

Reliability and Validity Analysis of Test Paper Quality Based on True Score Theory -- A Case Study of Linear Algebra

Lili Wang

College of Science, Nanchang Institute of Technology, Nanchang, China

Abstract. In this paper, two undergraduate majors were randomly selected as samples from the final examination of Linear Algebra in Nanchang Engineering College, and their final examination papers of Linear Algebra were used as the research object to analyse the reliability and validity of the papers respectively. Firstly, the reliability of the test papers was calculated to be 0.731 by the alpha coefficient method, and the degree of influence of each test question on the reliability of the test papers was analysed. Secondly, the content validity of the test papers is qualitatively analysed by using the two-way breakdown table of the examination content and competence requirements, with the average of the midterm grade and the usual grade as the validity score, and the correlation validity of the test papers obtained by using the correlation method is 0.669. Therefore, the reliability and validity of the final paper of "Linear Algebra" meet the requirements of the two undergraduate majors for the professional certification of engineering education. The test results truly and objectively reflect the actual level of the candidates, and the measurement results can reflect the intended purpose of the test.

Keywords: True score theory; Reliability; Validity covered; test papers quality; Linear Algebra.

1. Introduction

Nowadays, colleges and universities are centred on teaching, and the main task of teaching is to take systematic scientific and cultural knowledge as the carrier, to cultivate students' abilities and qualities in various aspects, so that each student can get different development [1]. An important way to evaluate the quality of teaching is to assess the performance of students, and the key to effectively reflecting the quality of teaching performance assessment lies in the preparation of test questions, that is, the quality of the examination paper [2,3].

To assess the quality of a test papers, it is necessary to consider whether the test results objectively and accurately reflect the real ability of the candidates, but also to consider whether the test results can achieve the expected test objectives; and to consider whether the degree of difficulty of the test questions and the actual level of the test subjects match, and whether it is possible to identify candidates of different levels of ability. Whether or not a test papers has achieved its intended purpose can only be known after a post-examination analysis of the quality of the test papers. Often the results of an examination are in the form of data. By collating and analysing these directly obtained data, valuable information can be extracted to enable a scientific evaluation of the quality of the examination.

However, in some colleges and universities, after the midterm and final examination papers are corrected, the total score is calculated and a table of total scores is finally obtained, and from then on the papers are put on the shelf, and not enough attention is paid to the work of statistical analysis of the examination results. From the preparation of test questions to the design and layout of the test papers, to the implementation of the test, marking, a test papers cohesion of the wisdom of the proposer. The whole process should be more informative than just getting a table of ranked scores. Some schools also analyse the test data after the test, but the analysis is basic and simple, and does not fully interpret and use the useful information, not to mention the deep analysis and use of test data resources. Based on the test results, we can analyse the quality of the test papers and the knowledge and ability level of the candidates, find out the problems and deficiencies in the preparation of the test papers, the test work and the teaching work of the teachers from the collection and analysis of the test information, and find out the deviation between the current situation and the expected situation, so as to inspire the learning and teaching activities in the future,

and then formulate the corresponding remedial measures. In-depth and thorough analysis of test papers is undoubtedly beneficial to the improvement of teachers' problem-solving and proposing abilities, as well as to their teaching and professional development. In addition, as a form of evaluation, the use of test data to carry out evaluation work is an effective combination of achieving changes in educational testing and promoting educational evaluation [4,5].

The main purpose of the quality analysis of examination papers is to make accurate judgements on the problems of examination papers and give appropriate feedback. Through the use of examination data, feedback and exchange of examination information, students can understand their own knowledge mastery and their own performance level in the examination, and make corresponding compensatory learning plans for their own course learning deficiencies; it can also provide teachers with more feedback information for teaching, so as to guide the teaching.

2. The True Score Theory

Classical test theory, also known as classical true score measurement theory, consists of a set of assumptions [6]. According to the theory, the test score obtained by the test taker, that is, the real score (observed score), is composed of the real score (true score) plus error score (error score). The standard error of measurement (standard error of measurement) is the index mark of the degree to which the real score is affected by the error score [7,8]. Because the purpose of a test is to obtain a reliable true score that is as close as possible to the true score, much of the effort that goes into the construction of the test is aimed at improving the reliability of the test.

Design and scientific papers to be measuring theory guidance. The index to measure the scientific design of test papers is the reliability and validity of the test papers, that is, the degree of reliability and validity of the test papers as a measurement tool. The test content structure, question type structure, question size structure, score structure, difficulty structure and other factors will affect the reliability and validity of the test papers [9]. Therefore, in order to ensure the reliability and validity of the test papers, the question type of the test papers should be allocated reasonably, the content covers the examination objectives and the prescribed knowledge points and ability points, there is a reasonable difficulty level distribution and discrimination level distribution, the question size is moderate, the score distribution of each question and each part is reasonable, etc.

2.1 Credibility Theory

Test reliability refers to a group of parallel test for the observation points from the same object for measuring the consistency and reliability. Here, the reliability of the test results is the reliability of the test. It reflects the ability to control the error [11,12].

Reliability in the true score theory is generally expressed as the correlation coefficient between the results of two or more tests. According to different calculation methods, reliability can be divided into test-retest reliability, duplicate reliability and internal consistency reliability. Test-retest reliability and duplicate reliability both need to be tested twice. However, in practice, it is difficult to use the same test or duplicate test to test the candidates independently. When it is impossible to measure the test repeatedly and there is no duplicate, two methods of internal consistency reliability are usually used, the split-half method and the α coefficient method. The calculation formulas and mathematical significance of credibility theory are detailed in References [11,12].

2.2 Validity Theory

The reliability of the test indicates the stability and reliability of the test results [12,13]. The high reliability of a test papers does not necessarily indicate that it is a good test papers. For example, if a balance measures the mass of an object multiple times, the difference in the results will be small, but if the balance itself has errors, then the measurement is obviously invalid. Therefore, when assessing the quality of a test papers, there is another important evaluation index: test validity. Test validity is the primary index for selecting and evaluating the quality of educational tests. What's

more the Validity is a quality index that measures the validity of test results, and it is an estimate of whether the object to be measured by a test is measured or to what extent. Validity can be divided into the content validity, structure validity and criterion related validity. For content validity, there is no directly available formula to calculate content validity. The key to evaluate the content validity of a test is to understand how well the test content represents the expected test content. Generally, each question in the test papers is required to have a clear and effective examination goal, which should not only complement other questions in the test papers, but also have low correlation with other questions in the test papers. General can prepare the exam content knowledge and capacity requirements of two-way breakdown to test content validity.

For evaluating criterion related validity, mainly using the Pearson correlation method to calculate the correlation coefficient. Since the criterion should be objective, reliable, practical and effective, the commonly used criterion includes academic achievement, actual learning performance and other ready-made valid tests, etc. The criterion validity is estimated by calculating the correlation coefficient between the test score and the criterion score, and the calculation formula is the same as Formula 1. But it is worth recalling that X represents the test score of the examinee, and Y represents the criterion score of the examinee. The calculation formulas and mathematical significance of and validity theory are detailed in references [12-14].

3. Material

In order to ensure the representability of the research objects, this study randomly selected the exam-takers of two undergraduate majors (agricultural and hydraulic engineering, automation) from the sopsopes of grade 2017 in Nanchang Institute of Engineering as the sample, and took their final exam papers of Linear Algebra as the research object, a total of 136 papers. The test content of papers is the whole book “Linear Algebra” (The Third Edition) written by the School of Mathematical Science of Yangzhou University. The test papers is composed of four parts with a full score of 100. Among them, fill in the blank (1-9), 3 points for each blank; Multiple choice (1-5), 3 points for each question; Calculation questions (1-5), with 7, 8, 10 and 14 points in turn, for a total of 22 questions/blank.

According to the requirements of teaching and assessment, the total score of Linear Algebra course will take the final score, the mid-term score and the regular score into consideration. In this paper, while analyzing the final exam papers of Linear Algebra, we not only collected the final exam data of Linear Algebra, but also collected the midterm exam data and the qualitative data of the completion of the Linear Algebra homework of these students in a semester. By collecting the scores of each question and the total score of the examinees, we can obtain their midterm and final examination data.

Since the true score model assumes that the true value of the examinees is a constant, that is, the real ability level of the examinees in a certain period is constant, which is in line with the actual situation, the final test data meet the theoretical hypothesis. After obtaining the final test data of Linear Algebra, this chapter will use SPSS and Excel software to analyze the reliability, validity, difficulty and discrimination of the test papers one by one according to the theory and technology of the previous chapter, and analyze and explain the difficulty and discrimination of various types of questions, as well as the number of questions and score distribution of each difficulty section and discrimination section.

4. Results

4.1 Reliability Analysis of the Test Papers

The data of the final exam of Linear Algebra were input into SPSS software, and the reliability of test papers was calculated by the split-half method and the Cronbach α reliability coefficient method respectively. The calculated results are as follows:

Table 2. The reliability of test papers is calculated by half method

Cronbach's Alpha	Section 1	value	0.541
		Number of terms	8 ^a
	Section 2	value	0.522
		Number of terms	8 ^b
	Total number of terms		16
Correlation between tables			0.702
Spearman-Brown coefficient	Equal length		0.822
	Unequal length		0.822
Guttman Split-Half coefficient			0.809

a. These items are 1,3,5,7,9,11,13,15. b. These items are: 2,4,6,8,10,12,14,16.

As can be seen from Table 2, the reliability of the whole test papers is 0.822 after dividing the test papers in half according to the parity of question number with the Spearman-Brown formula.

Table 3. Reliability of test papers calculated by α coefficient method

Cronbach's Alpha	Cronbachs Alpha based on standardized items	Number of terms
0.731	0.757	16

Table 4. Reliability analysis of each test item

	The entry has been deleted Degree mean	The entry has been reversed Degree variance	Corrected item total Relevance	Multiple correlations are flat Square	Item has been deleted Cronbach's Alpha value
1	55.90	408.211	0.231	0.211	0.735
2	57.69	331.478	0.341	0.376	0.747
3	55.90	329.874	0.428	0.396	0.728
4	59.37	346.179	0.031	0.178	0.747
5	56.82	333.303	0.254	0.269	0.736
6	56.88	340.948	0.100	0.264	0.709
7	56.94	331.521	0.347	0.210	0.720
8	57.37	323.408	0.486	0.407	0.717
9	55.79	350.670	0.398	0.138	0.736
10	58.77	323.468	0.444	0.296	0.719
11	51.16	322.701	0.352	0.270	0.713
12	51.31	262.896	0.412	0.349	0.721
13	51.26	215.012	0.695	0.603	0.657
14	50.74	229.124	0.632	0.442	0.687
15	54.75	311.073	0.381	0.297	0.712
16	55.87	318.639	0.302	0.194	0.723

As can be seen from Table 3, the Cronbach α reliability coefficient value is 0.731, indicating that the reliability of the whole test papers is high. From the last column in Table 4, it can be found that the reliability of test papers increases after the deletion of question 4, which indicates that the reliability of test papers can be improved after the deletion of this question. The reliability coefficient of question 13 is significantly smaller after deletion, which indicates that question 13 has a greater impact on the reliability of test papers.

4.2 Validity Analysis of the Test Papers

According to the standards and teaching requirements of course "Linear Algebra" of Nanchang Engineering College, it is a public university mathematics course with matrix theory as a tool, linear equations as the main line, matrices, determinants, n-dimensional vector spaces, linear systems of equations, eigenvalues and eigenvectors, quadratic forms as the basic content, and the course is relatively high in the degree of abstraction and logic. Through the study of this course, students can

cultivate their logical reasoning and abstract thinking ability, computer science computing ability, as well as the ability to apply algebraic methods to analyse and solve practical problems, laying a good foundation for students to further study the subsequent courses and engage in related work and practice.

The final examination of this course covers chapters 1 to 6 of the textbook, the knowledge and skills involved in the test papers cover the whole book, the sampling of test questions is highly representative, the knowledge and skills of test are presented in various forms, and the knowledge and skills of test are examined in a variety of ways. Each test question has a clear test objective, unique test function and the test function of test questions to complement each other, the important knowledge points of each chapter are examined, some of the test questions are comprehensive and can take into account the different learning levels of candidates, the distribution of marks basically in line with the allocation of teaching hours, in general, in line with the objectives of course and teaching requirements. Overall, the content validity of test papers is high.

Since the midterm grades and regular grades represent the candidates' learning status of course, this paper takes the average of midterm grades and regular grades as the validity score, which meets the characteristics of reliable and valid valid validity standard, objective standard, simple and practical. The correlation method is used to calculate the correlation coefficient between the final grade and the validity mark:

Table 5. Correlation method to calculate criterion association validity

		Final grade	Criterion score
Final grade	Pearson correlation	1	0.672
	significance (bilateral)		.000
	<i>N</i>	111	111
Criterion score	Pearson correlation	0.672	1
	significance (bilateral)	.000	
	<i>N</i>	111	111

**. Significant association at 0.01 level (bilateral).

It can be seen from Table 5 that the final scores are significantly correlated with the criterion scores, and the criterion correlation validity of final paper is $0.672 > 0.45$, indicating that the papers are valid and can achieve the expected purpose of the examination.

5. Conclusion

In this paper, the final examination papers of Linear Algebra of students majoring in agricultural and hydraulic engineering and automation in Nanchang Institute of Technology were used as the research objects. Based on the student score data, the paper quality was analyzed based on the true score theory, and the quantitative analysis and qualitative analysis were used to analyze the measurement indexes of paper quality, including the reliability and validity, so as to obtain the overall evaluation of the paper quality. The results show that the reliability and validity of the final examination paper meet the requirements. The test results truly and objectively reflect the actual level of the examinees, and the measurement results can reflect the expected purpose of the test, which meets the certification requirements of the two undergraduate engineering education majors.

In the era of big data, the paradigm of test papers evaluation based on experience and fixed thinking needs to be changed. Examination evaluation based on educational data is a frontier topic of examination evaluation. This paper takes the test paper quality analysis as the research perspective and uses the method of true score theory to analyze the sample test data. Although some meaningful exploration has been carried out, the data analysis method used is relatively normal. It is necessary to use more advanced models and methods for processing educational data in the future.

Acknowledgements

This work was supported by the program of school-level educational reform (2019JG043) and graduate education and teaching reform (NGYJG-2023-002) of Nanchang Institute of Technology.

References

- [1] Miller MB. Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling, 1995, 2(3):255-273.
- [2] Settlage J, Southerland, S. Teaching science to every child: Using culture as a starting point. New York, 2012. Routledge.
- [3] Pigott TD, Polanin JR. Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research*, 2020, 90(1):24-46.
- [4] Bradley RA. Some statistical methods in taste testing and quality evaluation. *Biometrics*, 1953, 9(1): 22-38.
- [5] Heyneman SP, Ransom AW. Using examinations and testing to improve educational quality. *Educational Policy*, 1990, 4(3):177-192.
- [6] Swart AJ. Evaluation of final examination papers in engineering: A case study using Bloom's Taxonomy. *IEEE Transactions on Education*, 2009, 53(2):257-264.
- [7] DeVellis RF. Classical test theory. *Medical care*, 2006, 44(11):S50-S59.
- [8] Brossman BG, Lee WC. Observed score and true score equating procedures for multidimensional item response theory. *Applied Psychological Measurement*, 2013, 37(6):460-481.
- [9] Raykov T, Dimitrov, DM, Marcoulides GA, et al. On true score evaluation using item response theory modeling. *Educational and Psychological Measurement*, 2019, 79(4):796-807.
- [10] Öz H, Özturan T. Computer-based and paper-based testing: Does the test administration mode influence the reliability and validity of achievement tests?. *Journal of Language and Linguistic Studies*, 2018, 14(1), 67-85.
- [11] Liu B. A survey of credibility theory. *Fuzzy optimization and decision making*, 2006, 5:387-408.
- [12] Norberg R. Credibility theory. *Encyclopedia of Actuarial Science*, 2004, 1:398-406.
- [13] Sireci SG. On validity theory and test validation. *Educational Researcher*, 2007, 36(8):477-481.
- [14] Moss PA. The role of consequences in validity theory. *Educational measurement: Issues and practice*, 1998, 17(2):6-12.