers of objective items, machine scoring can be used to ensure the objectivity and consistency of test results. Therefore, the higher the percentage of objective items enjoy higher reliability.

Table 1

Descriptive Statistics of Grade 7 Midterm Test

	N Statistic	Minimum	Maximum	Mean	Std. Deviation
Total	44	15.5	116.0	84.318	28.0013

Referring to Table 2, a high standard deviation (28) and maximum value (116), while a low minimum value (15.5) indicate a significant gap in students' English proficiency and a serious polarization problem. According to scores, it can be seen that the most concentrated score range is 70-110, with the 90-110 score range having the highest number of people.

Reliability

Reliability is the measure of consistency, stability, and dependability of test results, typically quantified by internal consistency to assess the reliability of the test. The higher the reliability, the more reliable the test results will be. Cronbach's alpha is a method of measuring the reliability of a scale or test, which increases with the number of items in the survey or test (Richards, Ross, & Seedhouse, 2012).

Table 2

Reliability of the Test Paper

Reliability Statistics	
Cronbach's Alpha	N of Items
.947	7

With the help of SPSS 27.0, we obtained the reliability indicators of this test. According to Table 2, the overall reliability of the test paper is 0.947, within the range of excellent internal consistency. Therefore, this test has very high reliability.

Analysis of Cloze

As a typical representative of objective questions, Cloze seems to be an indispensable section in both large and small English exams. The Cloze mainly tests students' mastery of grammar and vocabulary, their understanding of context, and their analytical and discriminative skills. The diversity of options and question stems can easily confuse students, therefore, this study specifically explores Cloze as an objective item representation based on SPSS data analysis.

Table 3

D.I. and F.V. of Each Cloze Items

Item NO.	16	17	18	19	20	21	22	23
F.V.	0.84	0.84	0.73	0.77	0.80	0.68	0.84	0.09
D.I.	0.54	0.46	0.69	-0.23	0.54	0.69	0.54	0.08
Item NO.	24	25	26	27	28	29	30	
F.V.	0.73	0.82	0.57	0.68	0.89	0.66	0.86	
D.I.	0.54	0.54	0.62	-0.08	0.31	0.62	0.46	

Table 3 lists the corresponding Facility Value (i.e. F.V.) and Discrimination Index (i.e.D.I.) for each question in the cloze test. Normally, the range of F.V. should be between 0 and 1, and the higher the F.V., the simpler the question. But to ensure that the problem is neither too simple nor too difficult in difficulty, the ideal F.V. is controlled between 0.3 and 0.7. However, from the data in the table above, it can be seen that most of the questions are not within this ideal range, and some questions are even set too simply to effectively test students'

language abilities. The range of D.I. is between -1 and 1, and the ideal range is above 0.35, the higher, the better. According to Table 3, most of the questions are belong to this ideal range, some questions, such as questions 19, 23, 27 and 28 have lower D.I., so it is necessary to optimize these questions.

Conclusion

This study takes a specific exam as the research object and provides a detailed analysis of the structure of the test paper and test results. Generally speaking, the mid-term exam for seventh grade in Hangzhou is generally reliable. The Cronbach alpha of this test is 0.947, which is higher than 0.8, indicating a high level of internal consistency reliability. This test can accurately measure the candidate's English proficiency. And the author puts forward the following thoughts and suggestions on English teaching and testing at the school:

Firstly, the difficulty level of the test questions should be moderate. The mid-term exam belongs to the achievement test, which is not to determine the level of students, but to understand their mastery of language knowledge and skills, and to check whether the teaching effect has achieved the expected goals. Therefore, the difficulty of the test questions should be moderate. Teachers should design effective teaching methods for students at different levels, in order to improve the overall level of the class.

Secondly, in future teaching, we should strengthen the training of students' reading skills and the cultivation of critical thinking. At the same time, we should pay more attention to their writing learning situation and encourage them to accumulate more words and sentences.

Finally, as a teacher, one should master necessary language testing knowledge and basic research methods to ensure that each exam has substantive significance. Teachers not only need to pay attention to students' exam scores, but also collect various types of exam data, organize, calculate, analyze and interpret the data, in order to obtain new discoveries and valuable information from the exams, and further improve our teaching methods and academic exams.

References

- Bachman, L. R., & Palmer, A. S. (1999). *Language testing in practice*. Oxford: OUP.
- George, D., & Mallery, P. (2019). *IBM SPSS statistics 26 step by step: A simple guide and reference*. London: Routledge.
- Richards, K., Ross, S., & Seedhouse, P. (2012). *Research methods for applied language studies: An advanced resource book for students*. New York: Routledge.
- 李筱菊. (2002). *语言测试科学与艺术*. 长沙: 湖南教育出版社.
- 吴小珊. (2007). 论英语测试信度和效度的平衡关系. *广西民族大学学报 (哲学社会科学版)*.