JOURNAL OF INTERDISCIPLINARY PERSPECTIVES

ISSN Print: 2984-8288, ISSN Online: 2984-8385
Vol. 2, No. 11, pp. 362-372, November 2024

# Ensuring Validity and Reliability in Algebra Midterm Assessment: A Comprehensive Approach to Test Development and Analysis

**Matthew E. Cañeda*, Arl Joshua F. Gamaya, Manuelin C. Baring**
College of Teacher Education, Agusan del Sur State College of Agriculture and Technology, Agusan del Sur, Philippines

*Corresponding Author Email: mcaneda@asscat.edu.ph

**Date received**: September 14, 2024
**Date revised**: October 1, 2024
**Date accepted**: October 14, 2024

**Originality**: 92%
**Grammarly Score**: 99%
**Similarity**: 8%

**Recommended citation:**
Cañeda, M., Gamaya, A.F., & Baring, M. (2024). Ensuring validity and reliability in Algebra midterm assessment: A comprehensive approach to test development and analysis. *Journal of Interdisciplinary Perspectives*, 2(11), 362-372. https://doi.org/10.69569/jip.2024.0497

**Abstract.** First-year mathematics education students often face challenges with College and Advanced Algebra assessments. This study aimed to develop and validate a 100-item multiple-choice midterm test for College and Advanced Algebra, ensuring the test's validity and reliability. The test was designed following a structured process that included creating a Table of Specifications (TOS) based on the approved syllabus. To ensure content accuracy and relevance, the test was reviewed by three subject matter experts and evaluated for clarity by 15 students. Pilot testing was conducted with 82 fourth-year Bachelor of Secondary Education Mathematics (BSEd Mathematics) students. The pilot test results underwent detailed item analysis, focusing on metrics such as the difficulty index, discrimination index, and overall reliability using the Kuder-Richardson Formula 20 (KR-20). Of the 100 items, 22 were retained, 48 were revised, and 30 were discarded. The test achieved a reliability coefficient of 0.876, indicating strong internal consistency. The findings suggest that the validated test questionnaire is a dependable tool for accurately assessing students' knowledge in College and Advanced Algebra, providing valuable feedback for educators and students.

**Keywords:** College Algebra; Advanced Algebra; Item analysis; Mathematics education; Test development.

## 1.0 Introduction

Several studies have emphasized the importance of developing valid and reliable assessment tools, particularly in mathematics education, to measure student learning outcomes effectively (Kilic, 2016; Mamolo, 2021). These studies highlight the necessity of aligning test items with learning objectives and ensuring they assess lower-order cognitive and higher-order thinking skills, as outlined in Bloom's Taxonomy (Anderson & Krathwohl, 2001). While using multiple-choice formats is common in algebra assessments due to their efficiency in evaluating a broad range of knowledge, issues such as poorly designed questions and a lack of context often limit the accuracy of these tests (Quaigrain & Arhin, 2017).

A key gap identified in the literature is the lack of validated test questionnaires specifically designed for College and Advanced Algebra that incorporate a thorough item analysis to ensure reliability and validity. Many existing tests do not undergo the rigorous process of expert validation and empirical item analysis, leading to assessments that may not accurately reflect student comprehension, particularly in higher-order thinking skills such as analysis

and problem-solving (Stephens et al., 2015; Kunwar, 2018). Moreover, while there has been significant research on general assessment development, few studies focus on applying specific psychometric methods like the Kuder-Richardson Formula 20 (KR-20) for measuring internal consistency in mathematics tests.

Additionally, previous research has often failed to consider the specific needs of first-year mathematics education students, particularly those enrolled in College and Advanced Algebra, where the conceptual complexity and vast number of topics can overwhelm students. This gap in specialized assessments for these demographics underscores the need to develop tailored, validated tools that accurately measure student understanding and provide actionable feedback to learners and educators (Ocampo & Usita, 2015). Thus, this study aims to fill these gaps by developing a midterm assessment for College and Advanced Algebra that aligns with course objectives and undergoes a rigorous validation process, including expert review, face validity assessment, and item analysis using psychometric methods. This approach ensures that the test provides reliable and accurate measures of student knowledge, addressing the deficiencies found in previous assessments.

On the other hand, designing an efficient test can transform a multiple-choice questionnaire into a useful instrument for evaluating, providing feedback, and preparing learners. Educators can develop trustworthy, valid tools that adequately measure students' competence against their knowledge, considering the alignment of content and difficulty with the purpose of a given examination (Bilyakovska, 2022). The formal procedure for developing tests begins by clearly specifying learning objectives that correspond with what the test intends to achieve (Irwing & Hughes, 2018). According to universal design principles, creating inclusive evaluations helps eliminate obstacles and promote equity among all learners (Lazarus et al., 2022). Furthermore, validity and reliability are important aspects that contribute to accurate and stable test outcomes, hence used in diverse student populations or different situations (Sullivan, 2011). This approach enhances the quality of assessments and improves educational experiences for all learners.

Validity and reliability are essential components of developing effective assessment tools. Validity, which includes face and content validity, ensures that a test measures what it is intended to measure. Students' responses and expert input are crucial in refining test items to ensure they are relevant and accurately reflect the subject matter (Cañeda, 2024a; Mamolo, 2021). Content validity, in particular, is strengthened through multiple expert reviews, with a mean score of 2.60 or higher considered acceptable (Ocampo & Usita, 2015). Reliability, conversely, refers to the consistency of a test's results across different administrations. Tools like Cronbach's alpha and the Kuder-Richardson 20 (KR-20) help measure this consistency, with values above 0.70 indicating good reliability (Taherdoost, 2016; Bobbit, 2022). Reliable tests ensure student performance is accurately measured, leading to fair and meaningful evaluations (Jhangiani et al., 2015).

To maintain the validity and reliability of the test, groups of researchers propose a detailed five-phase process for developing tests. The first phase is identifying key characteristics of the test, such as purpose, audience, and difficulty level. Next comes making a table of specifications that will act as one of the components ensuring content validity by providing information on what the test covers. Then, the emphasis shifts towards collecting and evaluating tests, ensuring difficulty and content coverage continuity. In addition, clear procedures are developed along with materials for administering or scoring tests; they include instructions for who administers the test and rubrics for consistent marking. Finally, revision incorporates feedback from pilot testing into the necessary changes and improvements. The table of specifications, also called a detailed test plan, guides the construction of the test and ensures it is aligned with learning objectives (American Educational Research Association, 2014).

When creating multiple-choice questions to assess higher-order thinking, it is essential to target advanced cognitive skills as outlined in the revised Bloom's Taxonomy. This framework categorizes cognitive skills from basic knowledge recall to more complex tasks like analysis, synthesis, and evaluation (Dwyer et al., 2014). Originally developed by Benjamin Bloom in 1956, the taxonomy has evolved to emphasize the importance of moving beyond memorization to encourage critical thinking and creativity (Andreev, 2024). The revised version is widely used in modern education to support flexible, systematic teaching and assessment practices that promote deeper learning (Metzgar, 2023). The taxonomy's six levels—remembering, understanding, applying, analyzing, evaluating, and creating—provide educators with a structured way to assess students' cognitive processes, particularly in complex subjects like mathematics (Wilson, 2016). By utilizing this framework, teachers can foster

critical thinking and ensure balanced learning outcomes that address student development's cognitive, emotional, and physical aspects (University of Waterloo, 2024).

Multiple-choice questions are often preferred due to their ease of grading and ability to cover various topics (Quaigrain & Arhin, 2017). However, test items must align with instructional goals and be worded to avoid confusion, ensuring students' responses reflect their understanding (Yaddanapudi & Yaddanapudi, 2019). Test questionnaires assess learning and provide valuable feedback, helping students identify areas of strength and weakness. Unfortunately, many assessments, particularly in advanced algebra, fail to capture complex concepts like polynomial functions and logarithmic equations, leading to a misleading representation of students' abilities (Tejeda & Gallardo, 2017). Poorly designed questions or insufficient context can further distort the accuracy of these evaluations, resulting in unfair judgments of student competence (Cañeda, 2024a).

When a test questionnaire is designed purposefully and strategically implemented, it transcends the simple function of a rating tool (Chigonga, 2020). It serves as a prime source for assessment, feedback, and preparation, allowing learners to gain an insight into Algebra, thereby equipping them with the necessary abilities to excel academically and personally. In like manner, Jain et al. (2016) stress that defining the purpose of a test is central during the process of designing a questionnaire. Being able to define what the test seeks to measure enables educators to choose appropriate content and set the right level of difficulty for test items. Such an organized approach also helps develop a test blueprint that ensures an exhaustive coverage of specific topics aligned with learning objectives.

In Agusan del Sur State College of Agriculture and Technology (ASSCAT), this represents a formative step in making a validated questionnaire that evaluates students' competence in College and Advanced Algebra, especially designed for those planning to be high school teachers. This study aimed to develop a test questionnaire focusing on midterm topics. The process consisted of formulating questionnaires aligned with the Revised Bloom's Taxonomy to provide a complete assessment across multiple cognitive domains. It was also necessary for the researcher to assess the face and content validity of these instruments and establish their reliability to ensure that they consistently represent correct measurements of students' knowledge and skills in College and Advanced Algebra. Such validated instruments are reliable and valuable resources for curriculum developers, educators, students, and future researchers. They offer crucial materials for evaluation, teaching purposes, and further academic pursuits.

## 2.0 Methodology

### 2.1 Research Design

This research focuses on instrumentation, which in academic contexts typically refers to the creation and use of measurement tools like surveys, tests, and questionnaires (Biddix, 2018). The instrumentation process encompasses designing, evaluating, and applying these tools to ensure they accurately measure the desired constructs, emphasizing validity and reliability. Key aspects of this process include ensuring item clarity, content validity, and scoring consistency (Kline, 2000). This study used instrumentation to develop and validate test questionnaires covering specific College and Advanced Algebra topics, focusing on syllabus-based mid and final-term content.

### 2.2 Research Respondents

This study involved three teacher validators and fifteen student examinees. The teacher validators were instructors from the Agusan del Sur State College of Agriculture and Technology (ASSCAT), all of whom had a master's degree in mathematics and at least two years of teaching experience. These instructors were tasked with assessing the content validity of the test questionnaire. According to Gilbert and Prion (2016), utilizing three experts can provide reliable evidence of good content validity.

The student examinees included 82 4th-year Bachelor of Secondary Education major in Mathematics (BSEd-Mathematics) students who had completed courses in College and Advanced Algebra. During the pilot testing phase, these students took the developed test questionnaires, and fifteen served as student validators for assessing face validity. As Taherdoost (2016) noted, face validity refers to the extent to which a test appears to measure a specific construct as judged by non-experts or test-takers.

The selection of both validators and examinees was guided by their interest and involvement in developing and validating the test questionnaire, specifically for selected topics in College and Advanced Algebra. A universal and probability sampling technique was employed for the examinees, ensuring a representative sample of the population. A non-probability purposive sampling method was used to select the validators, focusing on their qualifications and relevance to the study. This careful selection process was crucial for collecting data pertinent to the research objectives.

**2.3 Research Instrument**

The instrument used in this study was adopted by Oducado (2020) researchers for content validation and Patel and Desai (2020) for face validation. These tools evaluated the development of the test questionnaire in selected topics of the College and Advanced Algebra midterm in terms of content validity and face validity. The rating of each criterion for face validity was "yes or no." The rating of each criterion for content validity is as follows: 5 – Strongly Agree, 4 – Agree, 3 – Neither/Nor agree, 2 – Disagree, and 1 – Strongly Disagree. A 5-point Likert scale was used as the basis for the mean ranges per factor in the evaluation rating sheet, along with the corresponding description and interpretation used in this study. Tables 1, 2, and 3 are the rating scales utilized for face validity, content validity, and reliability coefficient. Tables 4, 5, and 6 are rating scales for the difficulty index, discrimination index, and the decision.

**Table 1.** Interpretation of face validity

| % of agreement | Strength of Agreement per question or overall | Action for each Question / entire tool |
|---|---|---|
| < 80 | Poor | Restructure |
| 80 - 90 | Substantial | Substantial Revise |
| 90 - 100 | Full | Retain |

(Patel and Desai, 2020)

**Table 2**. Interpretation of content validity

| Mean Range | Verbal Description | Qualitative Interpretation |
|---|---|---|
| 4.21 – 5.00 | Very High | This means that the validity of the developed test questionnaire is very much accepted. |
| 3.41 – 4.20 | High | This means that the validity of the developed test questionnaire is much accepted. |
| 2.61 – 3.40 | Moderate | This means that the validity of the developed test questionnaire is accepted. |
| 1.81 – 2.60 | Low | This means that the validity of the developed test questionnaire is poor. |
| 1.00 – 1.80 | Very Low | This means that the validity of the developed test questionnaire is not accepted. |

(Ocampo and Usita, 2017)

**Table 3**. Reliability coefficients value interpretations

| Reliability | Interpretation |
|---|---|
| 0.90 and above | Excellent |
| 0.80 - 0.89 | Good |
| 0.70 – 0.79 | Average |
| 0.60 – 0.69 | Questionable |
| 0.50 – 0.59 | Poor |
| 0.50 or below | Unacceptable |

(Longe and Maharaj, 2023)

**Table 4**. Discrimination index (d-level) interpretation range

| D value Range | Interpretation |
|---|---|
| -1.00 – -0.60 | Questionable Item |
| -0.59 - -0.21 | Not Discriminating |
| -0.20 – 0.20 | Moderately discriminating |
| 0.21 – 0.59 | Discriminating |
| 0.60 – 1.00 | Very Discriminating |

(Padua and Santos, 1997)

**Table 5**. Difficulty level (p-level) interpretation range

| P value Range | Interpretation |
|---|---|
| 0.00 – 0.20 | Very Difficult Item |
| 0.21 – 0.40 | Difficult Item |
| 0.41 – 0.60 | Moderately Difficult Item |
| 0.61 – 0.80 | Easy Item |
| 0.81 and above | Very Easy item |

(Padua and Santos, 1997)

**Table 6.** The decision for discrimination index and difficulty index.

| Discrimination Index (D-Value) | Difficulty Index (P-Value) | Decision |
|---|---|---|
| Acceptable | Acceptable | Retained |
| Not Acceptable | Acceptable | Revised |
| Acceptable | Not Acceptable | Revised |
| Not Acceptable | Not Acceptable | Rejected |

(Padua and Santos, 1997; Cañeda et al., 2024b)

### 2.4 Development Process

This research adopted the framework of Cañeda et al. (2024b), which is rooted in the early development and validation by Mamolo (2021). The framework consists of four stages, as shown in Figure 1, in this study: conceptualization, development of the test, validation of the test, and pilot testing.

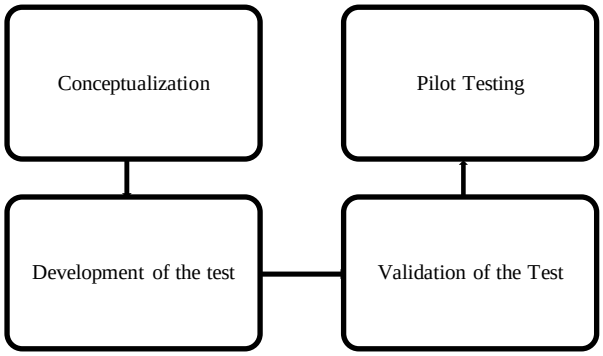


**Figure 1.** Process in the development and validation of test

#### *Conceptualization*

The initial step in this study involved identifying key concepts from selected midterm and final term topics in College and Advanced Algebra. When developing term examinations, it is crucial to ensure that the items constructed align with the constructs outlined in the approved syllabus. The test questionnaire created for the midterm contained one hundred (100) items, covering all relevant learning competencies.

The midterm test questions included number systems, functions and relations, algebraic expressions, the plane coordinate system, systems of equations, second-degree equations, radicals and rational exponents, and logarithms. These topics were all selected based on the approved syllabus from the Agusan del Sur State College of Agriculture and Technology (ASSCAT), referenced to CMO No. 75, series 2017. The representation of each topic in the exams was proportional to the instructional time and importance indicated in the syllabus.

#### *Development of the Test*

The study centered on developing midterm test questionnaires for College and Advanced Algebra courses. The first step in this process was creating a Table of Specifications (TOS), a blueprint for designing the test items. The midterm questionnaire comprised 100 items, meticulously categorized according to the Revised Bloom's Taxonomy. It covered eight distinct topics, each with a specific time allocation and percentage of questions. For example, the Real Number System was allocated 6 hours and 22.22% of the questions. At the same time, other topics such as functions, algebraic expressions, and plane coordinate systems were each given 3 hours and 11.11% of the questions. This detailed approach ensured a balanced and comprehensive assessment of students' understanding.

After drafting the test items, the next steps involved reviewing and refining them to ensure clarity and accuracy. Establishing scoring guidelines and setting criteria for the test's completeness were also crucial to ensure a thorough evaluation process. The tests were designed using a multiple-choice format deemed suitable for the assessment level and content (Kara & Çelikler, 2015). The final set of test items included various questions with varying difficulty levels, carefully chosen and refined through several rounds of review to ensure they effectively measured students' understanding.

***Validation of the Test Questionnaire***

The validation process for the draft test questionnaires in College and Advanced Algebra involved thorough checks for both content and face validity. To ensure content validity, three experts reviewed the test items—a number deemed sufficient for a comprehensive evaluation. According to Gilbert and Prion (2016), having at least three experts assesses whether the test items accurately reflect the subject matter and cover the intended content. These experts examined how well the test items aligned with the learning objectives and whether they effectively represented the measured domain.

Face validity focuses on how relevant and clear the test appears from the perspective of those who will take it, such as students. As explained by Taherdoost (2016), face validity looks at whether the test seems to measure what it is supposed to in a way that is understandable and appropriate for the test takers. In this study, fifteen students evaluated face validity, which is more than the minimum of ten participants. This larger sample size helped ensure that the feedback was reliable, ensuring the test items were seen as relevant and clear by a broad range of respondents.

The experts were given detailed materials to ensure a thorough evaluation, including the draft test questionnaires, a Table of Specifications, and a syllabus outlining the learning competencies. This comprehensive documentation allowed the validators to carefully review how well the test items matched the instructional content and educational objectives. The in-depth review process ensured that the test items were relevant, appropriately challenging, fair, and unbiased, following the best practices in educational assessment.

***Pilot Testing***

The draft test questionnaire for the midterm exams in College and Advanced Algebra was given to fourth-year BSEd Mathematics students at Agusan del Sur State College of Agriculture and Technology (ASSCAT) after getting approval from the dean of the College of Teacher Education. Eighty-two students took part in this pilot testing phase, which is crucial for evaluating the quality and effectiveness of the test items. This phase assesses the empirical validity of the test and its overall suitability (Syahfitri et al., 2019). The feedback gathered from this pilot testing is vital for refining the questionnaire and addressing clarity, difficulty, and effectiveness issues. This process helps ensure that the test aligns with the learning objectives and accurately measures the intended knowledge and skills, ultimately enhancing the reliability and validity of the assessment tool.

After completing the pilot testing, the next step was to evaluate the reliability of the test items. To do this, the Kuder-Richardson 20 (KR-20) formula was used, which calculates reliability on a scale from 0 (indicating poor reliability) to 1 (indicating excellent reliability), with higher values reflecting greater reliability (Bobbit, 2022). An item analysis was then conducted, essential for tests where students choose the correct answer from multiple options. According to Kunwar (2018), item analysis is particularly important for multiple-choice exams, which often significantly impact students' grades or other key outcomes. This analysis helps identify any problematic questions that might be ambiguous, irrelevant, too easy or too difficult, or unable to differentiate between high and low performers effectively.

Through item analysis, educators can enhance the quality of an exam by identifying and refining or removing problematic questions. This process also helps improve classroom instruction by pinpointing areas where students may struggle and guiding targeted remediation efforts. For high-stakes exams like midterms, item analysis ensures that the assessments are fair, valid, and accurately measure students' knowledge and skills (Kunwar, 2018).

## 3.0 Result and Discussion

### 3.1 Content and Face Validity

The overall mean given by the three experts about content validity was 4.27 (Table 7), which is described as very high. This implies that the quality of the developed test questionnaire is very much accepted and that the test questionnaire adequately covers the full range of content that is intended to be measured. Suppose the mean range is greater than or equal to 2.60 as an overall mean for content validity. In that case, it indicates "moderate to very high" or "acceptable to very much acceptable" to the experts (Ocampo and Usita (2015). This means that the content validity of the developed test questionnaire for the College and Advanced Algebra Midterm was essential and accurately measured what it intended.

**Table 7.** Content validity of the test questionnaire

| Content Validity | Mean | Verbal Description |
|---|---|---|
| Expert 1 | 4.00 | High |
| Expert 2 | 4.18 | High |
| Expert 3 | 4.63 | Very high |
| **Overall Mean** | **4.27** | **Very high** |

Face validity obtained an overall mean percentage of 97%, meaning the developed test questionnaire is full. This implies that the developed test questionnaire measures the extent to which respondents and experts feel the questionnaire measures what it is intended to measure (Yaddanapudi & Yaddanapudi, 2019). If the range is within 90%–100%, the strength of agreement per question or overall is almost perfect, and the actions for each question or entire tool will be retained (Desai & Patel, 2020). This means that the face validity of the developed test questionnaire in College and Advanced Algebra in midterm was satisfied and acceptable.

**Table 8.** Face validity of the test questionnaire

| Face Validity | Percentage | Verbal Description |
|---|---|---|
| Rater 1 | 100% | Full |
| Rater 2 | 100% | Full |
| Rater 3 | 100% | Full |
| Rater 4 | 100% | Full |
| Rater 5 | 100% | Full |
| Rater 6 | 100% | Full |
| Rater 7 | 100% | Full |
| Rater 8 | 100% | Full |
| Rater 9 | 100% | Full |
| Rater 10 | 100% | Full |
| Rater 11 | 90% | Full |
| Rater 12 | 90% | Full |
| Rater 13 | 90% | Full |
| Rater 14 | 90% | Full |
| Rater 15 | 90% | Full |
| **Overall Mean** | **97%** | **Full** |

### 3.2 Item Analysis

Table 9 shows the difficulty distribution for the College and Advanced Algebra Midterm examinations. A significant portion of the items is categorized as very easy. Specifically, 39% of the items were found to have a difficulty index ranging from 0.81 and above, indicating that most students answered these items correctly. A difficulty index within this range typically suggests that the items are straightforward for the test-takers, making the assessment less challenging overall.

In addition to the very easy items, 29% fell into the easy category, with a difficulty index ranging from 0.61 to 0.80. This classification also implies that a substantial number of students managed to answer these items correctly. Together, both the very easy and easy categories account for a significant 68% of the total items, highlighting that a majority of the test consists of items that students find accessible. Conversely, 11% of the items are considered moderately difficult, with a difficulty index falling between 0.41 and 0.60. This indicates that while some students may struggle, a fair number still answer these questions correctly, showcasing a balanced approach to assessing students' understanding at varying levels of challenge.

**Table 9**. Mid-term item distribution (difficulty index)

| Difficulty Index | Test item number | Frequency | Percentage | Verbal Interpretation |
|---|---|---|---|---|
| 0.81 and above | (2, 3, 6, 7, 8, 9, 10, 11, 12, 13, 24, 25, 32, 33, 34, 39, 42, 43, 45, 47, 52, 53, 54, 56, 59, 68, 74, 78, 82, 83, 84, 87, 90, 91, 92, 93, 94, 98, and 99) | 39 | 39% | Very Easy item |
| 0.61 – 0.80 | (1, 4, 5, 16, 17, 19, 20, 26, 27, 30, 31, 36, 37, 40, 41, 46, 50, 51, 57, 61, 63, 64, 65, 67, 80, 88, 89, 95, and 96) | 29 | 29% | Easy Item |
| 0.41 – 0.60 | (18, 21, 29, 35, 38, 55, 71, 72, 79, 85, and 100) | 11 | 11% | Moderately Difficult Item |
| 0.21 – 0.40 | (14, 48, 69, 70, 75, 76, 77, and 97) | 8 | 8% | Difficult Item |
| 0.00 – 0.20 | (15, 22, 23, 28, 44, 49, 58, 60, 62, 66, 73, 81, and 86) | 13 | 13% | Very Difficult Item |

Among the test items, 8% were classified as difficult, with difficulty indices ranging from 0.20 to 0.40. These items might cover less familiar content or could be poorly constructed, undermining the assessment's effectiveness. Additionally, 13% of the items were categorized as difficult, with indices below 0.20. These particularly challenging items will likely test students significantly, potentially highlighting areas where more instructional focus may be needed.

The findings emphasized the need to adjust item difficulty and enhance the effectiveness of distractors used in assessments (Rezigall et al., 2019). Identifying the areas where students struggle provides valuable insights for educators, guiding them on where to concentrate their teaching efforts. This understanding allows teachers to develop tailored strategies that address these challenges, ultimately leading to improved learning outcomes in mathematics (Aguhayon et al., 2023).

Table 10 shows how the items in the midterm test questionnaire for College and Advanced Algebra are distributed based on their discrimination index. This index measures how well each item can differentiate between high and low performers on the test. Impressively, over 90% of the items are considered acceptable for their ability to distinguish between different levels of student performance. In particular, 11% of the items are classified as very discriminating or excellent, meaning they effectively differentiate between top-performing and lower-performing students.

Furthermore, 54% of the items fall under the category of discriminating or good, which suggests that these items adequately identify variations in student performance. An additional 26% of items are moderately discriminating or reasonably good, signifying that while they may not be the most effective, they still possess some capacity to differentiate levels of understanding among students. Conversely, only 8% of the items are characterized as not discriminating or marginal, which implies they do not effectively distinguish among student performance. Alarmingly, 1% of the items were identified as poor or questionable, indicating a need for revision or elimination from future assessments. This categorization underscores the importance of continuous evaluation and improvement of assessment tools to ensure they fulfill their intended purpose effectively.

**Table 10.** Mid-term Item Distribution (Discrimination Index)

| Discrimination Index | Test item number | Frequency | Percentage | Verbal Interpretation |
|---|---|---|---|---|
| -1.00 – -0.60 | (14) | 1 | 1% | Questionable Item |
| -0.59 - -0.21 | (21, 35, 58, 66, 69, 79, 85, and 97) | 8 | 8% | Not Discriminating |
| -0.20 – 0.20 | (7, 8, 11, 15, 16, 19, 22, 23, 28, 41, 44, 48, 49, 59, 60, 62, 70, 72, 73, 75, 76, 77, 81, 84, 86, 100) | 26 | 26% | Moderately discriminating |
| 0.21 – 0.59 | (1, 2, 3, 4, 5, 6, 9, 10, 12, 13, 17, 20, 24, 25, 26, 27, 30, 31, 32, 33, 34, 36, 37, 38, 39, 40, 42, 43, 45, 46, 47, 52, 53, 54, 55, 56, 57, 64, 68, 71, 74, 78, 82, 83, 87, 89, 90, 91, 92, 93, 94, 96, 98, and 99). | 54 | 54% | Discriminating |
| 0.60 – 1.00 | (18, 29, 50, 51, 61, 63, 65, 67, 80, 88, 95) | 11 | 11% | Very Discriminating |

The distribution of items in the College and Advanced Algebra Midterm reflects a strong overall performance, with most items meeting acceptable discrimination standards. This high percentage of effective discrimination indicates that the test items are of good quality (Ferrando et al., 2023). However, it is important to note that high

discrimination does not always equate to effective measurement. Factors like redundancy, shared residuals, and biased distributions can influence these results (Kılıç & Uysal, 2022). This underscores the need to identify and address potential issues that might skew discrimination estimates. Continuous refinement of assessments is crucial for improving educational outcomes and better-supporting student learning.

Table 11 offers important insights into item discrimination and difficulty. The analysis shows that out of all the items evaluated, 22 were retained, indicating a high level of acceptance for these questions. However, 48 items—making up 48% of the total—needed revision. This suggests a significant need for adjustments to assess the intended concepts better. Additionally, 30 items, or 30% of the total, were slated for rejection. This indicates that while a substantial number of items either met the required standards or could be improved, there is still room for refinement to ensure the effectiveness of the assessment.

**Table 11**. Decision for discrimination index and difficulty index

| Final Evaluation | Item Number/s | Frequency | Percentage |
|---|---|---|---|
| Items to be Retained | (1, 4, 5, 17, 18, 26, 29, 30, 31, 38, 50, 51, 55, 57, 61, 63, 65, 57, 71, 80, 88, and 95) | 22 | 22% |
| Items to be Revised | (2, 3, 6, 7, 8, 9, 10, 12, 13, 16, 20, 24, 25, 27, 32, 33, 34, 36, 37, 39, 40, 42, 43, 45, 46, 47, 52, 53, 54, 56, 64, 68, 74, 78, 82, 83, 84, 87, 89, 90, 91, 92, 93, 94, 96, 98, 99, and 100) | 48 | 48% |
| Items to be Rejected | (11, 14, 15, 19, 21, 22, 23, 28, 35, 41, 44, 48, 49, 58, 59, 60, 62, 66, 69, 70, 72, 73, 75, 76, 77, 79, 81, 85, 86, and 97) | 30 | 30% |

These findings highlight the crucial role of thorough item analysis in creating effective assessments. By sorting items into retained, revised, or rejected categories, educators can ensure that their tests stay valid and reliable, leading to better measurement outcomes (DeVellis, 2003). Striking the right balance between keeping valuable items and revising those with issues is essential for upholding the quality and effectiveness of assessment tools. This careful approach helps maintain the integrity of the evaluation process and ultimately supports improved student learning and assessment accuracy.

### 3.3 Reliability of the Test

The validated midterm test questionnaire was pilot-tested with fourth-year BSEd Mathematics students, leading to refinements that removed overly difficult, too easy, or misleading questions. The revised 99-item test was evaluated for reliability using the Kuder-Richardson 20 (KR-20) formula, which assesses internal consistency for binary data (correct/incorrect answers). The KR-20 coefficient of 0.876 indicates a high level of reliability and consistency for the questionnaire (University of Washington, 2020). This score, close to 1.0, confirms that the test meets the criteria for acceptable reliability (Kilic, 2016). The questionnaire is therefore considered effective for evaluating students' knowledge and skills in College and Advanced Algebra. However, item 11 was removed from the test due to zero variance in student responses.

## 4.0 Conclusion

The study's findings led to the creation of midterm test questionnaires for College and Advanced Algebra, carefully aligned with the approved syllabi. The aim was to enhance existing assessment tools, and the newly developed questionnaires showed strong face and content validity, with ratings of "Full" and "Very High," respectively. They also demonstrated high reliability and consistency, confirming their effectiveness in accurately measuring the intended content and evaluating student performance in these advanced math courses. The item analysis process refined the questionnaires further, resulting in 22 items being retained, 48 revised, and 30 rejected. This thorough evaluation ensures the assessments are well-suited for capturing students' understanding and skills in College and Advanced Algebra.

Items identified for revision will undergo a thorough refinement process. This involves carefully reviewing and modifying the items based on feedback and analysis. Adjustments will be made to improve clarity, ensure better alignment with learning objectives, and accurately measure the targeted knowledge and skills. The goal is to enhance the current assessment tool and contribute to the ongoing development of future test items. This iterative process helps the test evolve to assess student understanding and learning outcomes more effectively. By

continually refining the items, we foster a robust and reliable assessment framework that supports accurate evaluation of students' competencies in College and Advanced Algebra. The improved items will become essential components of a more effective and comprehensive assessment tool, ultimately enhancing the quality and fairness of future exams.

The findings of this study offer valuable contributions to mathematics education. By developing a reliable and validated test for College and Advanced Algebra, educators now have an effective tool to assess students' understanding of complex concepts and higher-order thinking skills more accurately. The item analysis provides insights to help teachers pinpoint areas where students struggle, enabling more targeted instruction to address learning gaps. The methodology, including tools like the Kuder-Richardson Formula 20 (KR-20), also sets a standard for creating similar assessments in other subjects. The study's results can guide curriculum improvements by ensuring teaching materials align with student needs. The test design's focus on fairness and accuracy also promotes equity in student evaluations. Lastly, this research paves the way for further studies on algebra assessments, encouraging the development of new methods for evaluating higher-order thinking skills.

Based on the findings of the study, the following recommendations are offered:

1. Implementation of the Revised Test Questionnaires: The revised test questionnaires for both the midterm and final terms should be implemented in College and Advanced Algebra courses to enhance assessment accuracy and reliability.
2. Continuous Evaluation and Improvement: It is essential to regularly evaluate and refine test questionnaires to ensure they continue to measure student learning outcomes effectively. This ongoing process helps maintain the quality of assessments and ensures they remain relevant and accurate in evaluating students' progress and understanding.
3. Utilization of the Test Questionnaires as a Benchmark: The developed test questionnaires can be a benchmark for creating similar assessment tools in other mathematics courses.
4. Further Research on Item Analysis: To further improve the test questionnaires, in-depth item analysis can be conducted to identify specific item characteristics that contribute to high or low discrimination and difficulty indices.
5. Teacher Training on Test Construction: Training teachers on test construction principles and item analysis techniques can enhance the quality of classroom assessments and student learning outcomes.

## 5.0 Contribution of Authors

Matthew E. Cañeda: Editing, writing, supervising, data analysis, conceptualization
Arl Joshua F. Gamaya: Writing, data analysis
Manuelin C. Baring: data gathering, checking, encoding

## 6.0 Funding

This research did not receive support from any funding agency.

## 7.0 Conflict of Interests

The research has no conflicts of interest, as the researchers do not have any financial, personal, or professional ties that could bias the study. The primary goal was to develop and validate a test questionnaire for assessing first-year students' knowledge of College and Advanced Algebra. The researchers independently funded the study, ensuring no external influence on the results. Furthermore, the experts involved in the evaluation were impartial, providing objective feedback without any vested interest in the study's outcomes.

## 8.0 Acknowledgement

The researchers expressed their gratitude to the students who participated in the study and extended their thanks to the experts for their invaluable suggestions, which significantly improved the test items.

## 8.0 References

Aguhayon, H., Tingson, R., & Pentang, J. (2023). Addressing students learning gaps in mathematics through differentiated instruction. International Journal of Educational Management and Development Studies, 4(1), 69–87. https://doi.org/10.53378/352967

Anderson, L., & Krathwohl, D.A. (2001). Taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of Educational Objectives. New York: Longman.

Andreev, I. (2024). Bloom's Taxonomy. Retrieved from https://tinyurl.com/42pmmk9b

Biddix, J.P. (2018). Research methods and applications for student affairs. John Wiley & Sons.

Bilyakovska, O. (2022). Test as an effective means of assessing the quality of students' knowledge. Academic Notes Series Pedagogical Science, 1(204), 16–20. https://doi.org/10.36550/2415-7988-2022-1-204-16-20

Bobbitt, Z. (2022). Kuder-Richardson Formula 20 (Definition & example). Retrieved from https://www.statology.-org/kuder-richardson-20/

Cañeda, M.E., Amar, R.P., & Lucin, E.L. (2024a). Development of test questionnaire on selected topics in calculus 1 (final term). International Journal of Research and Scientific Innovation, 9(8), 244-255. https://doi.org/10.51244/IJRSI.2024.1108020

Cañeda, M.E., Logroño, J.J.F., & Culibra, C.D. (2024b). Test questionnaire development on selected topics in calculus 1. Ignatian International Journal for Multidisciplinary Research, 2(8), 1363–1376. https://doi.org/-10.5281/zenodo.13371155
Chigonga, B. (2020). Formative Assessment in Mathematics Education in the Twenty-First Century. IntechOpen.
DeVellis, R.F. (2003). Scale development: theory and applications, applied social research methods. Sage Publications.
Dwyer, C. P., Hogan, M. J., & Stewart, I. (2014). An integrated critical thinking framework for the 21st century. Thinking Skills and Creativity, 12, 43–52. https://doi.org/10.1016/j.tsc.2013.12.004
Ferrando, P.J., Lorenzo-Seva, U., & Bargalló-Escrivà, M. T. (2023). Gulliksen's pool: A quick tool for preliminary detection of problematic items in item factor analysis. PloS one, 18(8), e0290611. https://doi.org/10.1371/journal.pone.0290611
Gilbert, G.E., & Prion, S. (2016). Making sense of methods and measurement: Lawshe's content validity index. Clinical Simulation in Nursing, 12(12), 530–531. https://doi.org/10.1016/-j.ecns.2016.08.002
Irwing, P., & Hughes, D.J. (2018). Test development. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development (pp. 3–47). Wiley Blackwell.
Jain, S., Dubey, S., & Jain, S. (2016). Designing and validation of questionnaire. International Dental & Medical Journal of Advanced Research, 2(1), 1–3. https://doi.org/10.15713/ins.idmjar.39
Jhangiani, R.S., Chiang, I.A., Cuttler, C., & Leighton, D.C. (2019). Research methods in psychology – 2nd Canadian edition. KPU.
Kara, F., & Celikler, D. (2015). Development of achievement test: Validity and reliability study for achievement test on matter changing. Journal of Education and Practice, 6(24), 21–26. https://eric.ed.gov/?id=EJ1078816
Kılıç, A. & Uysal, I. (2022). To what extent are item discrimination values realistic? A new index for two-dimensional structures. International Journal of Assessment Tools in Education, 9, 728-740. https://doi.org/10.21449/ijate.1098757
Kilic, S. (2016). Cronbach's alpha reliability coefficient. Journal of Mood Disorders, 6(1), 47. https://doi.org/10.5455/jmood.20160307122823
Kline, P. (2000). Handbook of psychological testing, second edition. Routledge.
Kunwar, R. (2018). Development and standardization process of mathematics achievement test for the students of grade x. International Journal of Current Research, 10(11), 75451–75455. https://doi.org/10.24941/ijcr.33168.11.2018
Lazarus, S.S., Johnstone, C.J., Liu, K.K., Thurlow, M.L., Hinkle, A.R., & Burden, K. (2022). An updated state guide to universally designed assessments (NCEO Report 431). Retrieved from https://tinyurl.com/y7tf3ty9
Longe, I.O., & Maharaj, A. (2023). Investigating students' understanding of complex number and its relation to algebraic group using and APOS theory. Journal of Medives : Journal of Mathematics Education IKIP Veteran Semarang, 7(1), 117. https://doi.org/10.31331/ medivesveteran.v7i1.2332
Mamolo, L.A. (2021). Development of an achievement test to measure students' competency in general mathematics. Anatolian Journal of Education, 6(1), 79-90. https://doi.org/10.29333/aje.2021.616a
Metzgar, M. (2023). Revised Bloom's taxonomy in a principles of Economics textbook. Acta Educationis Generalis, 13(3), 15–28. https://doi.org/10.2478/atd-2023-0019.
Ocampo, R., & Usita, N. P. (2015). Development of Lubeg (Syzygiumlineatum (Roxb.) Merr.& Perry) processed products. Asia Pacific Journal of Multidisciplinary Research, 3(4), 118-123. https://tinyurl.com/yxb9ccxn
Oducado, R. M. (2020). Survey instrument validation rating scale. Retrieved from https://doi.org/10.2139/ssrn.3789575
Padua, R.N., & Santos, R.G. (1997). Educational evaluation and measurement: Theory, practice, and application. KATHA Publishing: QC.
Patel, N., & Desai, S. (2020). Abc of face validity for questionnaire. International Journal of Pharmaceutical Sciences Review and Research, 65(1), 164–168. https://doi.org/10.47583/ijpsrr.2020.v65i01.025
Quaigrain, K., Arhin, A. K., & King Fai Hui, S. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. Cogent Education, 4(1). https://doi.org/10.1080/2331186X.2017.1301013
Rezigalla, A. A., Ibrahim, E. K., & ElHussein, A. B. (2019). Item Analysis: The impact of distractor efficiency on the discrimination power of multiple choice items. Retrieved from https://doi.org/10.21203/rs.2.15899/v1
Stephens, A., Blanton, M., Knuth, E., Isler, I., & Gardiner, A. M. (2015). Just say yes to early algebra! Teaching Children Mathematics, 22(2), 92–101. https://doi.org/10.5951/teacchilmath.22.2.0092
Sullivan G. M. (2011). A primer on the validity of assessment instruments. Journal of graduate medical education, 3(2), 119–120. https://doi.org/10.4300/JGME-D-11-00075.1
Syahfitri, J., Firman, H., Redjeki, S., & Srivati, S. (2019). Development and validation of critical thinking disposition test in Biology. International Journal of Instruction, 12(4), 381-392. https://doi.org/10.29333-/iji.2019.12425a
Taherdoost, H. (2016). 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, 5(3), 28-36. http://dx.doi.org/10.2139/ssrn.3205040
Tejeda, K., & Gallardo, G. (2017). Performance assessment on high school advanced algebra. International Electronic Journal of Mathematics Education, 12(3), 777-798. https://doi.org/10.29333/iejme/648
Wilson, L.O. (2016). Anderson and Krathwohl: Bloom's taxonomy revised. Retrieved from https://tinyurl.com/4s9vhnee
Yaddanapudi, S., & Yaddanapudi, L.N. (2019). How to design questionnaires. Indian Journal of Anaesthesia, 63(5), 335–337. https://doi.org/10.4103/ija.IJA_274_19

.