

A study on the Criterion-related Validity of Final Exam of College English

Haizhen Zhao

School of Foreign Studies, Guilin University of Electronic Technology, Guilin, China

Keywords: Criterion-related validity, Validation, Final exam, CET-4

Abstract: A quantitative method is adopted in the study to evaluate the criterion-related validity of final exam of English administered in Guilin University of Electronic Technology. Taking the CET-4 as a valid criterion, the author measured the normal distribution, difficulty, degree of differentiation and correlation between the final exam and CET-4. The result shows good criterion-related validity of the final exam. It indicates that the final exam is of good validity and can effectively evaluate students' English proficiency and therefore provide scientific basis for improving English teaching. At the same time, it is not sufficient to confirm the validity of a test only based on the correlation with CET-4. The construct validity and content validity of the final exam need to be further verified.

1. Introduction

Testing plays an imperative role in language learning and teaching. Language tests are used to assess learners' language ability, diagnose learners' learning situation and select learners for various purposes. In addition, as an important part of the teaching process, testing can provide rich information beneficial to teaching: if a test is designed scientifically and reasonably, it can provide accurate and objective description of the current situation of teaching, provide sufficient feedback information for teaching, help teachers improve language teaching, etc [1]. Therefore, it is rather important to analyse and evaluate the test paper after a test is taken. The process involves the collection, analysis and interpretation of the data from a test. Only in this way, can we evaluate a test properly and make a reasonable educational decision accordingly.

2. Validation and Criterion-related Validity

There are many ways to evaluate the quality of a test, among which validity is widely used. Kelly [2] defined validity as "the problem of validity is that of whether a test really measures what it purports to measure." The more consistent the measurement results are with the content to be examined, the higher the validity is; otherwise, the lower the validity is. With regard to a language test Weir [3] states that validity is "the extent to which a test can be shown to produce data, i.e., test scores, which are an accurate representation of a candidate's level of language knowledge or skills." That is to say, validity is not an abstract concept. It needs and can be tested. The validity of a test does not lie in what the test designers claim; rather, they need to produce evidence to support such claims starting from the design of a test. This process is called validation. Validation involves gathering evidence to support the inferences made from test scores. It is the inferences regarding specific uses of a test which are validated, not the test itself. In short, validation is a process or evaluation where evidence is generated to support the inferences made from test scores [4].

Three ways are mainly applied to measure the validity of a test, namely content validity, construct validity and criterion-related validity. In this study the author used criterion-related validity in particular to evaluate the validity of the final exam of English in a university. According to ALTE members [5]. "A test is said to have criterion-related validity if a relationship can be demonstrated between test scores and some external criterion which is believe to be a measure of the same ability." Criterion-related validity is applied to see how far results on a test agree with those provided by some independent and highly dependable assessment of the candidate's ability. This independent assessment is thus the criterion against which a test is validated. In this study, the

CET-4 is used to assess the validity of a final exam of college English. For one thing, the structure and content of the final exam is similar to that of CET-4. Listening comprehension, reading comprehension, translation and writing are included in both tests. So it is acceptable and reasonable to evaluate the correlation between the final exam and CET-4. For another thing, it was proved that CET-4 enjoys higher validity, reliability and stability which makes it an ideal criterion to evaluate the quality of English tests of the same type.

3. An Introduction to the Final Exam

Based on the classification of tests the final exam of English in Guilin University of Electronic University is an achievement test[6]. It is designed to evaluate participants' mastery of English in a semester in the aspects of listening, reading, translating and writing. Although a relatively small-scale school-based English test, the test scores are related to the Grade Point Average of students and taken as an important factor to get a bachelor's degree. Therefore it can be considered as a high-stakes test. The high risk of the test requires that each step, from the development of the test to its implementation, should follow strict quality control measures so as to ensure the high quality and publicity of it. So, it is imperative to evaluate the quality of the test and improve it correspondingly. Only in this way can the final exam effectively evaluate students' English level and at the same time better serve college English teaching and learning.

The test in this study is the final English exam of the first semester for non-English major freshmen. The test is designed by a group of college English teachers based on the syllabus of college English. The test has gone through a process of choices of questions, group discussion, modification and examination so as to ensure the quality of it. The structure, type, content and time of the exam are shown in the following table. Manual grading is adopted in subjective parts like writing and translation from English to Chinese. And automatic grading is used in objective parts like listening, reading and vocabulary.

Table 1 Structure and content of the final exam of English.

Test structure	Content	Question type	Scores	time
Listening comprehension	News report	Multiple choices	7	25 minutes
	Long conversations		8	
	Passages		10	
Reading comprehension	Banked cloze	Multiple choices	5	35 minutes
	In-depth reading		30	
Translation	E-C translation	Translation	15	15 minutes
Vocabulary	Multiple-choice Cloze	Multiple choices	15	15 minutes
Writing	Paragraph writing	writing	10	30 minutes

4. Methodology

The study analysed the normal distribution, difficulty, degree of differentiation and correlation between the final exam and CET-4 to evaluate the criterion-related validity of the final exam.

4.1 Participants

Convenience sampling is employed in the study. The author sampled 135 non-English major freshmen from the school of Mechanical and Electrical Engineering, the school of Information and Communication, and the school of Computer Science and Information Security in the Guilin University of Electronic Technology as the participants of this study. They are from level A classes of English with relatively higher English proficiency. They have been learning English for at least 6 years. Among them, 105 are male students and 30 are female students, ranging from 17 to 19 years old.

4.2 Data Collection

The scores of the participants' final exam of English were collected directly after the exam was finished. In order to avoid the interference of subjective factors, automatic grading is applied in objective questions of the exam (75% of the total). And the scores of the participants' CET-4 were collected after the results came out in February 2020.

4.3 Statistical Procedures

The study employed correlation analysis to evaluate the criterion-related validity of the final exam of English. The author used the software of SPSS to process the scores of the subject. Firstly, the normal distribution of the two tests is compared. Then the difficulty and differentiation of two tests are calculated according to the Brown's formula. Finally, the correlation of the two tests are analysed.

4.4 Results

4.4.1 Comparison of Normal Distribution between CET-4 and the Final Exam

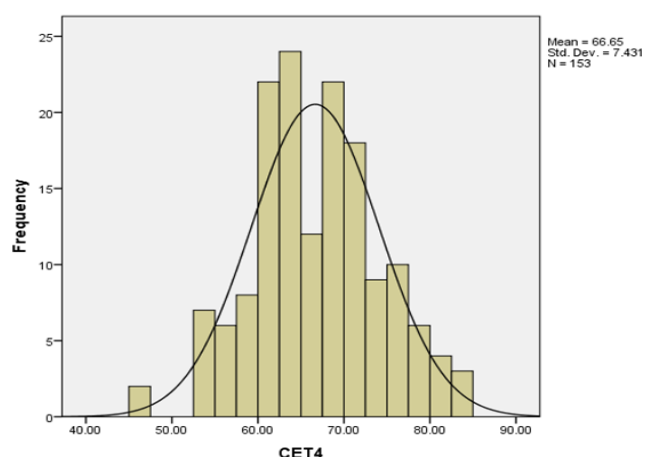


Figure 1 The frequency distribution of CET-4.

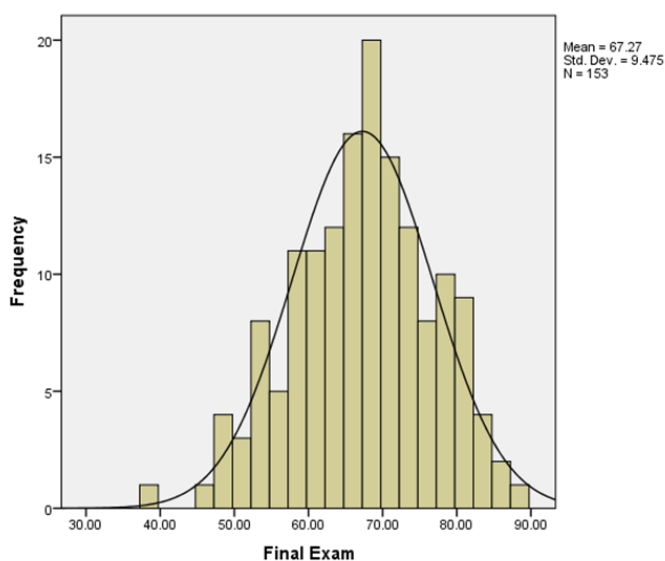


Figure 2 The frequency distribution of the final exam.

It can be seen from Figure 1-2 that the scores of the two tests basically show normal distribution. The standard deviation of the final exam is 9.475, while the standard deviation of CET-4 is 7.431, indicating that the data distribution of the final exam is more discrete than that of CET-4. In other

words, compared with CET-4, the final exam can better distinguish students of different levels.

Table 2 The skew and peak coefficient of both tests.

		CET-4	Final Exam
N	Valid	153	153
	Missing	0	0
Skewness		.082	-.219
Std. Error of Skewness		.196	.196
Kurtosis		-.025	-.180
Std. Error of Kurtosis		.390	.390

It can be seen from table 2 that the skew coefficient of CET4 is greater than zero, which means that the frequency distribution of CET4 test scores is in a normal skew form, with a small number of people higher than the average of 66.65 points. The skew coefficient of final exam is less than zero, which means that the frequency distribution of final exam scores is in a negative skew form, with a large number of people higher than the average of 67.27 points.

The peak coefficient of the two tests is less than zero, which indicates that the frequency distribution of the two exams is relatively low and broad, showing a negative peak. That is to say, the distribution of test-takers in the two exams does not reach the normal distributed concentration.

4.4.2 Comparison of Difficulty between CET-4 and the Final Exam

Difficulty means the difficulty degree of test questions. It is an important index parameter to measure the quality of test questions.

The full score of final exam is 100 points while the full score of CET-4 is 710 points. In order to compare the scores of them, the average score of CET-4 needs to be processed, that is, multiplied by the coefficient of 10/71. According to the Brown's difficulty formula: $P = \text{score}/\text{full mark}$, we can calculate the total difficulty of the two tests.

Table 3 Statistics.

		CET4	Final Exam
N	Valid	153	153
	Missing	0	0
Mean		66.6536	67.2712
Std. Deviation		7.43139	9.47521
Minimum		45.20	38.50
Maximum		84.10	88.50

According to the data in table 3, the average values of the final exam and CET-4 are 67.2712 and 66.6536, respectively. Therefore, the final exam's difficulty $P1=67.2712/100 \approx 0.673$, CET-4's difficulty $P2=66.6536/100 \approx 0.667$. Therefore, the difficulty of the final exam is slightly higher than that of CET-4.

4.4.3 Comparison of the Degree of Differentiation between CET-4 and the Final Exam

In addition to difficulty, another index to measure the quality of a test is the differentiation of it. Differentiation refers to the degree to which a test distinguishes candidates' abilities. In the measurement of differentiation, the total score of the test is usually taken as the actual level of the candidates, and the correlation coefficient between the score of the tested candidate and the total

score is taken as the differentiation of the test paper.

According to Brown's formula $D = P1$ (27% high grouping difficulty) - $p2$ (27% low grouping difficulty), we can calculate the differentiation between the final exam and CET-4. In the final exam, the average values of the 27% high and 27% low groups were 79 and 55, respectively. The average values of the 27 percent high and 27 percent low of CET-4 groups were 76 and 58, respectively. Therefore, the degree of differentiation in the final exam $D1 = 79/100 - 55/100 = 0.24$, while the degree of differentiation in CET-4 $D2 = 76/100 - 58/100 = 0.18$. The degree of differentiation in the final exam is significantly higher than that in CET-4. This is consistent with the results of the above analysis.

4.4.4 Correlation Analysis between CET-4 and the Final Exam

Criterion-related validity refers to the consistency between test results and criterion. The method of criterion-related validity is to analyse the correlation between the test results and the effective criteria. The larger the correlation coefficient is, the better the criterion validity is, and vice versa. It is generally considered that the correlation coefficient between 0.4 and 0.8 is ideal. In general, the correlation strength of variables can be determined by the following ranges: correlation coefficient of 0.8-1.0 represents extremely strong correlation, 0.6-0.8 represents strong correlation, 0.4-0.6 represents medium correlation, 0.2-0.4 represents weak correlation and 0-0.2 represents extremely weak correlation or no correlation [7].

Correlation analysis was conducted on the mean values of the two exams with SPSS (see table 4).

Table 4 Correlation between CET-4 and the final exam.

		CET-4	Final Exam
CET4	Pearson Correlation	1	.515
	Sig. (2-tailed)		.000
	N	153	153
Final Exam	Pearson Correlation	.515	1
	Sig. (2-tailed)	.000	
	N	153	153

**. Correlation is significant at the 0.01 level (2-tailed).

As can be seen from table 4, the correlation coefficient between the total score of the final exam and the total score of CET-4 is 0.515 (Sig.=0.000), indicating a significant positive correlation between the total scores of the two tests. This means the final exam is of higher validity and can test the English proficiency of students as expected.

5. Conclusion

The study evaluates the criterion-related validity of a final exam of English in a university. The findings are listed as follows: firstly, there is a high degree of consistency between the final exam and CET-4, which indicates that the final exam has a high criterion-related validity. That means that the quality of the final exam is relatively high and can effectively assess the English proficiency of participants. Secondly, the average difficulty of the final exam is slightly higher. The test needs to be improved and adapted to moderate difficulty.

However it is not sufficient enough to confirm the validity of a test only based on the correlation with CET-4. The construct validity and content validity of the final exam need to be further verified.

Acknowledgements

This work is supported by Research on the Backwash Effect of School-based English tests on English Teaching (Grant No. JGB201816), Guilin University of Electronic Technology education and teaching reform project.

This work is also supported by Research on Construction of Ecological System of English Smart Learning (Grant No. 2019KY0243), Guilin University of Electronic and Technology Teaching

Reform Project.

References

- [1] Yang, H. Z. (2015). Valid Testing, Effective Teaching, and Valid Test Use. *Journal of Foreign Languages*, no. 1, pp 2-26. (In Chinese)
- [2] Kelly, T. L. (1927). *Interpretation of Educational Measurements*. New York: New World Book Company.
- [3] Weir, C. J. (2005). *Language Testing and Validation*. London: Macmillan.
- [4] Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.
- [5] ALTE (Association of Language Testers in Europe) Members. (1998). *Multilingual Glossary of Language Testing Terms*. Cambridge: Cambridge University Press.
- [6] Liu, R. Q. and Han, B. C. (2000). *Language testing and Its Methods*. Beijing: Foreign Language Teaching and Research Press. (In Chinese)
- [7] Yang, D. H. (2004). *The Application of SPSS in Language Research*. Beijing: China Social Sciences Publishing House. (In Chinese)