ISSN: 2710-0170

A Quality Analysis of a College English Placement Test

Xiuzi Yang

University of Shanghai for Science and Technology, Shanghai, China

Abstract

This study systematically investigated the quality of a college English placement exam paper from the test theory perspective: reliability, validity, difficulty, and discrimination. The results emphasized the need to conduct scientific analyses of the scores of English tests to improve the quality and effectiveness of future teaching. These analyses are essential for enhancing the overall quality and effectiveness of future teaching practices in the field of college English. By identifying areas for improvement in the exam paper, educators can take targeted measures to refine their assessment tools, ultimately leading to more accurate evaluations and better-informed instructional strategies.

Keywords

Reliability, Validity, Difficulty Value, Discrimination Index.

1. Introduction

An exam is, by definition, a measurement tool designed to elicit specific samples of individual behavior (Bachman, 1990). Dafouz and Camacho-Miñano (2016) investigated whether or not university students improved their academic performance by analyzing the scores of final exams using SPSS software. Lee et al. (2014) explained in the book, "IBM SPSS for Intermediate Statistics", how SPSS software is used for educational research and its importance in the analysis of data. Amiri and Ghonsooly (2015) analyzed the relationship between students' anxiety about learning English and their performance in exams; through the correlation and t-tests using SPSS software, they found a significant relationship between the anxiety level and score in the exam.

This study drew on the methods of the aforementioned research, combined with the actual situation of the English exam, and utilized SPSS software to conduct a quality analysis of the exam scores, aiming to provide evidence for improving the quality of English teaching.

2. Theoretical Foundation

2.1. Reliability

Livingston et al. (2018) stated that reliability indicates the degree of consistency of test scores between different test situations, test versions, or raters. According to Rudner and Schafer (2001), there are several ways to estimate reliability, including test-retest reliability, split-half reliability, and internal consistency reliability. These methods use different statistical techniques to assess the consistency of test results and thus determine the level of reliability of the test tool (Rudner & Schafer, 2001). The level of reliability directly affects the interpretation and application of test results, so it is critical to ensure a high level of reliability for educational testing (Livingston, Carlson, and Bridgeman, 2018).

2.2. Validity

Validity represents "the extent to which evidence and theory support the interpretation of test scores" (Sireci, 2007). According to Shepard (2016), the validity of the test is not only the technical definition of the test design, but also the interpretation and use of the results. Validity

ISSN: 2710-0170

directly affects the acceptability and application of test results. Therefore, in any form of testing, especially educational testing, it always has to be high. By assessing validity, it is possible to ensure that the test results are scientific and reliable, thus providing strong support for educational decision-making (Chapelle & Lee, 2021).

2.3. Difficulty Value

Difficulty represents the test items' extent of challenging abilities that test candidates have. When analyzing difficulty, it is sure that test items will have a reasonable distribution. Therefore, their discrimination and effectiveness are enhanced for validity (Wright, 2007). The appropriate setting of difficulty enables teachers to more accurately reflect student learning levels with scientific evidence so as to guide teaching improvement. (Kubiszyn & Borich, 2024).

2.4. Discrimination Index

DI is used to indicate the discriminant ability of a test item. The higher the value of DI, the more discriminative the item. According to Burton (2001), discrimination is the difference in test item performance between different groups of test takers. By analyzing the discriminants, we can ensure a reasonable distribution of test items, thereby improving the validity and reliability of the test (Olutola, 2015).

3. Research Process

3.1. Subjects

The subjects of this test are students from a certain university in China. This study selected the test scores of ten classes for analysis, with a total of 300 students and 300 valid test papers.

3.2. Methods

This set of test papers was graded based on pre-established grading criteria and standard answers. The scores of the 300 test papers were encoded in an Excel spreadsheet by total score and scores of each sub-question. Then, the data was analyzed statistically with SPSS 25.0 statistical software.

4. Specific Analysis of the Test Paper Based on SPSS

4.1. Analysis of Test Scores

The full score of the test paper is 100 points, and the basic score distribution is shown in the table below:

Table 1 Statistics

		Total
N	Valid	300
	Missing	0
Mean		47.513
Std. Error of Mean		.8162
Median		47.000
Mode		59.0
Std. Deviation		14.1370
Skewness		-.142
Std. Error of Skewness		.141
Kurtosis		-.786

ISSN: 2710-0170

Std. Error of Kurtosis	.281
Range	62.0
Minimum	19.0
Maximum	81.0

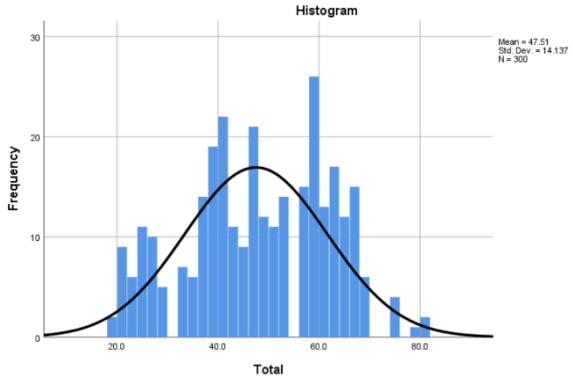

Figure 1 Normal distribution

Using SPSS 25 software, the results are shown in Table 4.1. The Skewness, Std. Error of Skewness, Kurtosis, and Std. Error of Kurtosis are all between -2 and 2, indicating that the total scores of the 300 students are normally distributed (Bachman, 2004). As shown in Picture 4.1, the number of students scoring between 75 and 85 is the lowest.

The mean for this class is 47.513 and the median is 47, which indicates that the distribution of scores is roughly symmetrical. A mode is the value that occurs most frequently in a set of data but does not meet the basic requirements of a central trend: accuracy and stability (Zou Shen, 2011). According to the statistical analysis of the SPSS 25 software, the test has a pattern of 59.

Above discussion depicts central tendency to describe a set of scores. However, the exact overall picture of scores cannot be depicted only with central tendency. Hence, this study has taken two important variability of scores that have been computed to examine dispersion: range and standard deviation. Range simply depicts the difference between the highest score and lowest score in a set of scores. In this test, the highest score is 81, and the lowest score is 19. The range of this test is 62, which reflects that it has a relatively large dispersion. The standard deviation of this test paper is 14.1370.

4.2. Analysis of Test Paper Quality

4.2.1. Reliability

Table 2 Reliability Statistics

Cronbach's Alpha	N of Items
.891	100

Reliability expresses the dependability, stability, and consistency of test results. The higher the reliability is, the more dependable the test is. Through reliability analysis, it is found that the reliability coefficient of this exam is 0.891, as shown in Table 4.2. The data from this exam indicated that it has good reliability.

4.2.2. Validity

The validity of a test refers to its effectiveness, that is, whether the test paper assesses the content it was originally intended to assess. Common methods for validating the construct

ISSN: 2710-0170

validity of second-generation system tests include correlation matrix analysis and factor analysis. Correlation matrix analysis involves calculating and arranging the correlation coefficients between the total score of the test and the scores of each major section into a matrix. The information revealed by this matrix is beneficial for studying the validity of the test paper. The value of the correlation coefficient ranges between -1 and 1, with 1 indicating a perfect positive correlation and -1 indicating a perfect negative correlation.

Table 3 Correlations

		Listening	Grammar	Total
Listening	Pearson Correlation	1	.628**	.830**
	Sig. (2-tailed)		.000	.000
	N	300	300	300

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

Table 4. 4 presents Pearson correlation coefficients among Listening, Grammar, and Total scores for a sample of 300 students. The Pearson correlation coefficient Listening and Grammar is 0.628, showing that there is a strong positive linear relationship between the listening and grammar scores. This correlation is statistically significant at the 0.01 level (two-tailed), which shows that higher listening scores are associated with higher grammar scores among the students. The Pearson correlation coefficient between Listening and Total score is 0.830, indicating a very strong positive linear relationship between the listening scores and total scores. This is significant at the 0.01 level-two-tailed-correlation, showing that performance on the listening part contributes a great deal to the general performance of the test.

4.2.3. Difficulty value

Difficulty is an indicator that measures the ease or difficulty of test items and the test paper. In fact, difficulty represents a measure of ease, which is exactly the opposite of the actual difficulty level of the test items (Zou Shen, 2011). The difficulty index is usually calculated using the following formula: FV = C(correct) / T(total), where FV is the difficulty value, C is the number of students who answered the item correctly, and T is the total number of students. The results are shown as follows:

Table 4 Facility Value

Item	FV
Item1	.77
Item2	.43
Item3	.48
Item4	.69
Item5	.65
Item6	.62
Item7	.62
Item8	.85
Item9	.72
Item10	.65
Item11	.68
Item12	.61
Item13	.43
Item14	.59
Item15	.61

ISSN: 2710-0170

Item16	.58
Item17	.46
Item18	.38
Item19	.41
Item20	.25
Item21	.69
Item22	.51
Item23	.65
Item24	.44
Item25	.54
Item26	.66
Item27	.67
Item28	.51
Item29	.46
Item30	.56
Item31	.71
Item32	.49
Item33	.53
Item34	.53
Item35	.52
Item36	.53
Item37	.59
Item38	.40
Item39	.47
Item40	.34
Item41	.37
Item42	.48
Item43	.33
Item44	.35
Item45	.48
Item46	.32
Item47	.41
Item48	.47
Item49	.24
Item50	.31
Item51	.37
Item52	.48
Item53	.47
Item54	.36
Item55	.45
Item56	.38
Item57	.55
Item58	.61
Item59	.27

ISSN: 2710-0170

Item60	.39
Item61	.36
Item62	.45
Item63	.49
Item64	.41
Item65	.31
Item66	.45
Item67	.39
Item68	.48
Item69	.31
Item70	.39
Item71	.54
Item72	.40
Item73	.51
Item74	.38
Item75	.46
Item76	.49
Item77	.42
Item78	.34
Item79	.49
Item80	.46
Item81	.38
Item82	.36
Item83	.44
Item84	.36
Item85	.42
Item86	.43
Item87	.49
Item88	.38
Item89	.51
Item90	.37
Item91	.50
Item92	.43
Item93	.52
Item94	.47
Item95	.32
Item96	.49
Item97	.51
Item98	.35
Item99	.28
Item100	.69

The difficulty index falling within the range of about 0.5 has a good level of difficulty. An index higher than 0.8 indicates that the test is too easy, while less than 0.3 reflects that the test is too

ISSN: 2710-0170

difficult. The difficulty index analysis reflects that the difficulty indices of the test items in this paper are mainly concentrated between 0.3 and 0.60, showing that the difficulty level of this test paper is moderate, and its quality is relatively good.

4.2.4. Discrimination

Table 5 Discrimination Index

	DI
Item1	0.36
Item2	0.32
Item3	0.55
Item4	0.41
Item5	0.47
Item6	0.34
Item7	0.35
Item8	0.24
Item9	0.4
Item10	0.33
Item11	0.39
Item12	0.35
Item13	0.27
Item14	0.19
Item15	0.27
Item16	0.4
Item17	0.06
Item18	0.29
Item19	0.26
Item20	0.16
Item21	0.37
Item22	0.32
Item23	0.45
Item24	0.38
Item25	0.32
Item26	0.29
Item27	0.33
Item28	0.17
Item29	0.01
Item30	0.35
Item31	0.39
Item32	0.26
Item33	0.3
Item34	0.29
Item35	0.35
Item36	0.13
Item37	0.36
Item38	0.42

ISSN: 2710-0170

Item39	0.41
Item40	0.31
Item41	0.33
Item42	0.45
Item43	0.26
Item44	0.16
Item45	0.49
Item46	0.27
Item47	0.41
Item48	0.38
Item49	0.13
Item50	0.1
Item51	0.26
Item52	0.21
Item53	0.47
Item54	0.24
Item55	0.2
Item56	0.22
Item57	0.33
Item58	0.36
Item59	0.21
Item60	0.3
Item61	0.31
Item62	0.35
Item63	0.16
Item64	0.34
Item65	0.26
Item66	0.24
Item67	0.25
Item68	0.3
Item69	0.17
Item70	0.2
Item71	0.33
Item72	0.37
Item73	0.43
Item74	0.37
Item75	0.39
Item76	0.24
Item77	0.34
Item78	0.35
Item79	0.36
Item80	0.26
Item81	0.19
Item82	0.34

ISSN: 2710-0170

Item83	0.35
Item84	0.13
Item85	0.31
Item86	0.26
Item87	0.22
Item88	0.21
Item89	0.52
Item90	0.5
Item91	0.32
Item92	0.51
Item93	0.41
Item94	0.32
Item95	0.18
Item96	0.39
Item97	0.51
Item98	0.29
Item99	0.33
Item100	0.13

It is generally believed that the discrimination index should be greater than 0.3. From the data results, it can be concluded that most of the test items have relatively good discrimination, effectively distinguishing students of varying proficiency levels.

5. Conclusion

By using SPSS 25 software to analyze the English test for students from dispersion to aggregation, it is concluded that the reliability and validity of the test are relatively reliable. In summary, it is a relatively reasonable test paper. This study also hoped that this analysis can provide scientific data and references for English teachers, enabling them to make targeted improvements in their teaching methods in future teaching work, thereby guiding students to face each exam more effectively.

The College English placement test plays a significant role in teaching, especially with analysis of language proficiency and skills for further improvement. Therefore, it works effectively for the teachers and their students to examine performances at either the individual or group level. Through this process of analyzing test results, teachers will be able to adjust their teaching strategies according to the different needs of students so that teaching becomes efficient and effective. In one word, it will provide a correct picture of their academic level.

References

[1] Amiri, M., & Ghonsooly, B. (2015). The relationship between English learning anxiety and the students' achievement on examinations. Journal of Language Teaching and Research, 6(4), 855.

[2] Bachman, L. F. (1990). Fundamental considerations in language testing. Oxford university press.

[3] Bachman, L. F. (2004). Statistical analyses for language assessment book. Cambridge University Press.

[4] Burton, R. F. (2001). Do item-discrimination indices really help us to improve our tests?. Assessment & Evaluation in Higher Education, 26(3), 213-220.

ISSN: 2710-0170

[5] Dafouz, E., & Camacho-Miñano, M. M. (2016). Exploring the impact of English-medium instruction on university student academic achievement: The case of accounting. English for Specific Purposes, 44, 57-67.

[6] Kubiszyn, T., & Borich, G. D. (2024). Educational testing and measurement. John Wiley & Sons.

[7] Livingston, S. A., Carlson, J., & Bridgeman, B. (2018). Test reliability-basic concepts. Research Memorandum No. RM-18-01). Princeton, NJ: Educational Testing Service, 8.

[8] Leech, N. L., Barrett, K. C., & Morgan, G. A. (2014). IBM SPSS for intermediate statistics: Use and interpretation. Routledge.

[9] Olutola, A. T. (2015). Item difficulty and discrimination indices of multiple choice biology tests. Liceo Journal of Higher Education Research, 11(1).

[10] Rudner, L. M., & Schafer, W. D. (2001). Reliability. ERIC Digest.

[11] Sireci, S. G. (2007). On validity theory and test validation. Educational Researcher, 36(8), 477-481.

[12] Shepard, L. A. (2016). Evaluating test validity: Reprise and progress. Assessment in Education: Principles, Policy & Practice, 23(2), 268-280.

[13] Wright, R. J. (2007). Educational assessment: Tests and measurements in the age of accountability. Sage Publications.

[14] Zou Shen. (2011). A Concise Course on English Language Testing. 3rd Edition. Higher Education Press.