Scientific Research Publishing

Creative Education, 2025, 16(8), 1150-1161
https://www.scirp.org/journal/ce
ISSN Online: 2151-4771
ISSN Print: 2151-4755

Design and Analysis of Mathematics Test Paper

Xiaomin An

School of Teacher Education, Zunyi Normal University, Zunyi, China
Email: 756566782@qq.com

How to cite this paper: An, X. M. (2025). Design and Analysis of Mathematics Test Paper. *Creative Education, 16,* 1150-1161.
https://doi.org/10.4236/ce.2025.168071

Received: July 8, 2025
Accepted: August 16, 2025
Published: August 19, 2025

Abstract

Assessment is an important process of improving teaching and learning quality. In this essay, the table of specifications (TOS) was designed according to the mathematics textbook of Grade 9. A math test paper was designed according to the TOS. 40 Year 9 students were selected to complete this math test paper, and classical test theory (SPSS) and item response theory (Winsteps) were used to analyze the test data. The analysis is conducted from several dimensions, such as item analysis, person analysis, and person-item map. This essay aims to identify problems in the teaching and learning process of the subject of mathematics through the design and analysis of the test paper. Then, optimize the design of the mathematics test papers to improve the teaching and learning quality.

Keywords

Assessment, Classical Test Theory, Item Response Theory, Mathematics Test Paper, Teaching and Learning

1. Introduction

Assessment is a process of improving quality, and using course assignments such as test papers to examine students' performance is an important embedded assessment strategy (Rockman, 2002). This essay selects the mathematics textbook Chapter 1, "Quadratic equations" and Chapter 2, "Quadratic Functions", Unit 1, Volume 1, Grade 9, People's Education Press, as the main topic. Because quadratic equations and quadratic functions are effective mathematical models for portraying certain quantitative relationships in the real world, they are important for solving real-life problems. Thus, it is necessary to teach and assess students' mastery of the knowledge in these two chapters. In this essay, the table of specifications (TOS) was designed according to the text content, and the content was divided into different parts with various learning activities based on the six cognitive domains of learning objectives and the learning difficulties. Subsequently, a math test paper was designed according to the TOS. 40 Year 9 students were selected to complete this math test paper, and classical test theory (SPSS) and item response theory (Winsteps) were used to analyze the test data. The analysis is conducted respectively from the item, the person, and the person-item map aspects. This essay aims to identify problems in the teaching and learning process of the subject of mathematics through the design and analysis of the test paper, so as to improve the quality of teaching and learning.

2. Content to Be Tested

The content to be tested is two chapters of the Grade 9 mathematics, Volume 1. The two chapters are "Quadratic equations" and "quadratic functions". The reason why both chapters were chosen to be tested at the same time is that they are related, and the learning of quadratic functions can help students to deepen their understanding of quadratic equations. Thus, it may be meaningful to design and assess these two chapters together.

Chapter 1: Quadratic equations. This chapter contains three sections: the conception of quadratic equations, solving quadratic equations, and practical problems and quadratic equations.

Section 1: The conception of quadratic equations. Students will be prompted to think through two practical problems and observe the characteristics of equations, and then the teacher will introduce the concept of quadratic equations. This learning activity aims to enable students to remember and understand the concept of quadratic equations.

Section 2: Solving quadratic equations. This section mainly introduced three methods for solving quadratic equations. These three methods are the matching methods, formula method, and factorization method. The teacher will explain each of the three methods using a practical problem and then set up a group activity to motivate students to discuss with each other and use the three methods to solve the problem and share their solutions with the class. These activities aim to promote students' interests in learning mathematics and collaboration skills.

Section 3: Practical problems and quadratic equations. This part is learning to apply. The teacher will use the COVID-19 scenario to illustrate the application of quadratic equations in the epidemic to calculate the number of people who were infected. The biggest difficulty in teaching is enabling students to understand how to apply what they have learned to solve real problems in daily life, rather than learning by rote. Accordingly, the purpose of this learning activity is to foster students' ability to apply knowledge and enhance their problem-solving skills.

Chapter 2: Quadratic functions. This chapter consists of three sections: Images and properties of quadratic functions, the relationship between quadratic functions and quadratic equations, and practical problems and quadratic functions.

Section 1: Images and properties of quadratic functions. The teacher will show a video of a fountain, prompting students to think about the shape of the sprayed water droplets, to further introduce the concept of quadratic functions. Then the teacher will use the geometry drawing board to present different graphs of quadratic functions for students to observe, and finally, the teacher and students will summarise the images and properties of quadratic functions together. These activities aim to enable students to remember and understand the meaning of quadratic functions.

Section 2: quadratic functions and quadratic equations. The teacher will use the example of playing golf to explain the relationship between the quadratic functions and quadratic equations. It aims to help students deeply understand the link between them. Then have students form groups to solve the problems presented in the slides. These exercises can increase students' problem-solving and critical thinking skills.

Section 3: Practical problems and quadratic functions. This section is similar to chapter 1's section 3. The teacher will play the video about the water level rise and fall in parabolic arch bridges and ask the relevant question, which can be solved by quadratic functions, to engage students to think and cope. After this section, students should be able to use quadratic functions to solve practical problems in life.

The purpose of these two chapters' content and learning activities is to enable students to understand the concept of quadratic equations and quadratic functions and apply them to solve real problems, rather than memorized learning. And it can also develop students' higher-order thinking skills, such as critical thinking, collaboration, and problem-solving skills.

3. Process of Test Design

Classroom tests provide teachers with essential information to use in making decisions about teaching and student achievement. A table of specifications (TOS) is a useful test blueprint that can help teachers align objectives, instruction, and evaluate students' performance (Fives & DiDonato-Barnes, 2013). In this essay, according to the TOS guideline, a mathematics test paper was designed with various types of items.

There are four types of questions in the designed test paper: fill-in-the-blank, true-false items, multiple-choice items, and problem-solving questions. Each type aims to check students' mastery of the knowledge of the two chapters. The test time is 30 minutes, with a total score of 50 points. The participants are 40 junior school students in grade 9. This math examination paper covers four cognitive domains based on the learning objectives, which are remembering, understanding, applying, and analyzing. All examination questions are arranged from simple to complex, which is consistent with cognitive development.

The first part of the test paper is to fill in the blanks. This type of item can usually provide an objective measurement of students' achievement or ability. There are three questions (Q1, Q2, Q3) in the part, with two marks for each. This part aims to investigate students' memories of the concept of quadratic equations. They should know the definition and features of quadratic equations and fill in the correct answer in the blanks.

True-false items are arranged in the second part. This kind of item is good for knowledge-level content, and it can evaluate students' understanding of popular misconceptions. Four questions were designed to test with three points for each. These items examine two aspects of knowledge, one is the solution of quadratic equations (Q4, Q5), and the other is the properties of Quadratic Functions (Q6, Q7). Students should not only remember the formula but also understand the meaning of this knowledge.

The following are the Multiple-Choice items. The MC question covers a broader range of learning objectives. There are five MC items with four marks for each in this test paper. This is a comprehensive part; it requires students to understand the three methods for solving quadratic equations and to apply them to solve problems. Q8, Q9, and Q10 were designed to test the formula method, matching method, and factorization method, respectively. The next item, Q11, investigated the figure of quadratic functions, and Q12 asked about the relationship between quadratic equations and quadratic functions. Students should be able to understand the key points of these two chapters.

The last part is the problem-solving questions. This is the most comprehensive part of the math test paper, which aims to test students' ability to apply and analyze. There are two questions with six marks for each, one of which (Q13) required students to use quadratic equations to calculate the number of people infected during the epidemic. Another question (Q14) was a practical question about the rise and fall of water levels in parabolic arch bridges. These two questions are both real problems in our daily lives. Students need to solve practical problems with what they have learned; it is a further stage that requires students to reach.

As discussed above, the distribution of marks is shown in Table 1.

Table 1. Distribution of scores.

Item number	Q1 - Q3	Q4 - Q7	Q8 - Q12	Q13 - Q14
Item type	Fill-in-the-blank	True-false	Multiple-choice	Problem-solving
Total score	6	12	20	12

4. Analysis of Test Results

To better analyze the test results, Classical Test Theory (CTT) and Item Response Theory (IRT) were adopted in this essay. The combined use of these two theories was intentional to leverage their complementary strengths. CTT (analyzed via SPSS) provides group-dependent measures such as item difficulty (p-value) and discrimination (Pearson correlation), which are straightforward to interpret but sensitive to sample characteristics. In contrast, IRT (analyzed via Winsteps) estimates group-independent item parameters and person abilities on the same scale, enabling more robust comparisons across different populations. By integrating both approaches, we could cross-validate results: for instance, CTT's overall reliability (Cronbach's α) supports test consistency, while IRT's fit statistics (e.g., infit/outfit) pinpoint misfitting items that may require revision. This dual approach aligns with recommendations by Sarı & Karaman (2018) for psychology data.

The standardized achievement test data can help educators evaluate the educational effect of instructional interventions, which is conducive to enhancing students' learning (Sussman & Wilson, 2019). In this essay, 40 students were invited to take a short quiz with the designed math test paper, and the 40 test paper results as the original data will be analyzed with CTT and IRT in three aspects. First is the item analysis. This part discusses the difficulty factor, discrimination index (compare and correlation), the distractors of multiple-choice items, and the reliability of the items. The second is the person analysis. It focuses on the Outfit MNSQ, Infit MNSQ, and the reliability of students. The third part is the Person-Item map. It investigates the location of students' ability and items' difficulty based on the same latent dimension in this test.

4.1. Item Analysis

The results of the analysis of the item data using SPSS and Winsteps are shown in **Table 2**. For the difficulty factor, in general, the appropriate range is 0.3 - 0.7. However, the difficulty factor of the 8 items is more than 0.7 in this test paper, which means that these 8 items are easy for students, and the remaining items are of moderate difficulty. Overall, the questions on this test paper were not too difficult for students.

The discrimination ability (compare) is an indicator that distinguishes the performance of students with high scores from those with low scores. An index greater than 0.2 means good differentiation. In this test paper, 8 items are good as their DI was between 0.2 and 0.5. However, 6 items' DI was less than 0.2, which means that these items can not distinguish between high score students and low score students, especially the Q1 and Q12, the DI of these two items is 0. It indicated that these items can not differentiate between students' ability levels.

The discrimination index (correlation) reflects the consistency of students' item scores and their total scores. It can be seen from **Table 2** that the DI for the 4 items' DI is less than 0.2, and it is worth noting that the DI for Q1 is negative, meaning that students in the lower-performing group answered this question correctly at a higher rate than those in the higher-performing group. This indicates that the item Q1 may have low validity; it should be strongly revised in this test paper. In addition, the student's performance on this item was not consistent with their performance on the test.

As for the Rasch analysis part, the Point Measure Correlation (PMC) shows the construct validity of the items. The PMC shows similar results to the DI; some items (Q1, Q2, Q3, Q9, Q12) were not good at distinguishing students' ability, and their performances on these items were not consistent with their total scores in this test.

The distractors of multiple-choice questions were analyzed in groups 1, 2, 3. Two main problems with these multiple-choice questions can be identified from

Table 2. Item data statistics.

SPSS							Winsteps			
Item	Full score	Mean	MeanH-MeanL	Difficulty factor	Discrimination Index (Correlation)	Discrimination ability (Compare)	Point measure correlation	Rasch measure	In-fit MNSQ	Out-fit MNSQ
Q1	2	1.80	0.00	0.90	−0.05	0.00	0.03	44.89	1.08	1.32
Q2	2	1.90	0.33	0.95	0.27	0.17	0.17	41.07	0.96	0.67
Q3	2	1.45	0.33	0.73	0.09	0.17	0.14	51.33	1.09	1.15
Q4	3	1.95	1.25	0.65	0.32	0.42	0.30	54.04	1.02	0.98
Q5	3	1.80	0.50	0.60	0.25	0.17	0.25	54.87	1.14	1.13
Q6	3	2.10	1.50	0.70	0.54	0.50	0.38	53.16	0.85	0.77
Q7	3	2.10	1.50	0.70	0.54	0.50	0.38	53.16	0.85	0.77
Q8	4	3.50	1.00	0.88	0.31	0.25	0.22	50.50	1.00	0.93
Q9	4	3.80	0.33	0.95	0.12	0.08	0.11	47.79	1.08	0.89
Q10	4	3.70	1.00	0.93	0.27	0.25	0.19	48.95	1.03	0.64
Q11	4	2.70	2.00	0.68	0.54	0.50	0.40	54.06	0.89	0.74
Q12	4	3.90	0.00	0.98	0.01	0.00	0.04	45.92	1.07	1.35
Q13	6	5.13	1.25	0.85	0.50	0.21	0.37	46.63	0.86	0.80
Q14	6	3.65	1.58	0.61	0.33	0.26	0.34	53.62	1.22	1.22

SPSS scale reliabity: 0.298

Rasch Item Reliability: 0.79

the group cross-tabulations. Firstly, invalid options were found in all MC items. Specifically, option B in Q8 (Table 3), option A, B in Q9 (Table 4), option D in Q10 (Table 5), option A in Q11 (Table 6), and option A, B in Q12 (Table 7). It means that these options are meaningless because none of the students would choose them; these options were unattractive distractors. Moreover, other distractors should also be considered for improvement, as the number of students who chose these distractors is particularly low.

Another major problem is that the correct options of these MC questions can not effectively distinguish students' ability levels, especially Q8, Q9, Q10, and Q12. It shows that a similar number of students in the high, medium, and low scoring groups chose the correct option in these four MC items, and the medium score group even a little bit more than the high score group. It reminds us that the correct option needs to be optimized to better differentiate between students' ability levels.

Furthermore, the lack of difficulty and differentiation of the questions also contributes to the low reliability measured by SPSS (Cronbach's α is 0.298). Additionally, the number of items in this test is small, with only 14 questions, which may affect the items' reliability.

Table 3. Q8 Group Cross-tabulation.

		Group			
		1	2	3	Total
Q8	1 (Correct)	12	14	9	35
	3	0	1	0	1
	4	0	1	3	4
Total		12	16	12	40

Table 4. Q9 Group Cross-tabulation.

		Group			
		1	2	3	Total
Q9	3 (Correct)	12	15	11	38
	4	0	1	1	2
Total		12	16	12	40

Table 5. Q10 Group Cross-tabulation.

		Group			
		1	2	3	Total
Q10	1	0	0	1	1
	2 (Correct)	12	16	9	37
	3	0	0	2	2
Total		12	16	12	40

Table 6. Q11 Group Cross-tabulation.

		Group			
		1	2	3	Total
Q11	2	0	3	2	5
	3	0	4	4	8
	4 (Correct)	12	9	6	27
Total		12	16	12	40

Table 7. Q12 Group Cross-tabulation.

		Group			
		1	2	3	Total
Q12	2 (Correct)	12	15	12	39
	3	0	1	0	1
Total		12	16	12	40

4.2. Person Analysis

The students' analysis is shown in Table 8. The Outfit MNSQ and Infit MNSQ are two indicators in the Rasch model, which can assess the item's difficulty in relation to the person's ability. A range of MNSQ values from 0.5 to 1.5 is productive for measurement. As shown in Table 8, the Outfit MNSQ of four students (NO.18, 19, 30, 32) was less than 0.5, which indicates that their test data is too ideal and is less productive for measurement. Moreover, seven students (NO.8, 21, 27, 34, 39, 40) whose Outfit MNSQ is greater than 1.5 suggest that these students may have been guessing, cheating, and carelessness. One of the students whose Outfit MNSQ is greater than 2.0 can not match the model well. Furthermore, the student's reliability as measured by the Rasch model was low at 2.2. In addition to the reasons discussed above, the small sample size and the small number of items may also have contributed to the low reliability.

Table 8. Person data statistics.

		Rasch		
id1	id2	Rasch measure	In-fit MNSQ	Out-fit MNSQ
0	1	76.53	1.00	1.00
0	2	57.92	1.00	0.59
0	3	76.53	1.00	1.00
0	4	60.50	0.98	0.37
0	5	60.50	0.51	0.55
0	6	54.82	0.75	0.61
0	7	60.50	0.46	0.54
0	8	55.25	1.55	1.95
0	9	55.25	1.24	0.89
1	0	56.72	1.17	0.79
1	1	56.20	0.77	0.47
1	2	57.92	1.16	0.79
1	3	54.82	0.94	0.74
1	4	57.29	0.95	0.52
1	5	57.29	1.31	1.36
1	6	57.29	0.95	0.52
1	7	57.29	0.87	0.59
1	8	57.92	0.77	0.43
1	9	61.84	0.70	0.25
2	0	52.88	0.92	0.75
2	1	59.48	0.68	1.51
2	2	52.88	0.64	0.56

Continued

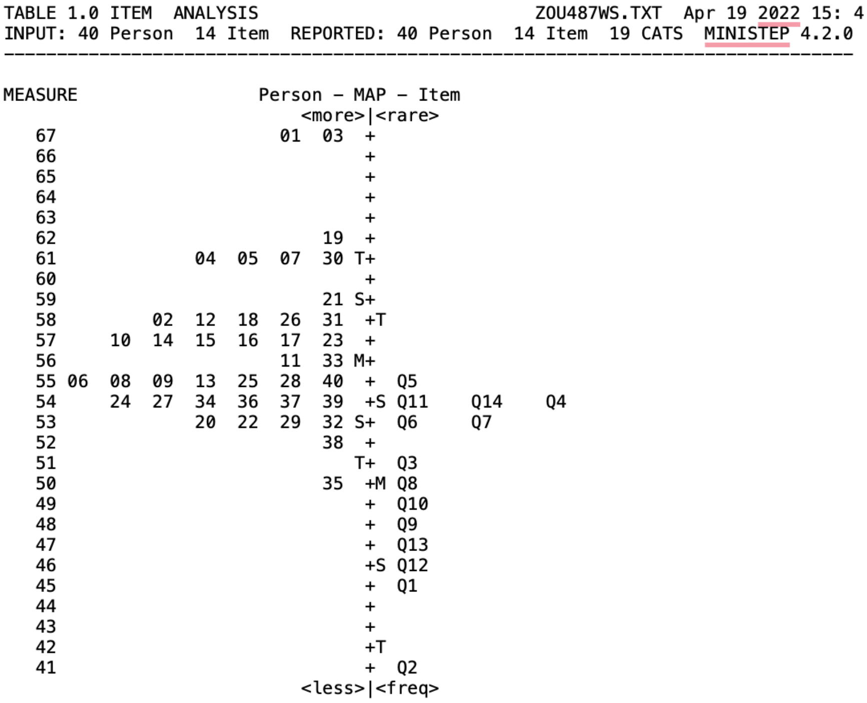
2	3	56.72	0.79	0.57
2	4	54.41	0.99	0.76
2	5	54.82	1.14	0.88
2	6	57.92	0.83	0.54
2	7	54.01	1.55	1.58
2	8	55.25	0.89	0.66
2	9	52.88	1.15	0.99
3	0	60.50	0.83	0.47
3	1	57.92	0.79	1.33
3	2	52.88	0.51	0.46
3	3	55.71	1.30	4.11
3	4	54.41	1.98	1.71
3	5	50.46	0.61	0.62
3	6	54.01	1.60	1.36
3	7	54.41	0.75	1.02
3	8	52.17	0.60	1.12
3	9	54.41	1.08	1.72
4	0	54.82	1.38	1.64
Rasch Person Reliability: 0.22				

4.3. Person-Item Map

The Person-Item map displays the correspondence between the person's abilities and item difficulties lying on the same potential dimension. It can be seen from Figure 1 that students No.01 and No.03 were the best on the test; they got the highest grade. Then followed by student No.19, while student No.35 was the lowest performer on this test. The ability distribution of the other students is relatively concentrated. For the item, the Q5 is the most difficult one in the test, while the Q2 is the easiest question. Overall, most of the students had similar performance, and their performance was higher than the difficulty of almost all items. It can be argued that, on the one hand, students have a better grasp of the two chapters' knowledge; on the other hand, this mathematics test paper should be designed with more difficult items to match students' abilities.

5. Discussion and Suggestions

The test revealed three key issues: a) Ambiguous wording in Q1. The negative discrimination of Item 1 suggests that higher-ability students performed worse, possibly due to ambiguous wording. Because this item requires students to list the area equation of a rectangle through the unknown x, but the question is not to explain the format of the answer, many students list the expression $x(x-2)$



```
TABLE 1.0 ITEM  ANALYSIS                          ZOU487WS.TXT  Apr 19 2022 15: 4
INPUT: 40 Person  14 Item  REPORTED: 40 Person  14 Item  19 CATS  MINISTEP 4.2.0
--------------------------------------------------------------------------------

MEASURE                 Person - MAP - Item
                            <more>|<rare>
   67                     01  03  +
   66                             +
   65                             +
   64                             +
   63                             +
   62                         19  +
   61             04  05  07  30 T+
   60                             +
   59                         21 S+
   58         02  12  18  26  31  +T
   57     10  14  15  16  17  23  +
   56                     11  33 M+
   55 06  08  09  13  25  28  40  +  Q5
   54     24  27  34  36  37  39  +S Q11    Q14    Q4
   53             20  22  29  32 S+  Q6     Q7
   52                         38  +
   51                            T+  Q3
   50                         35  +M Q8
   49                             +  Q10
   48                             +  Q9
   47                             +  Q13
   46                             +S Q12
   45                             +  Q1
   44                             +
   43                             +
   42                             +T
   41                             +  Q2
                            <less>|<freq>
```

Figure 1. Person-Item map.

instead of the equation x(x − 2) = 100. Actually, they know the concept of the area equation of a rectangle; however, as the test paper did not clearly specify the format of the answers, high-scoring students made mistakes in the answers. Teachers should clearly state specific requirements. b) The distractors in the multiple-choice questions are not attractive enough. For instance, the Q11 is to find the analytical expression of a quadratic function given that its graph passes through three points. The three distractors are quite different from the correct ones, and it is relatively easy to select the correct answer. On the other hand, this item also reflects that students are somewhat vague about the knowledge points of quadratic function expressions and graphs. During the teaching process, teachers can strengthen the knowledge points by comparing function graphs, and when setting options, they can combine the points that students are prone to make mistakes to enhance the setting of interfering options. And c) the difficulty of the question is too simple. A range of MNSQ values from 0.5 to 1.5 is productive for measurement. As shown in Table 3, the Outfit MNSQ of four students (NO.18, 19, 30, 32) was less than 0.5, which indicates that their test data is too ideal and is less productive for measurement. With the item data statistics, we can see that in this test, the items are easy for students, so teachers should increase the difficulty of the questions.

The analysis of test results shows that the critically low reliability scores (Cronbach's α = 0.298; Rasch person reliability = 0.22) primarily reflect the test's limited design. First, small sample size (N = 40). With a small sample size, estimates of item parameters (e.g., discrimination) become unstable, artificially depressing reliability (Kaya et al., 2016). Second, short test length (14 items). Classical test theory notes that reliability increases with test length (Alger, 2016); a 14-item test is unlikely to achieve α > 0.7 unless items are highly homogeneous. Third, poorly performing items. The negative discrimination index (Q1: r = −0.05) suggests possible miskeying or ambiguous wording, as higher-ability students may have answered incorrectly. Such items reduce internal consistency.

In short, the small number of items in this test influences the reliability of the test, so additional questions need to be added to the further test paper. And the expression of the item should be precise and clear, without any ambiguous words. With the suitable number, clear expression, and appropriate difficulty of items, the validity and reliability of the test can be ensured. What's more, tests can be conducted in two classes to increase the sample size in the future.

6. Limitations

This study has several limitations that should be acknowledged: a) Sample size constraints: With only 40 participants, the estimates of item parameters (e.g., the negative discrimination of Q1) may be unstable. A simulation study by Aksu et al. (2019) showed that with a sample size over 1000, more consistent results can be obtained in the studies performed with artificial neural networks in the field of education. b) Short test length: The 14-item test is insufficient to cover the full breadth of "Quadratic equations" and "quadratic functions", resulting in low reliability (α = 0.298) and restricted validity evidence. c) Single-school context: All students were recruited from one urban school, which may not represent rural populations or diverse curricular approaches.

These constraints imply that the test is unsuitable for high-stakes decisions (e.g., placement or certification). While the analysis provides exploratory insights into item performance, any conclusions about individual student abilities should be drawn with caution. Future studies should increase the sample size, test length, and item quality to improve reliability. In the future, the diversity of schools can be further increased, such as by including students from both urban and rural schools in the test analysis to enrich the results of research.

7. Conclusion

This essay discussed the process of designing and analyzing a mathematics test paper for Junior High School students in grade 9 in mainland China. It aims to measure students' mathematics performance and find the teaching and learning problems through the test paper, so as to improve the quality of instruction. Based on the TOS, I designed a mathematics test paper with clear objectives, and then conducted the test paper and collected the test data. According to the assessment theories and analysis tools, an analysis of students and items has been discussed. I found that the average score of the students is concentrated, the overall difficulty of items is low, the test length is short, and the discrimination index should be increased to distinguish the students' abilities. Based on this analysis information, an optimized version of the math test paper has been redesigned. Though the study shows some limitations, such as a small sample size, a short test length, simple items, and a single-school context. It provides some useful and practical information for a teacher to identify the problems during the teaching and learning process. It can help teachers know how to design a good test paper, how to analyze test data, and discover potential problems. Then teachers can optimize teaching and learning. Of course, future research should be improved in light of these limitations to better assess students' mathematical knowledge.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Aksu, G., Güzeller, C. O., & Eser, M. T. (2019). The Effect of the Normalization Method Used in Different Sample Sizes on the Success of Artificial Neural Network Model. *International Journal of Assessment Tools in Education, 6,* 170-192. https://doi.org/10.21449/ijate.479404

Alger, S. (2016). Is This Reliable Enough? Examining Classification Consistency and Accuracy in a Criterion-Referenced Test. *International Journal of Assessment Tools in Education, 3,* 137-150. https://doi.org/10.21449/ijate.245198

Fives, H., & DiDonato-Barnes, N. (2013). Classroom Test Construction: The Power of a Table of Specifications. *Practical Assessment, Research, and Evaluation, 18,* 1-7.

Kaya, Y., Leite, W. L., & Miller, M. D. (2016). A Comparison of Logistic Regression Models for DIF Detection in Polytomous Items: The Effect of Small Sample Sizes and Non-Normality of Ability Distributions. *International Journal of Assessment Tools in Education, 2,* 22-39. https://doi.org/10.21449/ijate.239563

Rockman, I. F. (2002). The Importance of Assessment. *Reference Services Review, 30,* 181-182. https://doi.org/10.1108/00907320210435455

Sarı, H. İ., & Karaman, M. A. (2018). Gaining a Better Understanding of General Mattering Scale: An Application of Classical Test Theory and Item Response Theory. *International Journal of Assessment Tools in Education, 5,* 668-681. https://doi.org/10.21449/ijate.453337

Sussman, J., & Wilson, M. R. (2019). The Use and Validity of Standardized Achievement Tests for Evaluating New Curricular Interventions in Mathematics and Science. *American Journal of Evaluation, 40,* 190-213. https://doi.org/10.1177/1098214018767313