ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue VIII August 2024

Development of Test Questionnaire on Selected Topics in Calculus 1 (Final Term)

Matthew E. Cañeda, Roland Jr P. Amar, El Francis L. Lucin

College of Teacher Education, Agusan del Sur State College of Agriculture and Technology San Teodoro, Bunawan 8506, Agusan del Sur, Philippines

DOI: https://doi.org/10.51244/IJRSI.2024.1108020

Received: 16 July 2024; Revised: 22 July 2024; Accepted: 25 July 2024; Published: 31 August 2024

ABSTRACT

Future math teachers often struggle with their final calculus exams, highlighting the need for improved assessment methods. This study developed a new multiple-choice calculus test specifically for college students training to be high school math teachers. The primary aim was to create and validate a Calculus 1 final exam for second-year students in the Bachelor of Secondary Education program, majoring in Mathematics, at Agusan del Sur State College of Agriculture and Technology. A 100-item multiple-choice test was designed based on the approved syllabus and evaluated by three experts for content refinement and 15 student validators for structure. The test's validity was deemed almost perfect and highly acceptable. It was pilot-tested with 77 students, and item analysis determined the difficulty index, discrimination index, and reliability. Following adjustments, 54 questions were included in the final test form. The reliability coefficient of 0.895 indicates high internal consistency, suggesting that the developed test questionnaire is an excellent tool for classroom assessment.

Keywords: test questionnaire development, item analysis, Mathematics Education

INTRODUCTION

In higher education, the process of education and instruction is commonly known as "lecturing" (Mutakin, 2013). Pursuing a mathematics education degree, especially to become a math teacher, is a common path in this field (Fatimah & Yerizon, 2019). However, many students find mathematics challenging due to its complexity (Sugiarti, 2016). To overcome learning obstacles and achieve educational goals, educators need to understand the diverse challenges students face when learning mathematics (Yuwono, 2016). Specifically, in the Philippines, calculus is particularly challenging for college students, with Calculus 1 forming the essential foundation for more advanced concepts in Calculus 2 (Angeles et al., 2015; Nuraeni, 2018).

Students often face difficulties with reading, writing, and accounting in calculus, primarily due to challenges in identifying questions and using integral symbols (Raupu, 2020). A substantial number of students fail their assessment exams, largely due to inadequately designed questionnaires (Bee & Murdoch Eaton, 2016). While exams remain a preferred method for evaluating student capabilities, poorly constructed questionnaires can lead to inaccurate assessments of students' true abilities.

While multiple-choice exams are popular for their ease of use and scoring (Jovanovska, 2018; Kumara et al., 2019), crafting effective questions remains a challenge for educators. These exams, though convenient, may not fully assess the depth of a student's understanding according to Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001). This framework categorizes cognitive skills into remembering, understanding, applying, analyzing, evaluating, and creating. Ideally, assessments should target higher-order thinking skills like analysis and evaluation, which are crucial in mathematics education (Oktaviana, & Susiaty, 2020). After all, math helps students develop valuable abilities like problem-solving and critical thinking that benefit them throughout their academic and professional careers.

A test questionnaire is a structured tool used to evaluate students' abilities, knowledge, and skills (Harlen, 2013). Moreover, in Filipino education, culturally relevant assessment tools are necessary to align tests with regional

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue VIII August 2024

goals and student experiences, ensuring accurate evaluations (Santos & Reyes, 2019). To ensure an accurate assessment of student knowledge, educators must prioritize well-written multiple-choice questions (Nedeau-Cayo et al., 2013; Przymuszała et al., 2020). The American Educational Research Association (2014) and other organizations emphasize the importance of fair and accessible testing. This means incorporating universal design principles during test development. These principles involve considering factors like question format and content type to avoid creating barriers for any students (AERA et al., 2014). Fairness is achieved through meticulous design and review processes (Gholami & Soleimani, 2022). A well-designed test goes through a multi-step process outlined by AERA et al. (2014), including defining the test's purpose, creating a detailed blueprint, and piloting the test before finalizing it. This ensures the test accurately measures the intended learning objectives (Çetin, 2019; Şahin et al., 2023).

In foundational courses like Calculus 1, well-designed assessments are essential for gauging student understanding. These assessments should align with instructional goals to ensure students are mastering the intended concepts and to identify areas needing improvement (Smith & Johnson, 2018). The development and validation of these assessments are crucial to guarantee their accuracy (validity) and consistency (reliability).

Validity refers to the extent to which collected data accurately reflects the research topic, emphasizing the need to measure what is intended (Taherdoost, 2016). It is considered the most fundamental aspect of test development and evaluation, as highlighted by the Standards for Educational and Psychological Testing (AERA et al., 2014). Validity encompasses several components, including face and content validity. Face validity assesses whether respondents and experts believe the questionnaire measures its intended purpose, often through informal, non-quantitative methods (Yaddanapudi & Yaddanapudi, 2019). Content validity measures how well the instrument's contents match the broader field of application, ensuring that all pertinent items are covered (Straub et al., 2014). Experts also evaluate the readability of the questionnaire, with their feedback incorporated before finalizing the test (Mamolo, 2021).

Reliability measures the consistency of the findings, ensuring that the questionnaire consistently evaluates the intended items. Validating test questionnaires can involve methods such as item analysis, expert review, and pilot testing (Effandi Zakaria, 2013). Face validity, as judged by non-experts, indicates how closely a measure aligns with a certain construct (Taherdoost, 2016). A minimum of 10 non-experts is recommended for a face-validity study, while three content validity experts can suffice for good content validity (Gilbert & Prion, 2016). The reliability coefficient, representing the correlation between two sets of test findings, reflects the stability or uniformity of scores over time or among raters (Bolarinwa, 2015). Higher reliability is achieved with more multiple-choice questions focused on a learning objective (Brame, 2013).

Cronbach's Alpha coefficient is the most frequently used measure of internal consistency or reliability between various items, measurements, or ratings (Taherdoost, 2016). Although research on Cronbach’s Alpha's application beyond questionnaire formulation is limited, it remains a crucial tool for assessing reliability (Bujang et al, 2018). Importantly, for a test to be considered reliable, it must first be valid (Taherdoost, 2016).

Kilic (2016) notes that Cronbach’s alpha coefficient is used in statistics to estimate the reliability of psychometric tests, calculated either as an average value for each scale item or separately for each item, with a value of 0.70 or higher indicating good reliability. However, an alpha greater than 0.90 might suggest redundancy among items, while a low alpha could be due to a small number of questions. The KR-20 formula, a special case of Cronbach’s alpha, is used to assess the internal consistency of test scores (Kara & Celikler, 2015). Reliability coefficients generally range between 0.00 and 1.00 and should not be negative (Fraenkel & Wallen, 2009). A misleadingly low KR-20 value can result from fewer questions, fewer examinees, or multiple key modifications, with a value of 0.70 or higher being acceptable with at least 30 exam takers (Ermie, 2017). Post-pilot testing, and reliability testing help determine item consistency, with higher KR-20 values indicating greater reliability (Bobbit, 2022).

In developing test questionnaires in Calculus 1 the following two primary theories can be utilized: classical test theory (CTT) and item response theory (IRT). Hambleton and Jones (1993) state that both CTT and IRT can be applied to creating test questionnaires for Calculus 1. CTT provides a basic framework for analyzing test items and evaluating the reliability and validity of a questionnaire, focusing on the relationship between observed

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue VIII August 2024

scores, true scores, and measurement errors. This allows for assessing item difficulty, discrimination, and overall test reliability. In contrast, IRT offers a more advanced approach by modeling the relationship between students' responses and their abilities, using models like the Rasch model or the 2-parameter logistic model to provide detailed information about item characteristics and student abilities, resulting in a more precise assessment of knowledge and skills in Calculus 1 (Hambleton & Jones, 1993).

Several studies have demonstrated the effectiveness of CTT and IRT in developing mathematics assessments. Butakor (2022) used both methods to analyze the psychometric properties of a mathematics test, showing how these approaches can complement each other to provide a comprehensive evaluation of test items. Ding and Beichner (2009) highlighted the advantages of IRT in developing physics assessments, which share common topics with Calculus 1, such as integrals, and emphasized IRT's ability to provide more accurate estimates of student abilities and item parameters. Hambleton and Jones (1993) also argue that combining CTT and IRT can lead to the creation of high-quality, reliable, and valid test questionnaires in mathematics, including those focusing on Calculus I. Using these theories, educators can develop effective tools to measure student knowledge and skills in specific mathematical topics.

Despite extensive research on student assessment, there remains a gap in developing and validating exams for specific topics like Calculus 1, particularly for final-term evaluations. Focused research is required to create valid and reliable test questionnaires for evaluating students' knowledge in Calculus 1. Current tests do not strictly follow the conventional test development process (Nasreen et al., 2019). Locally, there are no existing studies on the development and validation of Calculus 1 test questionnaires for the final term. A comprehensive calculus test questionnaire is vital as it enables a deeper assessment of key concepts, identifies common learning gaps, and supports targeted instructional modifications, thereby enhancing the overall educational quality (Zapata-Rivera & Suescun, 2015; Black & Wiliam, 2018).

This study addresses the deficiencies of using unvalidated tests for final-term examinations in Calculus 1. Recognizing the negative impact of these tests, the researchers aimed to develop reliable and validated questionnaires. Mastery of Calculus 1 is essential for success in subsequent courses like Calculus 2 and for understanding real-world applications of mathematics. Validated tests provide a more accurate assessment of student learning, thereby enhancing the overall educational experience (Kilic, 2016; Kara & Celikler, 2015).

The study had three primary objectives. First, it aimed to create test questionnaires aligned with Revised Bloom's Taxonomy, ensuring the tests assess various levels of thinking skills beyond rote memorization. Second, it sought to evaluate the content and face validity of the questionnaires. Content validity ensures the tests measure the intended learning objectives, while face validity confirms that the questions are clear and well-structured. Lastly, the study aimed to determine the reliability of the tests, ensuring they consistently measure what they are designed to measure. These validated tests benefit various stakeholders of the College of Teacher Education at Agusan del Sur State College of Agriculture and Technology (ASSCAT), helping curriculum developers refine learning experiences, aiding teachers in assessing student understanding and adjusting teaching strategies, and providing students with accurate assessments of their knowledge and areas for improvement. This study also serves as a valuable resource for future research in Calculus 1 assessment.

METHOD

Research Design

This study used instrumentation research. A measurement tool (such as a survey, test, questionnaire, etc.) is generally referred to by researchers as an instrument. To distinguish between an instrument and instrumentation, it is essential to note that an instrument is a physical device, whereas instrumentation refers to the overall process of developing, testing, and utilizing the device (Biddix, 2018). This refers to a strategy that a researcher has deliberately chosen before commencing data collection to effectively achieve the research objective. The process of instrumentation involves crafting items that accurately capture the intended construct, ensuring their validity and reliability (Sireci, 2016). This includes considerations like item clarity, content validity, and scoring consistency (Kline, 2013).

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue VIII August 2024

This research adapted the development and validation process designed by Graham (2012). The model consists of four stages that include conceptualization, development of the test, trial of the test, and testing (Mamolo, 2021). In this study, the process involved conceptualization, development of the test, validation of the test, and pilot testing.

Conceptualization

In this phase, is the understanding of the concepts of selected final-term topics in Calculus 1. In developing or compiling the Calculus 1 final-term test, the constructed items must represent each construct based on the approved syllabus. The test questionnaire is based on different learning competencies in the final-term coverage of Calculus 1. The identified topics in the final-term coverage of Calculus 1 are indefinite integrals-power, indefinite integrals-logarithm, indefinite integrals-trigonometric functions, definite integrals, and some application of the integrals. This is based on the approved syllabus.

To ensure that all learning competencies were reflected on the test questionnaire each topic are properly represented based on the number of hours and percentages. The test questionnaire is composed of 100 items covering the 3 topics: the indefinite integral with a 12 number of hours; the definite integral with a 12 number of hours; and lastly, the application of the definite integral with an 8 number of hours based on the approved syllabus. There are 38 out of the items, or 37.5%, on the topic of indefinite integral. On the other hand, there are only 37 out-of-item questions, or 37.5% for a definite integral topic. Finally, 25 out of the total number of item questions, or 25%, are on the topic of applying the definite integral. Moreover, the test questions were divided following the given percentage of difficulty of the Revised Bloom's Taxonomy. This ensures all students are assessed on the same material and have the opportunity to demonstrate their understanding of the key concepts covered in the final term.

Development of the Test

This includes drafting and compilation of the test. The test that was developed in this study is the final term test questionnaire in Calculus 1. According to Mamolo (2021), the design phase includes drafting the table of specifications (TOS), followed by writing test items. Subsequently, test item review and correction, scoring guidelines, compilation, and completeness criteria determination followed. This will utilize a multiple-choice type of test appropriate for the year level being assessed (Kara & Çelikler, 2015).

In drafting questions for the test, the table of specifications (TOS) was first crafted. Then, test items were made based on the topics reflected in the (TOS). Subsequently, thorough test items review and correction were done. Scoring guidelines, compilation, and completeness criteria determination followed.

In the compilation of the test items, the draft of the Calculus I final term test questionnaire consisted of questions with different levels of difficulty. Some questions coming from test banks or old or previously used questions were considered. The total items were developed through a series of rechecks, readings, and references to the learning competencies and table of specifications by the researchers.

Validation of the Test

The validation process involved content and face validity. For content validity, three experts in the field of Calculus were involved in the validation process. These experts have backgrounds in mathematics education or a related field and possess a deep understanding of the subject matter. Their expertise is crucial in evaluating the test's content validity. The validators were not just given the draft test questionnaire. They were also provided with additional resources to help them assess its content validity like the TOS and Syllabus. The instrument evaluation used in this study was adapted by the researchers from Oducado (2020) for content validation, and Desai and Patel (2020) for face validation.

The table of Specifications (TOS) document outlines the blueprint for the test, detailing the learning objectives, question types, difficulty levels, and how many questions address each topic. The syllabus Outlining Learning Competencies document specifies the specific skills and knowledge students are expected to have acquired by the final term of Calculus 1.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue VIII August 2024

With these resources in hand, the experts effectively evaluate the test by checking if the test questions align with the learning competencies outlined in the syllabus. Verify if the questions cover all the key topics and concepts listed in the TOS. Assess whether the difficulty level of the questions is appropriate for students at this stage of the course, and identify any questions that might be ambiguous, unclear, or not truly measure the intended learning objectives.

For face validity, fifteen 4th-year students were involved. This refers to the initial impression of a test or questionnaire. It is about whether the student validators believe the questionnaire accurately reflects the learning objectives of the Calculus 1 final term. Student Validators have completed Calculus 1 and are familiar with the subject matter. They are not necessarily experts in test development, but their perspectives can be valuable in assessing the face validity of the questionnaire.

Students provide valuable insights into whether the questionnaire feels relevant and appropriate from the perspective of someone who has taken the course. Also, they might flag questions that seem ambiguous or unclear, helping to improve the overall clarity of the questionnaire. They can identify any logistical issues with the questionnaire, such as time constraints or unclear instructions.

Using student validators to assess face validity can be a helpful step in test development. However, it is important to combine their feedback with input from content experts and consider the limitations of students' experiences due to limited expertise and personal biases.

Pilot Testing

The validated questionnaire was piloted-tested to seventy-seven (77) second-year students of BSEd Mathematics of Agusan del Sur State College of Agriculture and Technology (ASSCAT). This pilot testing (construct validation) aims to evaluate each test item's quality, the test's empirical viability, and the suitability of the test construct that was created (Syahfitri et al.,2019)

Data Gathering and Analysis

This study focused on the development and validation of a test questionnaire for the coverage of final-term topics in Calculus 1. The Table of Specifications (TOS) served as the foundation for creating the test questionnaire. Moreover, the approved syllabus also served as the basis for the coverage. The test questionnaire was developed and then validated by the experts, who determined if the questionnaire had to be modified or enhanced. Opinions of three (3) experts in the field of mathematics and mathematics education from Agusan del Sur State College of Agriculture and Technology (ASSCAT) served as the validators for the content validity of the test questionnaire. For face validity, there were 15 student validators.

Pilot testing was done after the dean approved the conduct of the study. After the pilot testing, item analysis was done using KR-20, discrimination index, and difficulty index.

Research Instrument

Rating scales used in this study were adopted from other sources. The rating of each criterion for face validity was “yes or no”. The rating of each criterion for content validity is as follows: 5 – very satisfactory, 4 – satisfactory, 3 – undecided, 2 – poor satisfactory, and 1 – not satisfactory. Likert scales were used as the basis for the mean ranges per factor in the evaluation rating sheet, along with the corresponding description and interpretation that was used in this study. Tables 1, 2, and 3 are the rating scales for face validity, content validity, and reliability. Tables 4, 5, and 6 are rating scales for the index of discrimination, index of difficulty, and the decision for discrimination index and difficulty index.

Table 1. Interpretation and Acceptability of Percentage of Agreement (Face Validity).

Percentage of agreement	Strength of Agreement per question or overall	Action for each Question / entire tool
< 80	Poor	Restructure

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue VIII August 2024

80 - 90	Substantial	Substantial Revise
90 - 100	Full	Retain

(Desai & Patel, 2020)

Table 2. Content Validity Interpretation.

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.

(Pemintel, 2010; Nyutu et al., 2021)

Table 3. Reliability Coefficient Value Interpretation.

Reliability	Interpretation
0.90 and above	Excellent reliability; at the level of the best-standardized tests
0.80 – 0.89	Very good for a classroom test
0.70 – 0.79	Average, Good for a classroom test. There are probably a few items which could be improved.
0.60 – 0.69	Questionable, somewhat low. This test needs to be supplemented by other measures (e.g., more tests) to determine grades. There are probably some items which could be improved.
0.50 – 0.59	Poor, suggests the need for revision of the test, unless it is quite short (ten or fewer items). The test needs to be supplemented by other measures (e.g., more tests) for grading.
0.50 and below	Unacceptable, this test should not contribute heavily to the course grade, and it needs revision.

(Longe & Maharaj, 2023; University of Washington, 2024)

Table 4. Index of Discrimination (D-Index).

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 & Santos, 1997)

Table 5. Index of Difficulty (P-Index)

P value range	Interpretation
0.00 – 0.20	Very Difficult Item
0.21 – 0.40	Difficult Item
0.41 – 0.60	Moderately Difficult Item

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue VIII August 2024

0.61 – 0.80	Easy Item
0.81 and above	Very Easy item

(Padua & Santos, 1997)

Table 6. The Decision for Discrimination Index and Difficulty Index

Discrimination Index (D-Value)	Difficulty Index (P-Value)	Decision
0.20 and above (Acceptable)	0.26-0.75 (Acceptable)	Retained
0.19 and below (Not Acceptable)	0.26-0.75 (Acceptable)	Revised
0.20 and above (Acceptable)	Not within 0.26-0.75 (Not Acceptable)	Revised
0.19 and below (Not Acceptable)	Not within 0.26-0.75 (Not Acceptable)	Rejected

(Bermundo et al, 2004)

RESULTS AND DISCUSSION

Content and Face Validity of the Test Questionnaires

The experts meticulously reviewed the 100 test questions. Their feedback was incorporated into the final version, which shows an overall mean score of 4.36 for content validity, described as very high. This indicates that the quality of the developed test questionnaire is highly acceptable and comprehensively covers the intended content. This finding aligns with a previous study by Ocampo and Usita (2015), which stated that a mean range of 2.60 or higher for content validity suggests a "moderate to very much acceptable" rating. Therefore, the content validity of the developed Calculus I (final term) test was essential and accurately measured its intended content.

Table 7. Content Validity of the Test Questionnaires.

Content Validity	Mean	Verbal Description
Expert 1	4.36	Very High
Expert 2	4.38	Very High
Expert 3	4.34	Very High
Overall Mean	4.36	Very High

The face validity scored a mean of 0.98, indicating that the developed test questionnaire is comprehensive. This suggests that both respondents and experts believe the questionnaire effectively measures its intended purpose (Yaddanapudi & Yaddanapudi, 2019). This result is consistent with Desai and Patel's (2020) study that a range of 0.90–1.00 indicates an almost perfect agreement among raters, implying that the questionnaire or tool should be retained. Thus, the face validity of the developed Calculus 1 (final term) test was highly acceptable, with most raters agreeing that it appears to measure its intended purpose accurately.

Table 8. Face Validity of the Test Questionnaires.

Face Validity	Mean	Verbal Description
Rater 1	0.90	Full
Rater 2	1.00	Full
Rater 3	0.80	Substantial
Rater 4	1.00	Full
Rater 5	1.00	Full
Rater 6	1.00	Full
Rater 7	1.00	Full
Rater 8	1.00	Full
Rater 9	1.00	Full
Rater 10	1.00	Full
Rater 11	1.00	Full
Rater 12	1.00	Full

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue VIII August 2024

Rater 13	1.00	Full
Rater 14	1.00	Full
Rater 15	1.00	Full
Overall Mean	0.98	Full

Item Analysis of the Test

Item analysis is relevant to test formats where students must select the correct or best answer from the provided choices. Consequently, item analysis is most applicable to multiple-choice tests. Examinations that significantly impact students' course grades, such as midterms and final exams, or serve other critical decision-making purposes, should ideally be devoid of deceptive or ambiguous items. Unfortunately, identifying such issues is often challenging before the test is administered (Kunwar, 2018).

Item analysis procedures enable teachers to identify items that may be ambiguous, irrelevant, excessively easy or difficult, and lacking in discriminatory power. These procedures also contribute to improving the technical quality of an examination by highlighting non-functional options and signaling areas that need enhancement or removal. Item analysis also serves to support classroom instruction. For instance, item analysis in diagnostic testing pinpoints a student's areas of weakness and provides details for targeted remediation (Kunwar, 2018). After the pilot testing was conducted, the results of the test were used to determine the reliability of the test items. The internal consistency and reliability of the test items were determined using Kuder Richardson 20. The higher values for KR-20, which has a value range of 0 to 1, indicate higher reliability (Bobbit, 2022). Test items need a thorough evaluation to establish their quality before being deemed effective. In computerized procedures, values are generated swiftly, while manual computation takes a longer time.

After the test was administered to 77 students, checking and coding took place for the item analysis. Students' correct answers were coded as one (1) and incorrect answers as zero (0). The scores the students obtained were sorted from highest to lowest. The upper group was selected by getting 27% of the 77 test-takers, which were 21 top-rated students. The lower group was also chosen by getting 27% of 77 test-takers, or 21 students, as the lowest-rated students. Item difficulty was determined using the p formula, $D = \frac{R}{N}$, and item discrimination through the D formula, D value $= \frac{\text{Difference}}{27\% \text{ of N}}$ (Kara & Çelikler, 2015).

Table 9 shows the distribution of the items in terms of item difficulty. Results show that none of the items were found to be very difficult. Only 7% of the items have a difficulty index ranging from 0.20 to 0.40, which makes them difficult items. On the other hand, 27% of the items in the test questionnaire in Calculus 1 (final term) have a difficulty index ranging from 0.41 to 0.60, which makes them moderately difficult items. However, 54% of the items have a difficulty index ranging from 0.61 to 0.80, which makes them easy items. Finally, only 13% of the items have a difficulty index ranging from 0.81 and above, which makes them very easy.

Table 9. Item Distribution of the Test Questionnaire in Calculus 1 (Final Term) in terms of Difficulty Index.

Difficulty Index	*f*	%	Test Item Number/s	Verbal Interpretation
0.00 – 0.20	0	0%		Very Difficult Item
0.21 – 0.40	7	7%	13, 14, 19, 37, 64, 83, 95	Difficult Item
0.41 – 0.60	27	27%	3, 9, 12, 15, 17, 20, 22, 23, 28, 30, 34, 36, 45, 46, 48, 60, 61, 62, 69, 71, 81, 82, 92, 94, 96, 97, 98	Moderately Difficult Item
0.61 – 0.80	54	54%	4, 5, 6, 7, 10,11, 16, 21, 24, 27, 28, 31, 32, 33, 38, 39, 40, 41, 42, 43, 44, 47, 49, 50, 51, 54, 55, 56, 57, 58, 59, 63, 65, 66, 67, 70, 72, 74, 77, 78, 79, 80, 84, 85, 86, 87, 88, 89, 90, 91, 93, 99, 100	Easy Item
0.81 and above	13	13%	1, 2, 8, 18, 25, 26, 35, 52, 53, 68, 73, 75, 76	Very Easy item
Total	100	100%	100	

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue VIII August 2024

Table 10 presents the distribution of the items of the test questionnaire in Calculus 1 (final term) in terms of the items' discrimination index. More than 90% of the items are acceptable: 22% are very discriminating or very good items, 32% are discriminating or good items, and 44% are moderately discriminating or reasonably good items. Only 2% are not discriminating or marginal, and none of the items were found to be poor or questionable.

Table 10. Item Distribution of the Test Questionnaire in Calculus 1 (Final Term) in terms of Discrimination Index.

Difficulty Index	*f*	%	Test Item Number/s	Verbal Interpretation
-1.00 – -0.60	0	0%		Questionable Item
-0.59 - -0.21	2	2%	8, 13	Not Discriminating
-0.20 – 0.20	44	44%	1, 2, 7, 11,14, 15, 16, 18, 23, 25, 26, 29, 35, 38, 39, 40, 41, 42, 46, 47, 52, 53, 54, 55, 57, 63, 65, 66, 67, 68, 70, 73, 75, 79, 80, 83, 84, 85, 86, 87, 89, 90, 93	Moderately discriminating
0.21 – 0.59	32	32%	3, 4, 5, 6. 10, 12, 22, 24, 28, 31, 32, 43, 44, 48, 49, 51, 56, 59, 61, 62, 64, 71, 72, 74, 76, 77, 78, 88, 94, 95, 99, 100	Discriminating
0.60 – 1.00	22	22%	9, 17, 19, 20, 21, 27, 30, 33, 34, 36, 37, 45, 50, 58, 60, 69, 81, 82, 92, 96, 97, 98	Very Discriminating
Total	100	100%	100	

Based on the results presented in Table 11, 54% of the items in the test questionnaire in Calculus 1 (final term) can be retained. Only 25 or 25% of the items needed to be revised, while only 21 or 21% of the items were to be rejected.

Table 11. Summary of Item Analysis for the Test Questionnaire in Calculus 1 (Final Term).

Final Evaluation/Remark	*f*	%	Item Number/s
Items to be Retained	54	54%	3, 4, 5, 6, 9, 10, 11, 12, 17, 19, 20, 21, 22, 24, 27, 28, 30, 31, 32, 33, 34, 36, 37, 43, 45, 48. 49, 50, 51, 56, 58, 59, 60, 61, 62, 64, 69, 71, 72, 74, 77, 78, 81, 82, 88, 91, 92, 94, 95, 96, 97, 98, 99, 100
Items to be Revised	25	25%	7, 13, 14, 15, 16, 23, 29, 39, 40, 41, 42, 44, 46, 47, 54, 57, 63, 67, 70, 76, 79, 83, 86, 87, 89
Items to be Rejected	21	21%	1, 2, 8, 18, 25, 26, 35, 38, 52, 53, 55, 65, 66, 68, 73, 75, 80, 84, 85, 90, and 93
Total	100	100%	100

Reliability of the Test

The validated test questionnaire was pilot-tested to the 2nd-year BSE Mathematics students, and items were analyzed to omit too difficult, very easy, and misleading questions. Kuder Richardson 20 (KR-20) was used to find the internal consistency of the tests. Descriptive statistics were obtained from the test, and KR-20 measures the reliability of the test with binary variables (1: correct answers and 0: incorrect answers). In addition, KR-20 ranges from 0.0 to 1.0; closer to 0 indicates very poor reliability, and closer to 1.0 indicates high reliability.

The reliability coefficient was calculated at 0.895, which is closer to 1.0, which indicates high reliability. This implies that the developed test questionnaire in Calculus 1 (final term) was "good" and consistently measured what it intended to measure, which means that if the same students were to retake the test under similar conditions, their scores would be relatively consistent over time. This result corresponds to the study by Ermie (2017): if the result is closer to 0, that result indicates very low and poor reliability; on the other hand, if the result of the Kuder Richardson 20 (KR-20) is closer or $\geq .070$ the result indicates acceptable to very high reliability. This means that the items have acceptable value (Kilic, 2016) and are very good for classroom use (University of Washington, 2020). Hence, the developed test questionnaire can be used for classroom assessment

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XI Issue VIII August 2024

and is reliable for measuring the knowledge and skills of the students in Calculus 1 (final term). In addition, a test must be valid for it to be reliable (Taherdoost, 2016).

CONCLUSION AND RECOMMENDATION

This study aimed to develop a reliable and valid test questionnaire for the final term examination of Calculus 1. The researchers based the test on the approved syllabus topics and ensured it aligned with student needs.

The analysis showed the test had strong validity. Experts deemed the content relevant and appropriate for the subject matter ("very high" content validity). Additionally, the overall structure and format of the test were judged to be clear and easy to understand ("full" face validity). The test also demonstrated good reliability, indicating consistent results when administered repeatedly considering similar conditions.

This study offers valuable insights for educators and future researchers. The constructed test can be used as an assessment tool for the Calculus 1 final examination for students taking up a Bachelor of Secondary Education major in Mathematics. The developed test questionnaire can serve as a helpful resource for instructors crafting their tests in Calculus 1. However, after the test is implemented, it is essential to continuously monitor its performance including collecting feedback from test-takers and administrators and conducting periodic re-analyses of item and test performance.

REFERENCES

1. American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). Standards for educational and psychological testing. https://www. apa.org/science/programs /testing/standards.
2. Anderson, L.W. & Krathwohl, D.R. (2001). Learning and teaching: New Bloom's taxon-omy for the 21st century. Center for the Study of Education at Illinois State University.
3. Angeles, M.R., Fajardo, A.C., & Tanguilig III, B.T. (2015). E-Math Version 2.0, a Learning Management System as a Math Reviewer Tool for Engineering Students in the Philippines. International Journal of Engineering and Technical Research, 3(2), 18-21. http://www.acade-mia.edu/11863180/E-aLearningManagement System as a Math Reviewer Tool for Engineering Students in the Philippines
4. Bee, D.T. & Murdoch-Eaton, D. (2016). Questionnaire design: the good, the bad and the pitfalls. Archives of Disease in Childhood-education and Practice Edition, 101(4), 210–212. https://doi.org/10.1136-/archdischild-2015-309450
5. Bermundo, C., Bermundo, A. & Ballester, R. (2004). Test checker and item analyzer ver. 2.0. Naga City, Philippines.
6. Biddix, J.P. (2018). Research methods and applications for student affairs. John Wiley & Sons.
7. Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. Assessment in Education: Principles, Policy & Practice, 25(6), 551–575. https://doi.org/10.1080/-0969594X.2018.1441807
8. Bobbitt, Z. (2022). Kuder-Richardson Formula 20 (Definition & examp le). Statol- ogy. https://www.statology.org/kuder-richardson-20/
9. Bolarinwa, O.A. (2015). Principles and methods of validity and reliability testing of ques-tionnaires used in social and health science researches. Nigerian Postgraduate Medical Journal, 22(4), 196-201.
10. Brame, C. (2013). Writing good multiple choice test questions. Center for Teaching Vanderbilt University.
11. Butakor P. (2022). Using Classical Test and Item Response Theories to Evaluate Psychometric Quality of Teacher-Made Test in Ghana. European Scientific Journal, ESJ, 18 (1), 139. https://doi.org/ 10.19044/esj.2022.v18n1p139
12. Bujang, M.A., Omar E.D., & Baharum N.A. (2018). A review on sample size determinat-ion for Cronbach's alpha test: a simple guide for researchers. The Malaysian journal of medical sciences: MJMS 25.6: 85.
13. Ding, L. & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. Physical Review Special Topics. Physics Education Research, 5(2). https://doi.org/10.1103/physrevs tper.5.020103

14. Effandi-Zakaria, E. (2013). The development and validation of conceptual and procedural understanding test for integral calculus. International Journal of Applied Mathematics and Statistics, 42(3), 49-60.
15. Ermie, E. (2017). Psychometrics 101: Know what your assessment data is telling you. Sli-des presented at the Exam Soft Assessment Conference 2017, Denver, Colorado.
16. Fatimah, S. & Yerizon (2019). Analysis of difficulty learning calculus subject for mathematical education students. International Journal of Scientific & Technology Research, 8(03). https://www.ijstr.org/fina-lprint/mar2019/Analysis-Of-Difficulty-Learning-Calculus-SubjectFo-r-MathematicalEducationStude nts.pdf.
17. Fraenkel, J.R. & Wallen, N.E. (2009). How to Design and Evaluate Research in Education (7th ed.). New York: McGraw-Hill Companies
18. Gholami, A.R. & Soleimani, H.R. (2022). Fairness in classroom assessment: development and validation of a questionnaire. Language Testing in Asia, 16(3), 274-299.
19. 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
20. Graham, M.J. (2012). A comprehensive Framework for the Development and Validation of Assessment Tools in Education. Educational Assessment, Education Program Evaluation, and Student Success, 23(4), 399-418.
21. Hambleton, R. K. & Jones, R. W. (1993). An NCME Instructional Module on Educational Measurement: issues and practice, Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development, 12(3),38-47. (n.d.). Retrieved from https://www.scie pub.com/referen-ce/115 927
22. Harlen, W. (2013). The role of assessment in the learning process. Assessment in Educa-tion: Principles, Policy & Practice, 20(1), 1-15.
23. Jovanovska, J. (2018). Designing effective multiple-choice questions for assessing learning outcomes. Infotheca, 18(1), 25-42.
24. 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. Retrieved from www.iiste.org
25. Kilic, S. (2016). Cronbach's alpha reliability coefficient. Psychiatry and Behavioral Sciences, 6(1), 47.
26. Kline, P. (2013). The Handbook of Psychological Testing (3rd ed.). Pearson.
27. Kumara, B.T.G.S., Brahmana, A., & Paik, I. (2019). Bloom's taxonomy and rules based question analysis approach for measuring the quality of examination papers. International Journal of Knowledge Engineering, 5(1), 2-6.
28. Kunwar, R. (2018). Development and standardization process of Journal of Current Research, 10, (11), 75451-75455.
29. 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/Journal of Medives : Journal of Mathematics Education IKIP Veteran Semarang, 7(1), 117. https://doi.org/10.31331/medivesveteran.v7i1.2332
30. Mamolo, L.A. (2021). Development of an Achievement Test to Measure Students' Competency in General Mathematics. Anatolian Journal of Education, 6(1), 79-90.
31. Mutakin, T.Z. (2013). Analisis kesulitan belajar kalkulus 1 mahasiswa teknik informatika. Jurnal Formatif, 3(1), 49–60. https:// dx.doi.org/10.30998/ formatif. V 3i1.113.
32. Nasreen, A., Ahmad, A.U., & Sabiha, I. (2019). Development and validation of multiple-choice test geometry part of mathematics for secondary class. Global social sciences reviews, 4(2), 283-292.
33. Nedeau-Cayo, R., Laughlin, D., Rus, L., & Hall, J. (2013). Assessment of item-writing flaws in multiple-choice questions. Journal for nurses in professional development, 29(2), 52-57.
34. Nuraeni, R. (2018). Perbandingan kemampuan komunikasi matematis mahasiswa antara yang mendapatkan pembelajaran group investigation dengan konvensional. Mosharafa: Jurnal Pendidikan Matematika, 7(2), 219–228. http://journal.insti-tutpendidikan.ac.id/index.php/mosharaf
35. Nyutu, E. N., Cobern, W. W., & Pleasants, B. A-S. (2021). Correlational study of student perceptions of their undergraduate laboratory environment with respect to gender and major. International Journal of Education in Mathematics, Science, and Technology (IJEMST), 9(1), 83-102. https://doi.org/10.46328/ijemst.1182

36. Ocampo, R. & Usita, N.P. (2015). Development of Lubeg (Syzygium Lineatum (Roxb.) Merr. & Perry) processed products. ResearchGate. https://www.researchgate.net/-publication/361901665.
37. Oducado, R.M. (2020). Survey instrument validation rating scale. Available at SSRN 3789575.
38. Oktaviana, D. & Susiaty, U.D. (2020). Development of test instruments based on revision of Bloom's taxonomy to measure the students' higher-order thinking skills. Jurnal Ilmiah Pendidikan Matematika (JIPM), 9(1), 21. https://doi.org/10.25273-/jipm. V 9i1.5638
39. Patel, N. & Desai, S. (2020). ABC of face validity for questionnaire. Int J Pharm Sci Rev Res, 65, 164-8.
40. Pimentel, J.L. (2010). A note on the usage of Likert Scaling for research data analysis. USM R&D Journal, 18(2), 109-112.
41. Przymuszała, P., Piotrowska, K., Lipski, D., Marciniak, R., & Cerbin-Koczorowska, M. (2020). Guidelines on writing multiple choice questions: a well-received and effective faculty development intervention. SAGE Open, 10(3), 2158244020-947432.
42. Raupu, S. (2020). Learning difficulties in solving calculus test. https:// ww w.semanticscholar.org/paper/LEARNINGDIFFICULTIESINSO-LVI NG-CALCULUS-TESTS Ra-upu-Thalhah/87177677 6d5d802b8-d80e-ff506275fac d15e4b7e
43. Şahin, Ş., Yıldırım, Y., & Boztunç, Ö. (2023). Examining the achievement test development process in the educational studies. https://dergipark.org.tr/en/download/-article-file/2626246.
44. Santos, A. & Reyes, B. (2019). Significance of precise and culturally relevant assessment tools in the Filipino educational setting. Journal of Educational Research, 45(3), 123-145.
45. Sireci, S.G. (2016). Test development and evaluation. In R. L. Brennan (Ed.), *Educational measurement* (4thed., pp.31-80). American Council on Education.
46. Smith, A.B. & Johnson, C.D. (2018). Assessing student learning: A guide to the principles, goals, and methods of successful classroom assessment. Academic Press.
47. Straub, D., Boudreau, M.C. & Gefen, D. (2014). Validation guidelines for IS positivist research. Communications of the Association for Information Systems, 13, 380-427.
48. Sugiarti, L. (2016). Kesulitan siswa dalam menyelesaikan soal operasi bentuk aljabar. Prosiding Seminar Nasional Etnomatnesia, 232–330.
49. 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
50. Taherdoost, H. (2016). Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research. How to test the validation of a questionnaire/survey in a research.
51. University of Washington (2024). Understanding item analysis. https:/-/ww w.washington.edu/assessment/scanning-scoring/scor-ing/reports/ ite-manalysis.
52. Yaddanapudi, S. & Yaddanapudi, L.N. (2019). How to design a questionnaire. Indian journal of anaesthesia, 63(5), 335.
53. Yuwono, M.R. (2016). Analisis kesulitan mahasiswa dalam menyelesaikan soal geometri berdasarkan Taksonomi Bloom dan alternatif pemecahannya. Beta Jurnal Tadris Matematika, 9(2), 111. https://doi.org/10.20414/ betajtm. V 9i2.7.
54. Zapata-Rivera, L.F. & Suescun, C.A. (2015). Game-based assessment for radiofrequency circuits courses in engineering. Proceedings of the 2015 IEEE Frontiers in Education Conference (FIE)https://doi.org/10.1109/FIE.2015.7344108