International Journal of Evaluation and Research in Education (IJERE)
Vol. 15, No. 1, February 2026, pp. 286~302
ISSN: 2252-8822, DOI: 10.11591/ijere.v15i1.35698

Development and validation of curriculum-based problem-solving MCQ classroom tests in mathematics

Purity V. Mata[1,2]**, Quennie Rose H. Abao**[1]**, Roxanne Kaye S. Bihasa**[1]**, Glenda N. Kiamco**[1]**, Anna Marie C. Neiz**[1]
[1]College of Education, Arts, and Sciences, Cebu Technological University-Carmen Campus, Cebu, Philippines
[2]Naveen Jindal Young Global Research Fellowship, O.P. Jindal Global University, Sonipat, India

Article Info

Article history:

Received May 5, 2025
Revised Dec 17, 2025
Accepted Jan 1, 2026

Keywords:

Curriculum-based assessment
Item analysis
Mathematics problem-solving
Reliability testing
Teacher-made MCQs
Test development
Test validation

ABSTRACT

Classroom assessment is a fundamental element in the teaching and learning process. The objective of this study is to determine the validity of the teacher-made multiple-choice questions (MCQs) used in General Education Course Mathematics in the Modern World (GEC-MMW). Employing a quantitative descriptive research design, four sets of teacher-made MCQ tests were developed and administered to over 500 students across two campuses of a state university. The tests underwent expert validation, tryouts, and analysis using difficulty index (DIF I), discrimination index (DI), distractor efficiency (DE) and reliability testing. Most items showed an average level of difficulty, acceptable to excellent DI, and functional distractors (FD), with reliability coefficients ranging from 0.54 to 0.78. Moreover, mean percentage scores ranged from 42.94 to 50.95 revealed below-average student performance. These findings confirm that the developed tests are valid, reliable, suitable for classroom assessment, and demonstrate that item analysis is essential for improving test quality, identifying learning gaps, and informing instructional decisions.

Corresponding Author:

Purity V. Mata
College of Education, Arts, and Sciences, Cebu Technological University-Carmen Campus
R.M. Durano Ave, Poblacion, 6005 Carmen, Cebu, Philippines
Email: mata.purity@ctu.edu.ph

1. INTRODUCTION

Classroom assessment is a fundamental element in the teaching and learning process. They provide evidence of student learning and enable educators to make informed instructional decisions [1]. Through assessments, teachers can identify students' learning gaps, track student progress, and adjust teaching strategies. Observations from classroom assessments suggest that students struggle with problem-solving and reasoning skills. The Organization for Economic Co-operation and Development (OECD) [2] reported that the Philippines has consistently performed poorly in the Programme for International Student Assessment (PISA) in mathematics. Generelao *et al.* [3] also disclosed the same outcome in the Trends in International Mathematics and Science Study (TIMSS). Recent evidence highlights that academic preparedness in secondary education remains one of the most reliable indicators of success for university mathematics performance [4]. This underperformance remains a critical challenge for the Philippine educational system that raises concerns on the effectiveness of learning and assessments. When students' scores do not meet the expected benchmarks, it calls into question how they advance to the next grade level despite not achieving the prescribed learning objectives. Tertiary-level mathematics is more advanced, which can lead to challenges when teaching the subject to students with weak mathematical background. These challenges often relate to integrating mathematical applications into non-mathematics disciplines. For example, research by Tsui and Khan [5] found that engineering students with weaker secondary mathematical backgrounds often struggle more with required mathematical components of their courses. Similarly, Romo-Vázquez and Artigue [6] discuss how presenting advanced mathematical content to students without deep prior preparation is particularly difficult, emphasizing that these challenges persist as educational demands and disciplinary requirements evolve. The students' learning gaps can be mitigated using targeted assessment tools, such as teacher-made tests, specific to the curriculum. However, teacher-made tests which are frequently used in tertiary classrooms often lack systematic validation, which may compromise the accuracy of measuring student learning. Despite their widespread use, very few studies in the Philippines have examined the development and validation of teacher-made mathematics tests at the tertiary level. This gap underscores the need for localized, empirically validated assessment tools that can provide accurate measures of student problem-solving skills and support data-driven improvements in instruction.

Teacher-made tests remain the most commonly used assessment tool in the classroom [7], particularly multiple-choice questions (MCQs). While several authors [8]–[10] collectively affirm the popularity of MCQs as an assessment tool in schools for its objectivity and ease in scoring, other authors have criticized the limitations of MCQs. For example, MCQs were found to only test what students can recall and fail to measure their ability to apply and analyze course material. In addition, MCQs were challenged for emphasizing memorization, making them ineffective for assessing critical thinking skills such as analysis and inference [11]. In contrast, other authors [10], [12]–[14], assert that well-designed MCQs promote deeper analytical thinking, allowing examiners to assess students' knowledge integration, problem-solving skills, and application of knowledge. Hassan and Hod [15] emphasize that MCQs can consistently assess student knowledge with a high degree of reliability and validity. These perspectives highlight the evolving recognition of MCQs not merely as tools for rote recall but as effective instruments for evaluating higher-order thinking skills when carefully constructed. This highlights the importance of thoughtful design and rigorous validation processes to maximize the pedagogical value of multiple-choice assessments, which is relevant to tertiary education, where college students are expected to think at a higher cognitive level. When properly designed and validated to incorporate higher-order cognitive skills, these tests not only reduce assessment biases but also enhance the quality of educational outcomes by ensuring content validity (CV) and alignment with curricular goals [16]. Unclear or poorly designed test items may compromise assessment validity and reliability, resulting in inaccurate measurements of student learning.

Poor student performance in assessments is often linked to low-quality teacher-made tests, which may lack validity and reliability due to ambiguous items, misalignment with learning objectives, or construction flaws, leading to unreliable measurement of true student ability [17]. Tertiary teachers as an administrator in the classroom are required to have the ability to formulate a qualified test; however, the procedure of constructing and standardizing tests requires sufficient knowledge, time, and financial resources [18]. Numerous studies have reported the continued utilization of teacher-made tests despite containing item-writing flaws [19], [20]. Furthermore, previous research reported that teacher-made tests commonly evaluate what might be characterized as lower-order thinking skills instead of higher-order thinking skills that require deeper analysis [21]. This can lead to poorly constructed questionnaires that potentially impact the overall assessment of the quality of learning in education [22].

Matazu and Julius [23] emphasized that strict adherence to the principles of test construction, administration, and analysis is essential to prevent the creation of flawed test items. Improving the quality of teacher-made tests can help identify specific learning gaps, allowing for targeted interventions that address students' needs. Among various test formats, MCQs are widely used in standardized assessments due to their efficiency, objectivity, and capacity to measure a wide range of cognitive levels. As noted by Elgadal and Mariod [10], well-constructed MCQs can effectively assess higher-order thinking skills when designed according to established item-writing guidelines. Similarly, Liu *et al.* [24] supports the validity of MCQs in educational measurement, particularly when aligned with learning outcomes. Furthermore, item analysis plays a crucial role in ensuring that test items possess both validity and reliability [25]–[27]. According to educational assessment experts, a quality test should possess the following key characteristics: objectivity, reliability, validity, and practicability [28]. Ensuring these characteristics is essential to the development of assessments that provide accurate and meaningful measures of student learning.

While many studies have examined the development and validation of tests, these investigations have primarily focused on disciplines such as the sciences and allied fields [25], [29]–[35] or have targeted secondary-level students [36]. Research on test construction in this domain is limited, and those that exist were primarily carried out outside the Philippines [37], highlighting the need for more localized research focused specifically on mathematics assessments for tertiary education within the Philippine context. This study aims to develop and validate a curriculum-aligned mathematics tests for tertiary students enrolled in the General Education Course Mathematics in the Modern World (GEC-MMW), specifically targeting problem-solving competencies. Problem solving, a core 21st century skill, involves observation and critical thinking which is vital cognitive process that students must develop to succeed in the modern world [38].

By collaborating with mathematics education experts and employing a structured test development framework, the primary objectives of the study are to: i) design a standardized multiple-choice tests that aligns to the course's intended learning outcomes on the topic problem-solving and logical reasoning skills; ii) ensure the tests' validity through expert reviews (content and face validation); iii) evaluate item quality using difficulty index (DIF I), discrimination index (DI), and distractor efficiency (DE); and iv) establish reliability through tryouts (preliminary, proper, and final phases). To guide the investigation, the following research questions were posed:

- What is the content and face validity of the developed test items?
- What are the indices of difficulty, discrimination, and DE of the developed test items?
- What are the reliability coefficients of the developed set of tests?

The study seeks to determine whether the developed curriculum-based MCQ tests meet acceptable standards of validity and reliability in measuring the problem-solving skills of tertiary-level students. Through this approach, it aims to generate empirical evidence that supports the effectiveness of teacher-made assessments.

This research addresses critical gaps in localized assessment practices, particularly within the Philippine tertiary education context by offering a validated model for mathematics test construction. Specifically, the study is expected to contribute: i) a practical framework for developing curriculum-based multiple-choice tests in higher education; ii) empirical evidence on the validity and reliability of teacher-made assessments; and iii) insights that can guide instructors in improving test quality and aligning assessments with learning outcomes. These contributions aim to strengthen evidence-based practices in mathematics education and enhance the evaluation of student achievement in tertiary settings.

2. METHOD

2.1. Sample and population

The participants in the study were acquired from two separate campuses of the same university through a purposive sampling method. It was utilized for the study because it allowed the deliberate inclusion of individuals who met the predefined criteria essential to the study: i) they must be currently enrolled at the university for SY 2024-2025 and ii) be enrolled in the course GEC-MMW. This ensured that the participants possessed the specific characteristics and contextual relevance necessary to address the research objectives. The ages of the participants range from 20 to 22 years old, with more males (52.88%) than females (47.12%). They were enrolled in various academic programs with 25.69%, 21.57%, 19.23%, 13.46%, 11.68%, and 8.38% representing industrial technology, education, marine engineering, mechatronics, hospitality management, and fisheries, respectively. The number of participants for each test varied depending on their availability as well as voluntary participation, and due to differences in the completeness of responses. Test papers which contained a significant number of unanswered items were excluded from the analysis to ensure accuracy and reliability of the results. Thus, from the 808 enrolled students, 515, 525, 514, and 513 student responses were analyzed for Test 1, Test 2, Test 3, and Test 4, respectively. Formal ethical approval was not obtained prior to data collection due to institutional constraints. However, all ethical procedures were strictly observed. Informed consent was obtained from all participants before test administration, ensuring their awareness of and compliance with the provisions of the data privacy act of 2012 [39].

2.2. Instrumentation

Four set of tests to measure students' problem-solving skills in mathematics were developed with the following topics: i) types of reasoning (Test 1) containing 40 MCQs; ii) problem-solving involving patterns (Test 2) with 20 MCQs; iii) Polya's problem-solving strategies (Test 3) involving 20 MCQs; and iv) recreational problem-solving (Test 4) with 20 MCQs. Each MCQ contains one correct response which is given one mark. Before the development of the tests, a test blueprint was designed using the university template of the table of specification (TOS). Rows represented the intended learning outcomes derived from the syllabus whereas columns corresponded to Bloom's taxonomy of cognitive domains. The allocation of test items was based on the proportion of instructional hours devoted to each competency, with the total number of items set at a minimum of 50 for the entire test. For instance, if 4 hours was spent on the competency "using different types of reasoning to justify statements and arguments made about mathematics and mathematical concepts," out of a total of ten instructional hours, the corresponding number of items was computed as $(4/10)x50$=20 items. This value was then multiplied by two to account for the expected reduction in items during validation and item analysis. In addition, items were distributed according to Bloom's taxonomy, with a substantial proportion designed to assess higher order thinking skills across all the competencies.

2.3. Procedure

The tests underwent content and face validation by eight subject matter experts (SMEs). The inclusion criteria for the SMEs were as: i) possession of at least a bachelor's degree in mathematics education and ii) a minimum of five years of teaching experience in mathematics. As presented in Table 1, the SMEs exceeded these requirements, possessing postgraduate qualifications and teaching experience ranging from 5 to 27 years with an average of 12 years. Their academic training and professional exposure ensure that the tests underwent thorough and informed evaluation. Table 1 shows the profile of the SMEs.

Face validation was performed after the content validation. Following expert validation, preliminary testing was conducted with ten students to identify any language difficulties, word ambiguities, unclear instructions, and to establish the time allocation for each test. During this phase, students recorded their start and end times for each test to estimate the proper duration for the subsequent try-outs. Only 10 students were involved in this phase, as this number was considered sufficient to gather meaningful feedback on formulation issues and linguistic clarity through a group interview [40], [41]. Next, a proper try-out was conducted with 50 students answering the tests. This is a sufficient sample size to obtain stable item estimates for initial psychometric analysis [42]. Item analysis was performed using the upper-lower (UL) index method to calculate the DIF I and the DI. The 27% cut-off was utilized because it maximizes the differences in a normal distribution while keeping enough cases for stable analysis. This balance, widely supported by education experts [43], ensures clearer discrimination between groups. Items that fall under the acceptable range were retained, and those that do not were revised. Lastly, the final try-out was carried out, administering the test to more than 500 students from the same population, exceeding the minimum of 400 respondents for reliability studies [44]. This suggests that the results of the reliability coefficients are stable, less prone to error, and more generalizable. Test administrations from the preliminary to the final try-outs were personally conducted by the researchers. A script was prepared to ensure uniformity of instructions. After the final try-out, an item analysis was conducted to evaluate the performance and quality of each test item. Based on the results of the item analysis, decisions to either retain or discard the items were then recommended. Items that were either easy or difficult and exhibited poor discriminating power were discarded, while items with average difficulty and a DI of 0.19 or higher were retained [45]. Subsequently, the internal reliability of the test was calculated using the Kuder-Richardson 20 (KR-20) formula. Figure 1 illustrates the successive stages we traversed during the development and validation process of the tests.

Table 1. Profile of the SMEs

Code	Age	Gender	Years of teaching	Bachelor's degree	Highest level of education
E1	30	Male	8	Bachelor in secondary education (BSED) major in mathematics	Master of arts in education
E2	38	Female	7	BSED major in mathematics	Doctor of development education
E3	50	Female	27	BSED major in mathematics	Master of arts in education
E4	31	Male	11	BSED major in mathematics	Ph. D. in research management and development
E5	41	Male	16	BSED major in mathematics	Master of arts in education
E6	41	Male	10	BSED major in technology and livelihood education and minor in mathematics	Master of arts in education
E7	34	Female	12	BSED major in mathematics	Doctor of development education
E8	27	Male	5	BSED major in mathematics	Master of arts in education

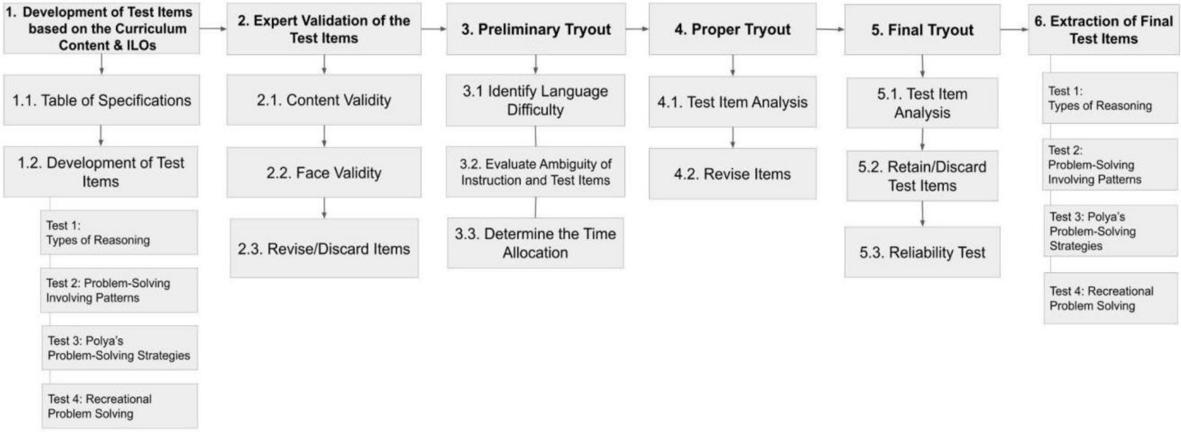

1. Development of Test Items based on the Curriculum Content & ILOs
1.1. Table of Specifications
1.2. Development of Test Items
Test 1: Types of Reasoning
Test 2: Problem-Solving Involving Patterns
Test 3: Polya's Problem-Solving Strategies
Test 4: Recreational Problem Solving
2. Expert Validation of the Test Items
2.1. Content Validity
2.2. Face Validity
2.3. Revise/Discard Items
3. Preliminary Tryout
3.1 Identify Language Difficulty
3.2. Evaluate Ambiguity of Instruction and Test Items
3.3. Determine the Time Allocation
4. Proper Tryout
4.1. Test Item Analysis
4.2. Revise Items
5. Final Tryout
5.1. Test Item Analysis
5.2. Retain/Discard Test Items
5.3. Reliability Test
6. Extraction of Final Test Items
Test 1: Types of Reasoning
Test 2: Problem-Solving Involving Patterns
Test 3: Polya's Problem-Solving Strategies
Test 4: Recreational Problem Solving

Figure 1. Flow of the study

2.4. Data analysis

Following data collection, two datasets were encoded in Microsoft Excel ver. 16 for analysis. The first dataset consisted of the students' original responses, recorded using letter options (e.g., A, B, C, D), and was used for efficient analysis of DIF I, DI, and DE. The second dataset was recorded in binary format, assigning '1' for correct and '0' for incorrect responses. This format enabled the expedient computation of the tests' reliability coefficient scores.

2.4.1. Content validity (CV)

CV refers to the assessment conducted by SMEs of whether the test items are representative of what the assessment tool intends to measure. A three-point scale was used to indicate a test item as 1=essential, 2=useful but not essential, and 3=not essential. In this study, CV was evaluated using the content validity ratio (CVR) of each item and was followed by computing the content validity index (CVI) of each test. The CVR was computed employing Lawshe's method [46], while the CVI was computed as the average CVR of all items within the test. The value of CVR varies from -1 to 1. A CVR value closer to 1 indicates a higher level of agreement among experts regarding an item's relevance. Therefore, items with CVR values approaching 1 were considered strong candidates for inclusion in the instrument [47]. The minimum acceptable CVR and CVI values depended on the number of SMEs involved in the evaluation. For this study, the threshold was set at 0.75, which is considered acceptable [48]. Following the computation of the CVR values, items that did not meet the threshold were discarded. This led to a reduction in the number of items, particularly for Tests 1 and 4. In Test 1, only 32 out of the original 40 items were retained, while in Test 4, 17 of the initial 20 items were kept. The number of items for Tests 2 and 3 remained the same. In addition, the CVI values of Tests 1 to 4 were 0.86, 0.93, 0.91, and 0.99, respectively, reflecting an excellent validity of the tests. Table 2 presents these findings.

Table 2. Results of the CVR and the CVI

CVR value	Test 1 (item number)	Test 2 (item number)	Test 3 (item number)	Test 4 (item number)	Decision
1.00	2, 5, 6, 9, 12, 13, 15, 17, 20, 25, 26, 35, 39, 40	1, 2, 3, 6, 7, 9, 11, 12, 13, 14, 15, 16, 17, 20	7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	2, 3, 9, 11, 14,	Retain
0.75	3, 4, 7, 8, 11, 16, 21, 22, 27, 29, 30, 31, 32, 33, 34, 36, 37, 38	4, 5, 8, 10, 18, 19	1, 2, 3, 4, 5, 6, 8,	1, 4, 5, 6, 7, 10, 12, 15, 17, 18, 19, 20	Retain
0.50	1, 10, 14, 18, 23			8, 13, 16	Discard
0.25	19, 24, 28				Discard
Total items	40	20	20	20	
CVI (items not discarded)	0.77	0.93	0.91	0.91	
Total items after decision	32	20	20	17	
CVI (items are discarded)	0.86	No items are discarded	No items are discarded	0.99	

2.4.2. Face validity

Face validity refers to the subjective evaluation of how relevant the test items are in measuring the intended construct and its overall presentation, including its feasibility, readability, consistency in style and formatting, and the clarity of language used [48], [49]. Although face validity can be assessed by non-experts, it remains a crucial component during the initial development of an instrument [50]. In this study, it was assessed using eight criteria: clarity of objective, appropriateness of the format, appropriateness of the font size, clarity of every item, clarity of instruction, adequacy of item count, cultural appropriateness of terms, and the range of response variation. For example, appropriateness of font size ensured readability for all participants and clarity of every item assessed whether the test questions were free from ambiguity. The item-level face validity index (I-FVI) and the scale-level face validity index using the average method were computed, with. I-FVI as the proportion of raters who rated the statement 1 (strongly disagree) to 5 (strongly agree) divided by the number of raters and S-FVI/Ave as the average of all I-FVI values. Acceptable I-FVI value is ≥0.80, while an acceptable S-FVI value would be ≥0.83 [51].

Table 3 shows that seven out of eight criteria achieved a perfect agreement among experts, with I-FVI values ranging from 0.875 to 1. Moreover, the S-FVI/Ave was 0.83, reflecting a strong overall face validity. This indicates that the assessment tool is generally perceived as appropriate, clear, relevant, and acceptable within the FV parameters.

Table 3. SMEs' agreement on face validity

Criteria	Experts' ratings								I-FVI
	E1	E2	E3	E4	E5	E6	E7	E8	
1. Clarity of objective	5	4	5	5	5	5	5	5	1
2. Appropriateness of the format	5	5	5	5	5	5	5	4	1
3. Appropriateness of the font size	5	5	5	5	5	4	5	5	1
4. Clarity of every item	5	5	4	5	5	5	5	5	1
5. Clarity of instruction	5	5	4	5	5	5	5	5	1
6. Adequacy of item count	5	5	5	4	5	4	4	5	0.875
7. Cultural appropriateness of terms	5	4	4	5	5	5	5	5	1
8. Range of response variation	5	5	4	5	5	5	5	5	1
Average I-FVI									0.98
Proportion relevance	1	0.75	0.50	0.86	1	0.75	0.875	0.875	
S-FVI/Ave									0.83

2.4.3. Inter-rater reliability

Inter-rater reliability refers to the degree of agreement among SMEs when evaluating the same data, with higher values indicating stronger consistency in their judgments. In this study, inter-rater reliability was assessed using percent agreement and the computation followed the method presented by Shabankhani [52]. While percent agreement is a straightforward measure, it does not account for agreement by chance. However, it is considered acceptable when raters share similar expertise and training, which minimizes variability in interpretation [53]. The use of SMEs with comparable domain knowledge strengthens the consistency of scoring and enhances the reliability of the test. Although future studies may consider complementing percent agreement with chance-corrected measures such as Cohen's or Fleiss' Kappa for more robust evidence of inter-rater reliability.

Table 4 presents the percent agreement values representing inter-rater reliability of SMEs ratings of the test items' content and face validity. The percent agreement across the four tests ranges from 86 to 93 percent which exceeds the conventional 0.80 threshold for acceptable agreement and is therefore indicative of strong agreement among SMEs regarding the appropriateness and relevance of the test items to the intended domain [53]. In contrast, the face validity percent agreement falls below this threshold, aligning more closely with the substantial agreement range in kappa-based interpretation scales [54].

Table 4. Inter-rater reliability for content and face validity ratings

Inter-rater reliability	CV				Face validity
	Test 1	Test 2	Test 3	Test 4	
Percent agreement	0.86	0.93	0.91	0.90	0.69

2.4.4. Item analysis

A key step in the test development cycle is item analysis [55]. It is a procedure that is performed after an examination or test is constructed and administered. Two theoretical approaches are widely used as frameworks for item analysis. Classical test theory (CTT) [56] focuses on the reliability of test scores, assuming that all items in a test contribute equally to students' overall performance [57]. In contrast, item response theory (IRT), emphasizes the quality of test items by accounting for variations in item difficulty, recognizing that some items are inherently more challenging than others [58]. IRT argues that the probability of success on an item is influenced by both the student's ability and the item's difficulty, offering a more nuanced approach to analyzing test performance. Item analysis consists of several components to determine how well each item in a test function and contributes to the overall reliability and validity of a test. Therefore, it can be used to revise and improve the effectiveness of test items [59].

a. Difficulty index

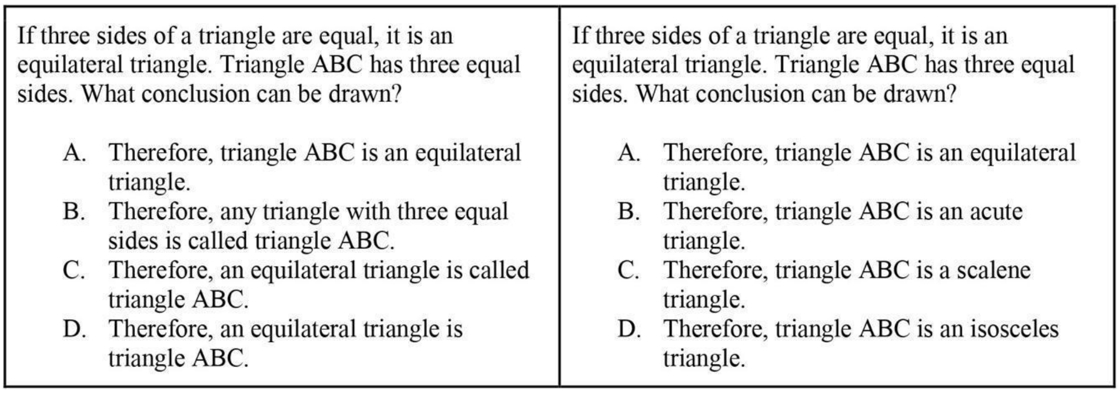
One of the primary metrics used in item analysis is the DIF I. DIF I is defined as the percentage of students who got an item correctly [35], [60]. It demonstrates how easy or hard an item is and ranges from 0 to 1. Values leaning towards 1 indicate that the item is becoming easier and that a higher proportion of individuals can answer it correctly [61]. Ideal items should have an average level of difficulty, with DIF I between 0.3 to 0.7. Previous studies [25], [26], [62] reported that the most effective items had difficulty indices clustered within this ideal range.

b. Discrimination index

Apart from item difficulty, the ability of an item to differentiate between high- and low-performing students must also be considered. This is assessed using DI. It compares the proportion of students who answer an item correctly among the top and bottom scorers [23], [33]. For example, students who scored high in the overall test should, logically, be able to answer the item correctly, while those who scored poorly should struggle with it. If in that item, more students in the low-performing group got it correctly than in the upper group, then the item is not able to discriminate between the two groups. Therefore, it has low discriminating power. Ideal items should have a DI greater than 0.29 [63], though items with DI between 0.19 to 0.29 were still considered acceptable but are recommended for revision [62]. Earlier researches [29], [64], [65] collectively suggests that effective items are typically characterized by discrimination indices falling within this range. Items whose DI are below 0.19—including the negative values—should either be removed or completely revised.

c. Distractor efficiency

The final metric that offers deeper insight into how individual options in MCQs affect item quality is DE. Several studies have specifically examined DE, emphasizing its role in enhancing the overall quality of assessment items. For example, Kumar *et al.* [26] proposed an automated system that generates distractors for MCQs in technical domains using both structured and unstructured data. Ali *et al.* [66] found that using common incorrect responses from free-response items to craft more effective MCQ distractors enhances the validity and reliability of the test item. In MCQs, incorrect options—called distractors—are categorized as functional distractor (FD) or non-functional distractor (NFD). A well-crafted distractor should appear like a correct answer and draw in students who are just guessing or who do not fully understand the material. An example is shown in Figure 2, in which the distractors on the left side presented related conclusions about triangle ABC, making the choices more closely connected. In contrast, the distractors on the right focused on the different types of triangles, making the correct answer easier to identify. A FD is an incorrect option selected by ≥5% of students. Otherwise, it is non-functional [66]–[68]. Prior studies [26], [68]–[70] reveal FDs ranging from 70 to 95%, suggesting that distractors are working well in attracting students who only have partial understanding. In addition, DE is determined for each test item on the basis of the number of NFDs and ranges from 0 to 100%. If an item contains three, two, one, or nil NFD, then DE will be 0% (poor), 33.33% (moderate), 66.6% (good), and 100% (excellent), respectively [26].



If three sides of a triangle are equal, it is an equilateral triangle. Triangle ABC has three equal sides. What conclusion can be drawn? A. Therefore, triangle ABC is an equilateral triangle. B. Therefore, any triangle with three equal sides is called triangle ABC. C. Therefore, an equilateral triangle is called triangle ABC. D. Therefore, an equilateral triangle is triangle ABC.	If three sides of a triangle are equal, it is an equilateral triangle. Triangle ABC has three equal sides. What conclusion can be drawn? A. Therefore, triangle ABC is an equilateral triangle. B. Therefore, triangle ABC is an acute triangle. C. Therefore, triangle ABC is a scalene triangle. D. Therefore, triangle ABC is an isosceles triangle.

Figure 2. Comparison of MCQ options for the same test item

2.4.5. Reliability coefficient

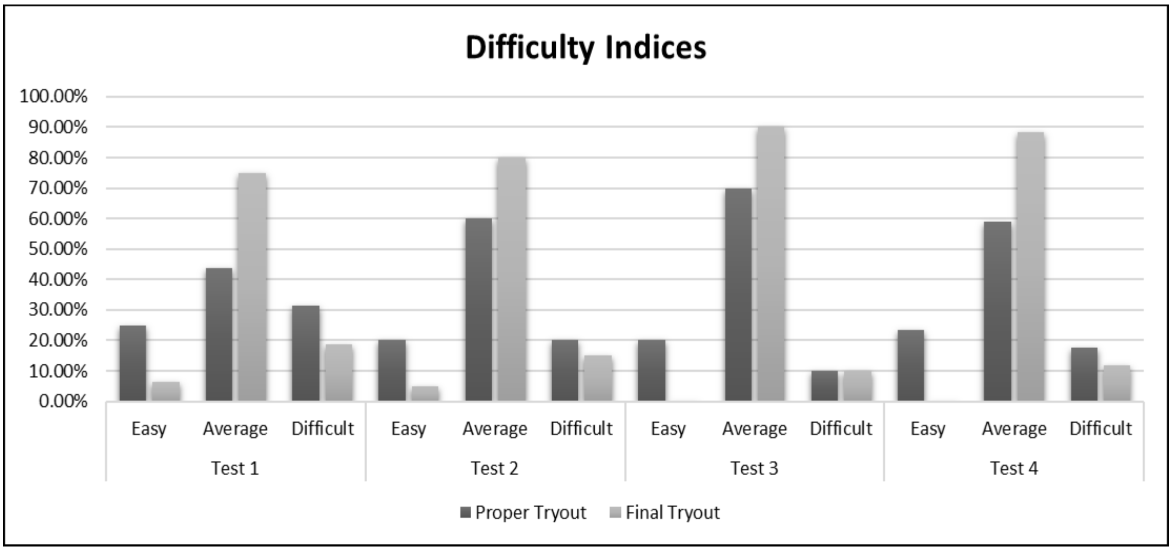
Reliability is defined as the consistency of scores obtained by an individual at different times on the same test, scored by the same or different examiners at the same or different times. For a test with binary variables (e.g., true/false, MCQs), the reliability coefficient is computed using the KR-20 formula [71]–[73]. Quaigrain and Arhin [74] note that teacher-made assessment tools generally have a value of 0.80 or lower. A KR-20 value above 0.70 is considered excellent; values between 0.6 to 0.7 are acceptable; values from 0.5 to 0.59 are deemed poor; and any value below 0.50 is regarded as unacceptable [40]. However, for classroom assessment tools, a KR-20 value between 0.50 to 0.60 is considered acceptable [26], [75].

3. RESULTS AND DISCUSSION

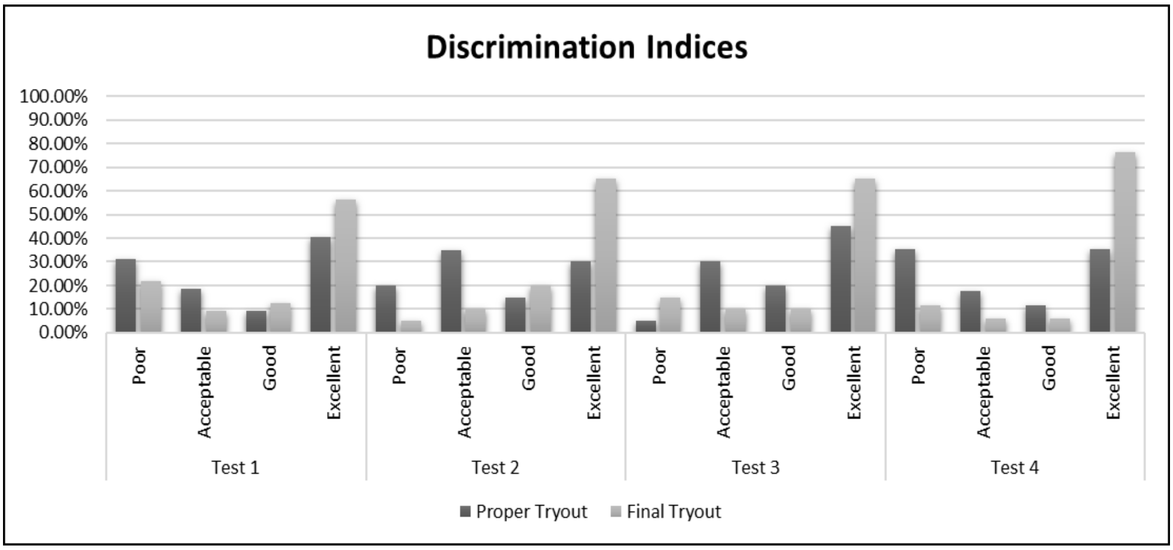
3.1. Comparative analysis of DIF I and DI between proper and final try-outs of the tests

To evaluate the effectiveness of item revisions, a comparative analysis in terms of DIF I, along with DI, was conducted between the proper try-out and the final try-out, as illustrated in Figure 3. Graphical comparison between proper and final try-outs of the tests are shown in Figure 3 (a) in terms of DIF I and Figure 3 (b) in terms of DI. Table 5 presents the DIF I and DI from Tests 1 to 4 across both phases. Across all four tests the proportion of easy items decreased substantially while average items also increased substantially in terms of DIF I. Meanwhile, difficult items slightly decreased in most tests except for Test 3 where it remained constant. This shift toward a greater number of items with an average level of difficulty better calibrates the tests to the intended level of competency, ensuring that items are neither too easy nor overly challenging for the students.

In addition, an improvement in item discrimination was evident in the final try-out, as the percentage of items classified as excellent discriminators increased across all tests. There was a reduction in poor discriminators in Tests 1, 2, and 4, suggesting that the revised test is more effective at distinguishing between students of different ability levels. In Test 3, there was an increase in poor discriminators and a decrease in both acceptable and good discriminators, but items with excellent discrimination improved substantially from 45% to 65%. From the proper to the final try-outs, the revised items improved, exhibiting a greater proportion of items with average difficulty and improved capacity to differentiate between varying levels of student performance across all tests.



(a)



(b)

Figure 3. Graphical comparison between proper and final try-outs of the tests in terms of (a) DIF I and (b) DI

Table 5. Comparative analysis of DIF I and DI between proper and final try-outs of the tests

Test instruments	DIF I			DI		
	Parameters	Proper tryout (%)	Final tryout (%)	Parameters	Proper tryout (%)	Final tryout (%)
Test 1	Easy	25	6.25	Poor	31.25	21.875
	Average	43.75	75	Acceptable	18.75	9.375
	Difficult	31.25	18.75	Good	9.38	12.5
				Excellent	40.625	56.25
Test 2	Easy	20	5	Poor	20	5
	Average	60	80	Acceptable	35	10
	Difficult	20	15	Good	15	20
				Excellent	30	65
Test 3	Easy	20	0	Poor	5	15
	Average	70	90	Acceptable	30	10
	Difficult	10	10	Good	20	10
				Excellent	45	65
Test 4	Easy	23.53	0	Poor	35.29	11.76
	Average	58.82	88.24	Acceptable	17.65	5.88
	Difficult	17.65	11.76	Good	11.77	5.88
				Excellent	35.29	76.47

3.2. Distribution of test items according to difficulty index in the final try-out

Table 6 shows that the majority of items across all four tests fell within the average level of difficulty range with 75% in Test 1, 80% in Test 2, 90% in Test 3, and 88.24% in Test 4. This was followed by items in the difficult range comprising 18.75% in Test 1, 15% in Test 2, 10% in Test 3, and 11.76% in Test 4. Only a minimal proportion of items were categorized as easy with 6.25% in Test 1, 5% in Test 2, and none in Tests 3 and 4. These results are consistent with those previous studies [65], [69], [70], where 81%, 70%, and 85% of the items were of average level of difficulty, respectively. The findings suggest that the majority of the assessment items presented an average level of difficulty for students with mean DIF I (SD) ranging from 51.47% (23.43) to 48.01 (16.95), indicating an appropriate alignment between item complexity and student ability. Compared to Rezigalla *et al.* [33] where the reported mean DIF I (SD) is 37.5 (19.1) and was considerably leaning towards the difficult range, the reported mean DIF I in our study fall within the generally accepted average difficulty range which is recommended for classroom assessments. Our study supports the findings of Konakcı [35]. From the perspective of IRT, this suggests that the items are well-targeted to the ability level of the students. Moreover, Test 1 showed the highest proportion of easy and difficult items. Greater variability in item difficulty was also observed among the difficult items. Despite prior revisions, items classified as difficult or easy were discarded after the final try-out as their performance continued to fall outside acceptable parameters. This decision aligns with established best practices to maintain assessment validity and reliability. This is supported by the study [65] which found a negative correlation between DIF I and DI, indicating that when DIF I increases, the DI decreases. As the test item becomes easier or more difficult, it would lessen its ability to discriminate between high and low performing students. Therefore, these items should be removed.

Table 6. Distribution of test items according to DIF I in the final try out

Test instruments	Parameters	DIF I		
		<0.30 (difficult)	0.30-0.70 (average)	>0.70 (easy)
Test 1	No. of items (%)	6 (18.75%)	24 (75%)	2 (6.25%)
	DIF I (mean±SD)	21.58±44.81	51.47±23.43	72.30±0
Test 2	No. of items (%)	3 (15%)	16 (80%)	1 (5%)
	DIF I (mean±SD)	22.65±41.02	45.42±21.80	79±0
Test 3	No. of items (%)	2 (10%)	18 (90%)	0
	DIF I (mean±SD)	26.08±12.69	47.66±22.09	
Test 4	No. of items (%)	2 (11.76%)	15 (88.24%)	0
	DIF I (mean±SD)	21.76±31.57	48.01±16.95	
Decision		Discard	Retain	Discard

3.3. Distribution of test items according to the discrimination index in the final try-out

Table 7 reveals that the results across the four tests in terms of DI showed that the majority of the items were clustered around an excellent DI with 56.25% in Test 1, 65% in Tests 2, 3, and 4. The mean DI (SD) of excellent discriminators range from 50.42 (17.04) to 57.04 (13.38). The results indicate a strong differentiating power between high and low performers. Test 1 had the highest proportion of items with a poor DI. Higher levels of variability were also observed among poor discriminators, particularly for

Tests 1, 3, and 4. Additionally, it is noteworthy that 5% of the items in Tests 2 and 15.39% in Test 4 have a negative DI, indicating that there are more students in the lower group who answered the items correctly than in the upper group. This observation is similar to the findings of [45], [65], [76] who reported that 8.4%, 9%, and 7.5% of the items in health science courses, respectively, exhibited negative DI values. This is indicative of a defective item, which can be attributed to insufficient coverage or emphasis of the course content. Given that students were drawn from different campuses, variations in teaching quality or content coverage across campuses may have influenced item responses. According to IRT, items that are not well-aligned with the target construct or are influenced by external factors can result in biased estimates of ability. For future test design, items with negative DI should be carefully reviewed for ambiguity, miskeyed answers, or ineffective distractors, and either revised or discarded as such items confuse high performers, benefit low performers, reduce reliability, and threaten validity. Addressing these issues will ensure that the test more accurately measures student learning [32], [77].

Table 7. Distribution of test items according to the DI in the final try-out

Test instruments	Parameters	DI			
		<.19 (poor)	0.19-0.29 (acceptable)	0.30-0.39 (good)	>0.39 (excellent)
Test 1	No. of items (%)	7 (21.875%)	3 (9.375%)	4 (12.5%)	18 (56.25%)
	DI (mean±SD)	11.10±69.01	23.02±19.50	35.07±5.90	52.16±17.37
Test 2	No. of items (%)	1 (5%)	2 (10%)	4 (20%)	13 (65%)
	DI (mean±SD)	-6±0	24.30±14.36	34.15±8.99	53.25±19.21
Test 3	No. of items (%)	3 (15%)	2 (10%)	2 (10%)	13 (65%)
	DI (mean±SD)	11.51±55.52	27.70±1.84	33.81±15.05	50.42±17.04
Test 4	No. of items (%)	2 (15.39%)	1 (5.88%)	1 (5.88%)	13 (65%)
	DI (mean±SD)	-7.39±-276.11	25±0	37±0	57.04±13.38
Decision		Discard	Retain	Retain	Retain

3.4. Distribution of distractors and categorization of MCQs according to distractor efficiency

A standard four-option MCQ format was used in all tests, comprising one correct answer and three distractors per item. As presented in Table 8, the majority of distractors across all four tests were functional, as evidenced by their choice frequencies meeting or exceeding the 5% threshold ranging from 95% to 100%. The results are higher compared to other studies [15], [26], [65], [69] and where FDs range from 59.1% to 73% but consistent with Rao *et al.* [70], where FDs was 95%. In terms of DE, 28 (87.5%) out of the 32 items in Test 1 achieved 100% DE and 4 (12.5%) achieved 66.6% DE with a mean of 95.825. In Test 2, 17 (85%) of the 20 items attained 100% DE and 15% achieved 66.6% DE with a mean of 94.99. Finally, all the items in Tests 3 and 4 have 100% DE. This indicates that the distractors are successfully engaging students with partial understanding. Compared to the other studies [15], [26], [34], [39], [53], [69], the results of the current study showed higher DE. Although crafting effective distractors is a rigorous and time-consuming process, it ensures that test items are more discriminative and better aligned with measurement goals that contribute to the item's overall discriminative power, a key parameter in IRT.

In the field of medicine, test items with DE less than 100% are often removed due to potential implications for patient safety [45]. Contrastingly, the context of the present study—a non-clinical discipline—allows for a more flexible approach. Retaining these items can help identify areas of student misunderstanding, inform instructional modifications, and support formative assessment practices without causing any direct real-world risks.

Table 8. Distribution of distractors and categorization of MCQs according to DE

Parameters	Test 1	Test 2	Test 3	Test 4
No. of MCQ items	32	20	20	17
Total no. of distractors	96	60	60	51
No. of FDs	92 (95.83%)	57 (95%)	60 (100%)	51 (100%)
No. of NFDs	4 (4.17%)	3 (5%)	0 (0.00%)	0 (0.00%)
MCQs with DE 100%	28 (87.5%)	17 (85%)	20 (100%)	17 (100%)
MCQs with DE 66.6%	4 (12.5%)	3 (15%)	0	0
Mean	95.825	94.99	100	100

3.5. Descriptive results and reliability coefficients for Tests 1 to 4 after refinement

Table 9 shows the questionnaire's descriptive results, normality, and reliability coefficients after several iterations. The remaining number of items across all tests range from 68.75% to 88.24%, where Test 4 has the lowest number of deleted items. The mean percentage scores of the respondents range from 42.94 to 50.95, which indicates that the means are relatively close to each other in the fourth and fifth percentiles.

The result fell within the threshold reported by other studies [26], [45], [68], [74], [78]. Based on the values, most of the respondents were similar in terms of their ability to deal with word problems and patterns. The relative standard deviations of the tests range from 39.43 to 48.25, indicating a wide variation in respondents' performances. This suggests that some students were doing well, especially in Test 1, while others were struggling, especially in Tests 2 and 4. This is supported by the consistent results of positive skewness and negative kurtosis values. Each test item was screened using skewness and kurtosis to determine the normality of item distribution [79], [80], and values less than ±2 are considered as a normal distribution [75]. Skewness values range from 0.17 to 0.45, suggesting that most of the scores are on the second quadrant. This implies that most of the respondents' scores were below the mean. This is aligned with the mean test scores across all four tests, resulting in scores less than 75%, which implies that respondents find the tests challenging. Kurtosis describes the extent to which data values cluster around the mean or deviate far from it [81].

From the test results, it has negative kurtosis values, suggesting a flat distribution with fewer high scores than the average. The combination of positive skewness and negative kurtosis across all tests suggests a consistent trend of generally below average student performance with moderate to high variability. This pattern raises concerns about the alignment between the assessments and student's preparedness. The test, especially Tests 2 and 4, may benefit from a review item difficulty and CV to ensure they accurately and fairly measure the intended learning outcomes. The results also align with the standard deviation results that few students were doing well in the tests. Standard error of measurement (SEM) was used to determine the precision and consistency of the test takers scores [82]. Reduced standard errors for large samples usually increase the kurtosis and skewness values [83]. The SEM yielded lower values, implying a higher precision in estimating true student performance, and this is supported by the KR-20 values. Tests 1 and 4 demonstrate the strongest reliability with KR-20 values of 0.78 and 0.72, respectively. Tests 2 and 3 were lower compared to Tests 1 and 4. This suggests that the tests may have been more difficult or less accessible to students, resulting in the reduction of internal consistency. Given the KR-20 values of Tests 2 and 3, literature supported that these tests have moderately good reliability with values of 0.68 and 0.54, respectively [84]. Therefore, all four tests establish good reliability coefficients and precision that may be used in assessing students' problem-solving and reasoning skills.

Table 9. Descriptive results and reliability coefficients for Tests 1 to 4 after refinement

Parameters	Test 1	Test 2	Test 3	Test 4
No. of items (%)	22 (68.75)	16 (80)	17 (85)	15 (88.24)
No. of students	515	525	514	513
Mean percentage test score	50.95	42.94	47.71	45.73
RSD	39.43	47.16	40.07	48.25
Median	11	6	8	6
Range of test scores	1-22	1-15	1-17	1-15
Skewness	0.17	0.43	0.30	0.45
Kurtosis	-0.74	-0.6	-0.33	-0.78
Kuder-Richardson 20	0.78	0.68	0.54	0.72
SEM	2.09	1.83	2.21	1.74

4. PRACTICAL RECOMMENDATION

Based on the findings of this study, it is recommended that mathematics teachers adopt the validated curriculum-aligned MMW problem-solving tests (MMW PST) as a supplementary tool to evaluate students' problem-solving skills. Doing so, can help ensure a fair, reliable, and evidenced-based measurement of student learning outcomes. In addition, they are encouraged to replicate the test development and validation process used in this study when designing their own tests. This can improve the quality of classroom tests by ensuring that each item is aligned with learning outcomes and supported by empirical evidence of validity and reliability. Furthermore, teachers should revise test items guided by the results of the item analysis. To guide the revision process, recommended actions prior to the final try-out are presented in Table 10.

Difficult items can be revised by simplifying the language or presenting a clearer context. For example, when assessing problem-solving with the tower of Hanoi, adding a diagram of the pegs and disks can help students visualize and focus on logical reasoning rather than struggling to interpret the verbal description. On the other hand, easy items can be revised by increasing their cognitive demand or replacing distractors with more plausible ones such as the example in Figure 2. To strengthen these efforts, school administrators may consider integrating empirically validated classroom assessments into their regular evaluation systems to enhance the quality and fairness of student performance measurement. It is also recommended to conduct item development training to enhance teachers' capacity in developing high-quality test items aligned with curriculum objectives. As a final recommendation, a bridging program is suggested for incoming freshmen students, especially those who are non-science, technology, engineering, mathematics (non-STEM) graduates. This will prepare them to cope with the demands of tertiary-level mathematics courses. Table 11 provides an overview of sample topics included in the bridging program. For students who continue to perform below the average and experience persistent difficulties, teachers can intervene through peer tutorials or supplementary enrichment activities, such as guided practice sessions or one-on-one consultations, to address individualized learning gaps.

Table 10. Recommended actions to be taken

Parameters	Index values	Interpretation	Problem	Action to be taken
DIF I	<0.30	Difficult	A limited proportion of examinees answered it correctly	Revise
	0.30-0.70	Average	None	Retain
	>0.70	Easy	Nearly all examinees answered the item correctly	Revise
DI	<0.19	Poor	The item showed little difference in the performance between top- and bottom-scorers	
	0.19-0.29	Acceptable	None	Revise
	0.30-0.39	Good	None	Retain
	>0.39	Excellent	None	Retain

Table 11. Sample bridging program topics

Main topics	Session titles
Number sense and arithmetic refresher	Session 1: diagnostic test and orientation Session 2: integers, fractions, and decimals Session 3: percentages and ratios Session 4: order of operations
Algebra fundamentals	Session 5: variables and algebraic expressions Session 6: linear equations and inequalities Session 7: word problems
Geometry	Session 8: perimeter, area, volume Session 9: angles and basic trigonometry
Critical thinking and recreational problem solving	Session 10: logical reasoning Session 11: recreational problem-solving Session 12: problem solving strategies Session 13: post-test

5. CONCLUSION

Prior studies on test construction and item analysis have highlighted the significance of validating classroom assessment tools to accurately measure students' learning and inform instructional decisions. However, research in this area has largely been limited to courses within the medical or science domains. To address this gap, the study aimed to develop and validate a comprehensive curriculum-based mathematics PST that effectively measures problem-solving and reasoning skills among tertiary-level students.

The results showed that the developed tests demonstrated strong content and face validity with CV achieving a high level of inter-rater reliability. Although face validity registered slightly below the recommended threshold, it provides a solid foundation for further enhancement. One of the salient findings of the study is that data-driven refinement of test items guided by the results of the proper try-out enhanced the item quality by increasing both difficulty and discrimination indices in the final try-out. Furthermore, it was revealed that the majority of the test items fell within the acceptable difficulty level, which aligns with the ability levels of the students, as predicted by IRT. This suggests that the items effectively discriminate between students of varying abilities, with difficulty levels calibrated to provide a reliable assessment of their problem-solving and reasoning skills. Additionally, the analysis of distractors showed that the majority were functioning effectively, successfully attracting students and contributing to the overall effectiveness of the test in distinguishing between correct and incorrect responses. Finally, the tests demonstrated acceptable to excellent reliability coefficients. Together, these findings confirmed that the developed tests are both valid and reliable for assessing problem-solving skills, and suitable for classroom use. The findings suggest that employing tests that have undergone this process ensures confidence in its validity, thereby providing reliable data that supports a fairer and more accurate evaluation of student performance and informs instructional decision-making. Through the use of validated assessment tools, instructors can strengthen students' 21st century problem-solving abilities and support meaningful learning.

As with any research, this study has its limitations. First, it affirms the limitations inherent in the current methodology since it only provides a snapshot of how items perform within a population over a specific time. The generalizability of the findings may be confined to similar institutions, as factors such as campus-specific contexts, student demographics, and variations in teaching quality may have influenced the outcomes. These potential confounding variables highlight the need for caution in extending the results to broader settings. Second, the use of purposive sampling limits the representativeness of the participants, which may restrict the extent to which the results can be applied to a wider student population. Future studies may address these limitations by including multiple institutions, larger and more diverse student populations, and additional variables to provide a more comprehensive understanding of teacher-made assessments.

To establish more valid and more generalizable results, future works should aim to replicate this study and conduct time-series correlational analysis. Longitudinal studies are also recommended to track item performance over time and observe how students' mastery develops across different cohorts. Lastly, future studies should not only identify which problem-solving-related topics students find most challenging but also incorporate qualitative insights, such as interviews or focus group discussions, to better understand the underlying reasons for students' low scores.

ACKNOWLEDGMENTS

The authors extend their gratitude to individuals whose contributions were instrumental to the successful completion of this study.

FUNDING INFORMATION

The current study received funding support from Cebu Technological University under the Center for Cloud Computing, Big Data, and Artificial Intelligence.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Purity V. Mata	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Quennie Rose H. Abao	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Roxanne Kaye S. Bihasa	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓
Glenda N. Kiamco	✓			✓		✓	✓	✓		✓				✓
Anna Marie C. Neiz						✓	✓			✓				✓

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject administration
Va	: **Va**lidation	O	: Writing - **O**riginal Draft	Fu	: **Fu**nding acquisition
Fo	: **Fo**rmal analysis	E	: Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

INFORMED CONSENT

Participation was voluntarily agreed to, compliant with data protection regulations.

DATA AVAILABILITY

The data that supports the findings of this study are readily available from the corresponding author, [PVM], upon reasonable request. The data, which contain information that could compromise the privacy of research participants, are not publicly available due to certain restrictions.

REFERENCES

[1] T. Abate and E. Mishore, "Alignment analysis between teacher-made tests with the learning objectives in a selected school of central regional state of Ethiopia," *Heliyon*, vol. 10, no. 11, p. e31869, Jun. 2024, doi: 10.1016/j.heliyon.2024.e31869.

[2] OECD, *PISA 2022 Results (Volume I)*. Paris: OECD Publishing, 2023, doi: 10.1787/53f23881-en.

[3] I. N. Generelao, G. Ducanes, K. M. Yee, and C. David, "Teacher Education in the Philippines: Are We Meeting the Demand for Quality?" *Philippine Journal of Public Policy: Interdisciplinary Development Perspectives*, vol. 2022, pp. 1–65, 2022, doi: 10.54096/IENE4805.

[4] M. Mosia, F. O. Egara, F. A. Nannim, and M. Basitere, "Factors Influencing Students' Performance in University Mathematics Courses: A Structural Equation Modelling Approach," *Education Sciences*, vol. 15, no. 2, p. 188, Feb. 2025, doi: 10.3390/educsci15020188.

[5] T. Tsui and R. N. Khan, "Is mathematics a barrier for engineering?" *International Journal of Mathematical Education in Science and Technology*, vol. 54, no. 9, pp. 1853–1873, Oct. 2023, doi: 10.1080/0020739X.2023.2256319.

[6] A. Romo-Vázquez and M. Artigue, "Challenges for Research on Tertiary Mathematics Education for Non-specialists: Where Are We and Where Are We to Go?" in *Practice-Oriented Research in Tertiary Mathematics Education*, R. Biehler, M. Liebendörfer, G. Gueudet, C. Rasmussen, and C. Winsløw, Eds. Cham: Springer International Publishing, 2022, pp. 535–557, doi: 10.1007/978-3-031-14175-1_26.

[7] S. Wellberg, "Teacher-made tests: why they matter and a framework for analysing mathematics exams," *Assessment in Education: Principles, Policy & Practice*, vol. 30, no. 1, pp. 53–75, Jan. 2023, doi: 10.1080/0969594X.2023.2189565.

[8] Y. Oc and H. Hassen, "Comparing the Effectiveness of Multiple-Answer and Single-Answer Multiple-Choice Questions in Assessing Student Learning," *Marketing Education Review*, vol. 35, no. 1, p. 44, 2025, doi: 10.1080/10528008.2024.2417106.

[9] S. Nundy, L. A. Cooper, and K. S. Mate, "The Quintuple Aim for Health Care Improvement," *JAMA*, vol. 327, no. 6, pp. 521–522, Feb. 2022, doi: 10.1001/jama.2021.25181.

[10] A. H. Elgadal and A. A. Mariod, "Item Analysis of Multiple-choice Questions (MCQs): Assessment Tool for Quality Assurance Measures," *Sudan Journal of Medical Sciences*, vol. 16, no. 3, pp. 334–346, Sep. 2021, doi: 10.18502/sjms.v16i3.9695.

[11] N. Haycocks *et al.*, "Assessing the difficulty and long-term retention of factual and conceptual knowledge through multiple-choice questions: a longitudinal study," *Advances in Medical Education and Practice*, vol. 15, pp. 1217–1228, Dec. 2024, doi: 10.2147/AMEP.S478193.

[12] C. D. Riggs, S. Kang, and O. Rennie, "Positive Impact of Multiple-Choice Question Authoring and Regular Quiz Participation on Student Learning," *CBE—Life Sciences Education*, vol. 19, no. 2, p. ar16, Jun. 2020, doi: 10.1187/cbe.19-09-0189.

[13] S. P. Stevens, S. W. Palocsay, and L. J. Novoa, "Practical Guidance for Writing Multiple-Choice Test Questions in Introductory Analytics Courses," *INFORMS Transactions on Education*, vol. 24, no. 1, pp. 51–69, Sep. 2023, doi: 10.1287/ited.2022.0274.

[14] D. D. Shin *et al.*, "Science for all: Boosting the science motivation of elementary school students with utility value intervention," *Learning and Instruction*, vol. 60, pp. 104–116, Apr. 2019, doi: 10.1016/j.learninstruc.2018.12.003.

[15] S. Hassan and R. Hod, "Use of Item Analysis to Improve the Quality of Single Best Answer Multiple Choice Question in Summative Assessment of Undergraduate Medical Students in Malaysia," *Education in Medicine Journal*, vol. 9, no. 3, pp. 33–43, Sep. 2017, doi: 10.21315/eimj2017.9.3.4.

[16] N.-G. Ugwu and S. O. Mkpuma, "Ensuring Quality in Education: Validity of Teacher-made Language Tests in Secondary Schools in Ebonyi State," *American Journal of Educational Research*, vol. 7, no. 7, pp. 518–523, 2019.

[17] P. Kissi, D. Baidoo-Anu, E. Anane, and R. K. Annan-Brew, "Teachers' test construction competencies in examination-oriented educational system: Exploring teachers' multiple-choice test construction competence," *Frontiers in Education*, vol. 8, p. 1154592, Apr. 2023, doi: 10.3389/feduc.2023.1154592.

[18] D. Lebagi, Sumardi, and Sudjoko, "The Quality of Teacher-made test in EFL Classroom at the Elementary School and Its Washback in the Learning," *Journal of English Education*, vol. 2, no. 2, pp. 97–104, 2017, doi: 10.31327/jee.v2i2.289.

[19] L. G. Jimoh and M. A. Aremu, "Prevalence of item-writing flaws in teacher-made economics multiple choice questions in secondary schools in Ekiti North Senatorial District, Nigeria," *Journal of Psychometry and Assessment Techniques*, vol. 1, no. 1, pp. 112–118, 2020.

[20] T. R. Guskey and L. J. Link, "Exploring the factors teachers consider in determining students' grades," *Assessment in Education: Principles, Policy & Practice*, vol. 26, no. 3, pp. 303–320, May 2019, doi: 10.1080/0969594X.2018.1555515.

[21] S. Musliha, D. Sudana, and Y. Wirza, "The Analysis of Higher Order Thinking Skills (HOTs) in the Test Questions Constructed by English Teachers," in *Proceedings of the Fifth International Conference on Language, Literature, Culture, and Education (ICOLLITE 2021)*, 2021, pp. 610–617, doi: 10.2991/assehr.k.211119.095.

[22] E. Sriyono, Sardi, and W. H. Putri, "Construction Service Competence Test and Training Center (CSCTTC): Small Business Unit Based on the Potential and Intellectual Creativity of the University," in *Proceedings of the 3rd International Conference on Banking, Accounting, Management and Economics (ICOBAME 2020)*, 2020, pp. 341–345, doi: 10.2991/aebmr.k.210311.068.

[23] S. S. Matazu and E. Julius, "Item Analysis: A Veritable Tool for Effective Assessment in Teaching and Learning," *Journal of Education and Practice*, vol. 12, no. 21, pp. 22–28, Jul. 2021, doi: 10.7176/JEP/12-21-04.

[24] Q. Liu, N. Wald, C. Daskon, and T. Harland, "Multiple-choice questions (MCQs) for higher-order cognition: Perspectives of university teachers," *Innovations in Education and Teaching International*, vol. 61, no. 4, pp. 802–814, Jul. 2024, doi: 10.1080/14703297.2023.2222715.

[25] D. D. Kurniawan, A. Syifa, N. Huda, and M. Kusuma, "Item Analysis of Teacher Made Test in Biology Subject," in *Proceedings of the 5th International Conference on Current Issues in Education (ICCIE 2021)*, 2022, pp. 312–317, doi: 10.2991/assehr.k.220129.057.

[26] D. Kumar, R. Jaipurkar, A. Shekhar, G. Sikri, and V. Srinivas, "Item analysis of multiple choice questions: A quality assurance test for an assessment tool," *Medical Journal Armed Forces India*, vol. 77, pp. S85–S89, 2021, doi: 10.1016/j.mjafi.2020.11.007.

[27] N. Hartati and H. P. S. Yogi, "Item Analysis for a Better Quality Test," *English Language in Focus (ELIF)*, vol. 2, no. 1, pp. 59–70, Sep. 2019, doi: 10.24853/elif.2.1.59-70.

[28] R. Wess, H. Klock, H.-S. Siller, and G. Greefrath, "Test Quality," in *Measuring Professional Competence for the Teaching of Mathematical Modelling: A Test Instrument*, R. Wess, H. Klock, H.-S. Siller, and G. Greefrath, Eds. Cham: Springer International Publishing, 2021, pp. 77–84, doi: 10.1007/978-3-030-78071-5_4.

[29] L. I. E. Carolino *et al.*, "Electricity Concepts' Test Construction, Validation, and Item Analysis for Senior High School General Physics 2," *Southeast Asian Journal of Science and Technology*, vol. 6, no. 1, pp. 47–53, 2021.

[30] H. Y. Lee, S. J. Yune, S. Y. Lee, S. Im, and B. S. Kam, "The impact of repeated item development training on the prediction of medical faculty members' item difficulty index," *BMC Medical Education*, vol. 24, no. 1, p. 599, May 2024, doi: 10.1186/s12909-024-05577-x.

[31] M. Al-lawama and B. Kumwenda, "Decreasing the options' number in multiple choice questions in the assessment of senior medical students and its effect on exam psychometrics and distractors' function," *BMC Medical Education*, vol. 23, no. 1, p. 212, Apr. 2023, doi: 10.1186/s12909-023-04206-3.

[32] N. A. Zubairi, T. S. AlAhmadi, M. H. Ibrahim, M. A. Hegazi, and F. U. Gadi, "Effective use of Item Analysis to improve the Reliability and Validity of Undergraduate Medical Examinations: Evaluating the same exam over many years: A different approach," *Pakistan Journal of Medical Sciences*, vol. 41, no. 3, pp. 810–815, Feb. 2025, doi: 10.12669/pjms.41.3.10693.

[33] A. A. Rezigalla *et al.*, "Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items," *BMC Medical Education*, vol. 24, no. 1, p. 445, Apr. 2024, doi: 10.1186/s12909-024-05433-y.

[34] M. W. Gebremichael, B. Baraki, M.-A. Mehari, and B. Assalfew, “Item analysis of multiple choice questions from assessment of health sciences students, Tigray, Ethiopia,” *BMC Medical Education*, vol. 25, no. 1, p. 441, Mar. 2025, doi: 10.1186/s12909-025-06904-6.

[35] S. Konakcı, “Item Analysis in Multiple Choice Questions: A Study on Question Difficulty and Authors’ Evaluation,” *Journal of Basic and Clinical Health Sciences*, vol. 8, no. 2, pp. 490–497, May 2024, doi: 10.30621/jbachs.1427464.

[36] D. A. Tan, C. C. Cordova, I. P. B. Saligumba, and L. L. B. Segumpan, “Development of Valid and Reliable Teacher-Made Tests for Grade 10 Mathematics,” *International Journal of English and Education*, vol. 8, no. 1, pp. 62–82, 2019.

[37] S. A. Widodo, I. Ibrahim, W. Hidayat, S. Maarif, and F. Sulistyowati, “Development of Mathematical Problem Solving Tests on Geometry for Junior High School Students,” *Jurnal Elemen*, vol. 7, no. 1, pp. 221–231, Jan. 2021, doi: 10.29408/jel.v7i1.2973.

[38] M. M. Rahman, “21st Century Skill ‘Problem Solving’: Defining the Concept,” *Asian Journal of Interdisciplinary Research*, vol. 2, no. 1, pp. 64–74, Mar. 2019, doi: 10.34256/ajir1917.

[39] National Privacy Commission, *Republic Act 10173 - Data Privacy Act of 2012*. 2012. [Online]. Available: https://privacy.gov.ph/data-privacy-act/

[40] A. Wutich, M. Beresford, and H. R. Bernard, “Sample Sizes for 10 Types of Qualitative Data Analysis: An Integrative Review, Empirical Guidance, and Next Steps,” *International Journal of Qualitative Methods*, vol. 23, pp. 1–14, Jan. 2024, doi: 10.1177/16094069241296206.

[41] T. O. Nyumba, K. Wilson, C. J. Derrick, and N. Mukherjee, “The use of focus group discussion methodology: Insights from two decades of application in conservation,” *Methods in Ecology and Evolution*, vol. 9, no. 1, pp. 20–32, Jan. 2018, doi: 10.1111/2041-210X.12860.

[42] N. H. Azizan, Z. Mahmud, and A. Rambli, “Accuracy and Bias of the Rasch Rating Scale Person Estimates using Maximum Likelihood Approach: A Comparative Study of Various Sample Sizes,” *Journal of Physics: Conference Series*, vol. 2084, no. 1, p. 012006, Nov. 2021, doi: 10.1088/1742-6596/2084/1/012006.

[43] B. S. P. Earnest, P. Bhargava, A. K. Das, D. M. T. M. Azhar, N. M. Ibrahim, and R. G. Sirisinghe, “Transforming Teaching-Learning Culture by Appropriate Use of Discrimination Index in Item Analysis,” in *Redesigning Learning for Greater Social Impact: Taylor’s 9th Teaching and Learning Conference 2016 Proceedings*, 2018, pp. 159–166, doi: 10.1007/978-981-10-4223-2_14.

[44] I. Kennedy, “Sample Size Determination in Test-Retest and Cronbach Alpha Reliability Estimates,” *British Journal of Contemporary Education*, vol. 2, no. 1, pp. 17–29, Feb. 2022, doi: 10.52589/BJCE-FY266HK9.

[45] A. I. O. Yahia, “Post-validation item analysis to assess the validity and reliability of multiple-choice questions at a medical college with an innovative curriculum,” *The National Medical Journal of India*, vol. 34, pp. 359–362, Jul. 2022, doi: 10.25259/NMJI_414_20.

[46] P. Miralles-Sánchez, C. J. Gómez-Carrasco, and J. Rodríguez-Medina, “Design and validation of two tools to observe and analyze history lessons in secondary education,” *Frontiers in Education*, vol. 8, Aug. 2023, doi: 10.3389/feduc.2023.1213358.

[47] A. Roebianto, S. I. Savitri, I. Aulia, A. Suciyana, and L. Mubarokah, “Content Validity: Definition and Procedure of Content Validation in Psychological Research,” *TPM - Testing, Psychometrics, Methodology in Applied Psychology*, vol. 30, no. 1, pp. 5–18, 2023.

[48] H. Taherdoost, “Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research,” *International Journal of Academic Research in Management (IJARM)*, vol. 5, no. 3, pp. 28–36, 2018.

[49] J. Wang, Y. Zheng, and Y. Zou, “Face validity and washback effects of the shortened PTE Academic: insights from teachers in Mainland China,” *Language Testing in Asia*, vol. 14, no. 1, p. 32, Aug. 2024, doi: 10.1186/s40468-024-00302-3.

[50] P. Ranganathan and C. Caduff, “Designing and validating a research questionnaire - Part 1,” *Perspectives in Clinical Research*, vol. 14, no. 3, pp. 152–155, Jul. 2023, doi: 10.4103/picr.picr_140_23.

[51] M. A. Rahman, M. S. B. Yusoff, N. S. Roslan, J. A.-M. Mohammad, and A. Ahmad, “Development and validation of the medical professionals resilience scale,” *BMC Health Services Research*, vol. 21, no. 1, p. 482, Dec. 2021, doi: 10.1186/s12913-021-06542-w.

[52] B. Shabankhani, “Assessing the inter-rater reliability for nominal, categorical and ordinal data in medical sciences,” *Arch Pharma Pract*, vol. 11, no. 4, pp. 144–148, 2020.

[53] I. Nurjannah and S. M. Siwi, “Guidelines for analysis on measuring interrater reliability of nursing outcome classification,” *International Journal of Research in Medical Sciences*, vol. 5, no. 4, pp. 1169–1175, Mar. 2017, doi: 10.18203/2320-6012.ijrms20171220.

[54] F. Madadizadeh, H. Ghafari, and S. Bahariniya, “Kappa Statistics: A Method of Measuring Agreement in Dental Examinations,” *The Open Public Health Journal*, vol. 16, no. 1, pp. 1–6, 2023, [Online]. Available: https://openpublichealthjournal.com/VOLUME/16/ELOCATOR/e18749445259818/FULLTEXT/

[55] C. Thompson, R. Byrne, J. Adams, and H. A. Vidgen, “Development, validation and item reduction of a food literacy questionnaire (IFLQ-19) with Australian adults,” *International Journal of Behavioral Nutrition and Physical Activity*, vol. 19, no. 1, p. 113, Sep. 2022, doi: 10.1186/s12966-022-01351-8.

[56] A. Gorham and J. Randall, *Classical Test Theory*. New York: Routledge, 2022, doi: 10.4324/9781138609877-REE26-1.

[57] R. Jabrayilov, W. H. M. Emons, and K. Sijtsma, “Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment,” *Applied Psychological Measurement*, vol. 40, no. 8, pp. 559–572, Nov. 2016, doi: 10.1177/0146621616664046.

[58] Y. O. Shogbesan, “Item characteristics of National Examination Council’s economics multiple-choice items: An item response theory exploration,” *Journal of Psychometric Research*, vol. 2, no. 1, pp. 8–18, 2024.

[59] M. Taufiq *et al.*, “Item Analysis and Teachers’ Factors in Designing a Test,” in *Proceedings of The 1st EAI Bukittinggi International Conference on Education, BICED 2019,* Bukitittinggi, West Sumatera, Indonesia, 2019, pp. 235–241, doi: 10.4108/eai.17-10-2019.2289772.

[60] S. Koçdar, N. Karadag, and M. D. Sahin, “Analysis of the difficulty and discrimination indices of multiple-choice questions according to cognitive levels in an open and distance learning context,” *TOJET: The Turkish Online Journal of Educational Technology*, vol. 15, no. 4, pp. 16–25, 2016.

[61] S. M. J. A. Marie and S. Edannur, “Relevance of Item Analysis in Standardizing an Achievement Test in Teaching of Physical Science in B.Ed Syllabus,” *Journal of Educational Technology*, vol. 12, no. 3, pp. 30–36, Dec. 2015, doi: 10.26634/jet.12.3.3743.

[62] D. Kheyami, A. Jaradat, T. Al-Shibani, and F. A. Ali, “Item Analysis of Multiple Choice Questions at the Department of Paediatrics, Arabian Gulf University, Manama, Bahrain,” *Sultan Qaboos University Medical Journal*, vol. 18, no. 1, pp. e68–74, Apr. 2018, doi: 10.18295/squmj.2018.18.01.011.

[63] J. S. Camara, “Development and Validation of a Criterion-Referenced Test in Qualitative Research for 11th graders,” *SSRN Electronic Journal*, pp. 1–10, 2021, doi: 10.2139/ssrn.3865856.

[64] J. I. Salim, “Discrimination and Difficulty Indices of a Senior High School Entrance Examination Using Classical Test Theory,” *Psychology and Education: A Multidisciplinary Journal*, vol. 6, no. 6, pp. 499–504, 2022. [Online]. Available: https://www.researchgate.net/publication/370415148

[65] N. Maulina and R. Novirianthy, “Item analysis and peer-review evaluation of specific health problems and applied research block examination,” *Jurnal Pendidikan Kedokteran Indonesia: The Indonesian Journal of Medical Education*, vol. 9, no. 2, pp. 131–137, Jul. 2020, doi: 10.22146/jpki.49006.

[66] S. H. Ali, P. A. Carr, and K. G. Ruit, “Validity and Reliability of Scores Obtained on Multiple-Choice Questions: Why Functioning Distractors Matter,” *Journal of the Scholarship of Teaching and Learning*, vol. 16, no. 1, pp. 1–14, Feb. 2016, doi: 10.14434/josotl.v16i1.19106.

[67] I. Burud, K. Nagandla, and P. Agarwal, “Impact of distractors in item analysis of multiple choice questions,” *International Journal of Research in Medical Sciences*, vol. 7, no. 4, pp. 1136–1139, Mar. 2019, doi: 10.18203/2320-6012.ijrms20191313.

[68] W. Mahjabeen *et al.*, “Difficulty Index, Discrimination Index and Distractor Efficiency in Multiple Choice Questions,” *Annals of Pakistan Institute of Medical Sciences*, vol. 13, no. 4, pp. 310–314, 2017, doi: 10.48036/apims.v13i4.9.

[69] A. P. Date, A. S. Borkar, R. T. Badwaik, R. A. Siddiqui, T. R. Shende, and A. V. Dashputra, “Item analysis as tool to validate multiple choice question bank in pharmacology,” *International Journal of Basic & Clinical Pharmacology*, vol. 8, no. 9, pp. 1999–2003, Aug. 2019, doi: 10.18203/2319-2003.ijbcp20194106.

[70] C. Rao, H. L. K. Prasad, K. Sajitha, H. Permi, and J. Shetty, “Item analysis of multiple choice questions: Assessing an assessment tool in medical students,” *International Journal of Educational and Psychological Researches*, vol. 2, no. 4, pp. 201–204, 2016, doi: 10.4103/2395-2296.189670.

[71] N. V. F. Liando, E. Serhalawan, and C. Wuntu, “Analysis of Teacher-Made Tests Used in Summative Evaluation at SMP Negeri 1 Tompaso,” *Jurnal Ilmiah Wahana Pendidikan*, vol. 7, no. 8, pp. 480–493, 2021, [Online]. Available: https://jurnal.peneliti.net/index.php/JIWP

[72] S. Fatimah, A. B. Elzamzami, and J. Slamet, “Item Analysis of Final Test for the 9th Grade Students of SMPN 44 Surabaya in the Academic Year of 2019/2020,” *JournEEL (Journal of English Education and Literature)*, vol. 2, no. 1, pp. 34–46, Jun. 2020, doi: 10.51836/journeel.v2i1.81.

[73] A. M. Obon and K. A. M. Rey, “Analysis of Multiple-Choice Questions (MCQs): Item and Test Statistics from the 2nd Year Nursing Qualifying Exam in a University in Cavite, Philippines,” *Abstract Proceedings International Scholars Conference*, vol. 7, no. 1, pp. 499–511, Dec. 2019, doi: 10.35974/isc.v7i1.1128.

[74] K. Quaigrain and A. K. Arhin, “Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation,” *Cogent Education*, vol. 4, no. 1, p. 1301013, Jan. 2017, doi: 10.1080/2331186X.2017.1301013.

[75] D. George and P. Mallery, *IBM SPSS Statistics Step by Step: A Simple Guide and Reference*. New York: Allyn and Bacon, 2014.

[76] A. S. Khare, A. Verma, A. A. Kadam, and K. Avachar, “Evaluation of Discrimination Index of Single Best Response Type of Multiple Choice Questions in Physiology by Item Analysis,” *Journal of Pharmaceutical Negative Results*, vol. 14, no. 2, pp. 1039–1044, 2023.

[77] R. Schmucker and S. Moore, “The Impact of Item-Writing Flaws on Difficulty and Discrimination in Item Response Theory,” Aug. 2025, [Online]. Available: http://arxiv.org/abs/2503.10533

[78] D. S. Christian, A. C. Prajapati, B. M. Rana, and V. R. Dave, “Evaluation of multiple choice questions using item analysis tool: a study from a medical institute of Ahmedabad, Gujarat,” *International Journal of Community Medicine and Public Health*, vol. 4, no. 6, pp. 1876–1881, May 2017, doi: 10.18203/2394-6040.ijcmph20172004.

[79] M. Szcześniak, A. H. Świątek, M. A. Świątek, and W. Rodzeń, “Positive downstream indirect reciprocity scale (PoDIRS–6): Construction and psychometric characteristics,” *Current Psychology*, vol. 41, no. 7, pp. 4379–4400, Jul. 2022, doi: 10.1007/s12144-020-00942-7.

[80] G. Hatem, J. Zeidan, M. Goossens, and C. Moreira, “Normality testing methods and the importance of skewness and kurtosis in statistical analysis,” *BAU Journal - Science and Technology*, vol. 3, no. 2, p. 7, Jun. 2022, doi: 10.54729/KTPE9512.

[81] S. Demir, “Comparison of Normality Tests in Terms of Sample Sizes under Different Skewness and Kurtosis Coefficients,” *International Journal of Assessment Tools in Education*, vol. 9, no. 2, pp. 397–409, Jun. 2022, doi: 10.21449/ijate.1101295.

[82] S. A. Livingston, J. Carlson, and B. Bridgeman, *Test Reliability—Basic Concepts, Research Memorandum No. RM-18-01*. Princeton, NJ: Educational Testing Service, 2018. [Online]. Available: https://www.cn.ets.org/Media/Research/pdf/RM-18-01.pdf

[83] A. Field, *Discovering Statistics Using IBM SPSS Statistics*, 6th ed. London: SAGE Publications Limited, 2024.

[84] A. Faradillah and L. Febriani, “Mathematical trauma students’ junior high school based on grade and gender,” *Infinity Journal*, vol. 10, no. 1, pp. 53–68, Jan. 2021, doi: 10.22460/infinity.v10i1.p53-68.

BIOGRAPHIES OF AUTHORS

Purity V. Mata is a licensed teacher and a dedicated assistant professor at Cebu Technological University, Carmen Campus. She holds a Doctorate in Development Education, a Master of Arts in Education with a major in Administration and Supervision, and a Bachelor’s degree in Psychology. She successfully passed the licensure examination for Psychometrician and is currently pursuing a Ph.D. in Psychology. Her research interests include but not limited to instrument development and validation, structural equation modeling, Filipino psychology, educational psychology, and social science research. Her contributions involve leading the study’s conceptualization, conducting the literature review, designing the methodological approach with emphasis on item analysis, and performing data analysis. She can be contacted at email: mata.purity@ctu.edu.ph; purity.mata@jgu.edu.in.

Quennie Rose H. Abao received her Doctorate in Development Education from Cebu Technological University, Danao Campus. She has been in the teaching profession for 9 years and is currently an associate professor at Cebu Technological University, Carmen Campus. She earned her Bachelor's degree in Secondary Education, major in Mathematics, and completed her Master of Arts in Education, major in Mathematics Education. Her current research interests include mathematics education, pedagogical strategies, classroom teaching, curriculum-based assessments, and interdisciplinary social science studies. She teaches mathematics major courses, actively engages in research and extension projects, and provides mentorship to students in higher education. She can be contacted at email: quenniehinaloc@gmail.com.

Roxanne Kaye S. Bihasa is a licensed professional mathematics teacher. Aside from holding a bachelor's degree in BSED in mathematics, she has a diploma in Hotel and Restaurant Management. She completed her Ph.D. in education major in Research Management and Development as a government-funded scholar. She has also taken units of coursework toward a doctoral degree in education with a focus on Educational Leadership and Management. Currently, she is the Chairperson of the College of Education, Arts, and Sciences at CTU Carmen Campus. She likes to work on instrument validation, and curriculum and learning assessment. She provides free research mentorship to undergraduate and graduate students. She can be contacted at email: roxannekayesantillan15@gmail.com.

Glenda N. Kiamco is an assistant professor teaching mathematics courses across all programs at Cebu Technological University-Carmen, Campus for over 30 years. She pursued her graduate studies in Mathematics Education at the University of the Visayas. She is passionate about teaching mathematics and supports the improvement of the quality of teaching and learning. She can be contacted at email: kiamco.glenda@ctu.edu.ph.

Anna Marie C. Neiz is a committed educator and dean of Instruction at Cebu Technological University, Carmen Campus, specializing in Mathematics and Educational Administration. She holds a Doctorate in Development Education, a Master of Arts in Education Major in Administration and Supervision, and a Bachelor of Science in Mathematics. With expertise in both pedagogy and leadership, she plays a key role in enhancing curriculum development, fostering academic excellence, and mentoring future educators and administrators at CTU. She can be contacted at email: anna.neiz@ctu.edu.ph.