Sino-US English Teaching, June 2019, Vol. 16, No. 6, 249-255
doi:10.17265/1539-8072/2019.06.004

A Reliability and Validity Analysis on Vocabulary and Structure of NEPCMSS Test in 2010

GAO Wencheng, WU Xiaohua
University of Shanghai for Science and Technology, Shanghai, China

Reliability and validity are two criteria to evaluate whether a test paper is successful. This paper analyzes the vocabulary and structure MCQ (Multiple Choice Question) of National English Proficiency Competition for Middle School Students (NEPCMSS) test in 2010, evaluating the reliability and content validity by using SPSS software. It is concluded that the 20 multiple choice questions are not difficult in terms of difficulty; the reliability is good but the content validity is not so satisfied.

Keywords: reliability and validity, SPSS, vocabulary and structure MCQ, NEPCMSS

Introduction

As an integral part of the teaching system, examinations are the main means of checking students' academic achievements in school. If examination is regarded as a measurement, the test paper is a tool of measurement. The reliability and validity of examination questions are related to the correct evaluation of the student's language level. How to evaluate the quality of a test paper scientifically is very important to the teachers and school administrators. An effective test paper analysis can objectively reflect the teaching level and effect, which can also help the teachers and students find weakness in their teaching activities, improving the quality of teaching and learning. Zou (2000) holds that reliability and validity are two important indicators used to measure the quality of examinations in terms of teaching. Through analyzing reliability and validity, this paper studies National English Proficiency Competition for Middle School Students (NEPCMSS) in 2010 (senior two group) in China, so as to give an objective and impartial evaluation to the the quality of the test paper, teaching work as well as the student's feedback.

Methodology

This material is selected from the vocabulary and grammar multiple-choice questions of National English Proficiency Competition for Middle School Students in 2010. The significance of this material is that the statistical results and data produced by the competition will provide reference and basis for all kinds of foreign language teaching and research projects all over the country, so as to check and evaluate the quality of English teaching in high schools, improve the quality of English teaching, and strengthen the student's English competence. On the basis of this, we need to cultivate the student's comprehensive ability to use English and raise English teaching to a new level. Therefore, the authors choose this test paper as the research material. The

GAO Wencheng, Ph.D., Professor, College of Foreign Languages, University of Shanghai for Science and Technology, Shanghai, China.
WU Xiaohua, Master, College of Foreign Languages, University of Shanghai for Science and Technology, Shanghai, China.

collection of data was conducted in a senior high school in Henan Province. One intact class of the second year students consists of 50 participants in the study. The test was administered in March 2018 during their regularly scheduled class. The authors designed 25 minutes to finish the vocabulary and structure multiple-choice questions. The research methods adopted in reliability analysis include internal consistency reliability, normal distribution, and descriptive analysis. The reliability and frequency analysis program of SPSS software was used for data analysis. As for the validity analysis, the authors will discuss the content validity in terms of national curriculum standard for reference. With the advent of the information age and the increasing frequency of international communication, English has become increasingly important as an international lingua franca. In order to enable high school students to better meet entrance, employment, and cultivate lifelong learning ability, the Ministry of Education has set the English Curriculum Standards of high school. The English Curriculum Standards for Senior High School comprehensively specify the nature, design ideas, curriculum objectives, content standards, teaching tasks, and learning methods, and put forward suggestions for the implementation of teaching and evaluation. Since the test paper was constructed in 2010, we need to test the validity according to the curriculum standard in 2010. Through the above analysis methods, we can evaluate the school teaching and improve the efficiency and level of teaching management. At the same time, it is of certain reference value to the teacher's self-evaluation.

Results

In this section we will analyze the reliability and validity of MCQ (Multiple Choice Question) of NEPCMSS in 2010 on the basis of testing results.

Reliability Analysis

Reliability analysis is made from the perspectives of facility value, mean and standard deviation of score, normal distribution, internal consistency reliability.

Facility value. The facility value is the data which can reflect the difficulty degree of the test questions. The greater the facility value is, the higher the score rate of the question is, and the less difficult it is. Usually, facility value can be understood as "easy degree". Its range is from 0 to 1, and 0.3-0.7 is an ideal value of difficulty. The FV value of Multiple Choice Question is shown in the following table:

Table 1

The Facility Value of 20 Items of MCQ

Item	N 1	N 2	N 3	N 4	N 5	N 6	N 7	N 8	N 9	N 10	N 11	N 12	N 13	N 14	N 15	N 16	N 17	N 18	N 19	N 20
No. of correct answers	42	29	36	42	43	30	36	38	35	32	20	24	30	40	40	43	24	41	40	37
No. of student	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50
Facility value	0.84	0.58	0.72	0.84	0.86	0.6	0.72	0.76	0.7	0.64	0.4	0.48	0.6	0.8	0.8	0.86	0.48	0.82	0.8	0.74

From Table 1, we find that six values of the 20 multiple-choice questions are located in the interval of ideal difficulty coefficient, which indicates that these six questions are of moderate difficulty. Among them, according to the principle that the higher the facility value is, the less the difficulty is, there is no question whose value is below 0.3, which shows that these 20 questions are not difficult as a whole.

Mean and standard deviation of score. Descriptive analysis by SPSS software generally includes dispersion and central tendency. The measures of central tendency include mean, mode, and median. The measures of dispersion include range and standard deviation. The bigger the range is, the bigger the standard deviation is.

Table 2

Mean and Standard Deviation of MCQ

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Total score	50	15.00	5.00	20.00	14.0400	3.42833
Valid N (listwise)	50					

From Table 2, it is found that the difference between the highest score and the lowest score of the 50 students in this class is 15, and the lowest score is 5, the highest score is 20. Its average score is 14.04. In addition, the standard deviation is 3.43. It is obvious that the values of the standard deviation and range are all very large, which indicates that the student's language level in the class is very varied and polarized. However, in terms of the average score of this testing item, most of the students in the class have good grades, and the students with poor grades are in the minority.

Normal distribution. Normal distribution is a theoretical hypothesis about how the scores should spread. The concept of normal distribution starts from people's observation of the nature. For instance, height of trees in a forest, intelligence of people, and human height all conform to normal distribution; the averages are in majorities, while the high or low extremes are in minorities. Since people's intelligence is normally distributed, their scores of language learning should also be normally distributed. Whether the scores obey normal distribution or not can be observed by a graph or by calculating the skewness and kurtosis values as shown by Table 3:

Table 3

Skewness and Kurtosis Value of 20 Multiple-Choice Questions

	N	Skewness		Kurtosis	
	Statistic	Statistic	Std. Error	Statistic	Std. Error
Total score	50	-0.482	0.337	-0.262	0.662
Valid N (listwise)	50				

When the skewness and kurtosis value are 0, they are completely normal. Positive and negative values of skewness suggest positively skewed and negatively skewed respectively. The positive and negative kurtosis indicate that the peak is "high and thin" (fractional concentration) and "short flat" (fractional dispersion). Through the values of Table 3, we can see that the kurtosis value is -0.26, whose fraction is over dispersed. The skewness is -0.48, which is negatively skewed distribution. It shows that more scores are above the mean and the test is easy. Nevertheless, as a rule of thumb, values for skewness and kurtosis of between -2 and +2 indicate a reasonable normal distribution. Therefore, we can draw a conclusion that the 20 multiple-choice scores are basically normal distributed.

Internal consistency reliability. Internal consistency reliability, also known as homogeneity reliability, refers to the consistency of all subjects within a test. The Cronbach's Alpha was used to test the internal consistency reliability of a test. The Cronbach's Alpha value is between 0 and 1. According to Gay (1996), the greater the alpha value is, the stronger the correlation between the test items is, and the higher the credibility of

internal consistency is. Generally speaking, if the Cronbach's Alpha is higher than 0.8, it indicates excellent internal consistency; if it is between 0.6 and 0.8, it is better, while if it is lower than 0.6 it indicates poor internal consistency. In practical applications, the Cronbach's Alpha value is at least higher than 0.5, preferably above 0.7.

Table 4a

Cronbach's Alpha Coefficient

Reliability statistics	
Cronbach's Alpha	N of items
0.704	20

Table 4b

Correlation and Cronbach's Alpha Coefficient

	Scale mean if item deleted	Scale variance if item deleted	Corrected item-total correlation	Cronbach's Alpha if item deleted
N1	13.1800	11.375	0.105	0.706
N2	13.4400	11.476	0.015	0.720
N3	13.3000	10.296	0.438	0.677
N4	13.1800	10.640	0.413	0.682
N5	13.1600	11.198	0.193	0.699
N6	13.4200	10.820	0.218	0.699
N7	13.3000	10.255	0.452	0.675
N8	13.2600	10.686	0.320	0.689
N9	13.3600	11.051	0.155	0.705
N10	13.3800	10.812	0.228	0.698
N11	13.6200	10.200	0.421	0.677
N12	13.5400	10.417	0.339	0.686
N13	13.4200	10.902	0.192	0.702
N14	13.2200	10.583	0.391	0.683
N15	13.2200	10.665	0.359	0.686
N16	13.1600	10.464	0.524	0.674
N17	13.5400	11.600	-0.023	0.724
N18	13.2000	10.857	0.300	0.691
N19	13.2200	11.481	0.048	0.712
N20	13.2600	10.074	0.553	0.666

There are five columns in the above Table 4(b). The first column are variables, which are the 20 multiple choice questions. The second and the third column are the average and variance of the rest items after the deletion of the item. The fourth column is the correlation coefficient between the item and the other items. The fifth column is the change of Cronbach's Alpha value after deletion of this item. The fifth column is very useful in evaluating the test questions of poor reliability. If the Cronbach's Alpha value becomes higher after deleting the item, the item influences the reliability of the test.

As shown in Table 4 (a), the Cronbach's Alpha value is 0.704, between 0.7 and 0.8, indicating that the 20 questions have good internal consistency. In addition, we can use the fifth column to add or delete questions to improve the reliability. Usually, we will delete the question whose Cronbach's Alpha value is higher than 0.704. After several rounds of deletion and screening, we get the final result.

Table 5a

Improved Cronbach's Alpha Coefficient

Reliability statistics	
Cronbach's Alpha	N of items
0.766	11

Table 5b

Improved Correlation and Cronbach's Alpha Coefficient

	Scale mean if item deleted	Scale variance if item deleted	Corrected item-total correlation	Cronbach's Alpha if item deleted
N3	7.2400	5.288	0.555	0.730
N4	7.1200	5.700	0.461	0.744
N7	7.2400	5.492	0.449	0.744
N8	7.2000	5.837	0.302	0.762
N11	7.5600	5.517	0.383	0.754
N12	7.4800	5.642	0.315	0.763
N14	7.1600	5.566	0.484	0.740
N15	7.1600	5.851	0.326	0.758
N16	7.1000	5.765	0.453	0.745
N18	7.1400	5.919	0.308	0.760
N20	7.2000	5.306	0.583	0.727

From the above Table 5(a), it is found that the reliability coefficient is raised from 0.704 to 0.766. After we deleted the N1, N2, N5, N6, N9, N10, N13, N17, and N19, the reliability coefficient of the remaining 11 questions is lower than that of the whole reliability 0.766, indicating that these questions do not affect the overall reliability. In this way, we can analyze the reliability of all the items and find out those that affect the whole reliability. Moreover, we can select excellent questions and build up high quality items bank.

Validity Analysis

Validity is one of the criteria for language test evaluation. In general, it consists of content validity, criteria validity, and construct validity. Content validity, as an aspect of validity study, refers to whether the exam outline stipulates the exam, or to what extent the test questions can represent the target to be measured. In this paper, the authors will discuss to what extent the MCQ is consistent with the curriculum standard.

According to the comparison and analysis of the consistency between the 20 multiple-choice questions and the curriculum standard, we obtain the following results:

From Table 6, we can see that there are 23 grammatical items put forward in the national curriculum standard. Among them, only nine grammatical items are included in the 20 multiple-choice questions; less than 50% of the prescribed grammatical items are tested. In addition, in the 20 questions, 16 of them are intended to examine grammatical rules and the rest are intended to test vocabulary. However, the grammatical points are not evenly covered, so the content validity of the 20 multiple choice questions is not so satisfied.

Table 6

Distribution of Grammatical and Lexical Items

Grammar item	Noun	Pronoun	Numeral	Adverb	Prep. & prep. phrase phrase	Conjunction	Adjective	Article	Verb	Tense	Non-finite verb	Passive voice
No.		1			3	2		1	1	2	1 + 1(5)	
Freq.		0.05			0.15	0.1		0.05	0.05	0.1	0.05+0.05	

Grammar item	Word-building	Types of sentence	Sentence constituent	Types of simple sentence	Emphasis	Compound complex sentence	Subordinate clause	Ellipsis	Inversion	Subjunctive mood	Total	Lexical item
No.		2					3				15 + 1	4 + 1(5)
Freq.		0.1					0.15				0.8	0.2 + 0.05

Note. The data “1 + 1(5)” and “4 + 1(5)” means Question 5 not only tests the grammar of non-finite verb, but also the vocabulary.

Discussion

Firstly, by analyzing the difficulty of 20 multiple-choice questions, we find that these 20 choices are not difficult. And the simple questions account for 60%. From this we can see that this set of questions is simple for the students in this class. From the descriptive analysis data, we can see that the students’ academic achievements are polarized. The teachers should pay more attention to this part of the students whose grades are not good, and find appropriate teaching methods to help them. From the kurtosis and skewness values, we can see that they are all negatively distributed, which fully shows that the vocabulary and structure MCQ of the test paper is simple, and most students can get good grades. In addition, we also analyze the reliability of the 20 questions through the internal consistency reliability. The results show that the reliability is good. In order to help the school teacher build a high quality test paper bank, we also selected 11 excellent questions for the school students through several rounds of deletion. Finally, according to the Table 6, we know that the 20 multiple-choice questions only cover nine grammatical items prescribed by the curriculum standard, which cannot fully reflect the requirements of the curriculum standard. Also, there are 16 multiple-choice questions intended to test grammar and only five questions intended to examine vocabulary. The proportion of testing contents is unbalanced. Therefore, the validity of the 20 multiple-choice questions is not high.

Conclusion

Through the analysis of the part MCQ of the test paper, based on the results of 50 students from a senior high school (senior grade two), it is concluded that the reliability and the facility value of MCQ are appropriate enough to achieve the goal of designing this competition, while the validity of MCQ is low.

According to the research results, the student’s language proficiency is polarized in the class. The teacher should pay more attention to those slow students and take appropriate teaching methods to change the unbalanced situation. Furthermore, the design of the choice questions can combine grammatical items with vocabulary, which will balance the proportion of the two types of testing points, which will improve the validity of the test paper. The findings will provide reference and basis for the improvement of English test papers construction and English teaching in high schools in China.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Chen, J. L. (2000). *Modern Chinese teaching organization and management*. Shanghai: Shanghai Foreign Language Education Press.

Gay, L. R. (1996). *Educational research: Competencies for analysis and application*. Englewood Cliffs, New Jersey: Merrill, Prentice Hall.

Gui, S. C. (1997). *Linguistic methodology*. Beijing: Foreign Language Teaching and Research Press.

Henning, G. A. (1987). *Guide to language testing*. USA: Newbury House Publishers.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Li, X. J. (1997). *Language testing science and art*. Changsha: Hunan Education Publishing House.

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Qin, Z. Q. (2005). *Quantitative analysis of academic examination results*. English Teaching and Research in Normal Universities.

Song, Z. Y. (1986). *Modern education measurement*. Beijing: Education Science Press.

Wang, H. L. (1987). *Educational measurement*. Kaifeng: Henan University Press.

Yang, H. Z. (1991). *Language testing and language teaching*. Beijing: Foreign Language Teaching and Research Press.

Zou, S. (2000). *A concise coursebook of English testing*. Beijing: Higher Education Press.