Kontinu: Jurnal Penelitian Didaktik Matematika
E-ISSN: 2656-5544
P-ISSN: 2715-7326
Vol. 10, No. 1: May 2026
Page 176-190

Evaluating Test Quality in English for Mathematics Course: An Item Analysis Approach

Hedia Rizki[1)*], Reri Seprina Anggraini[2)], Indra Bagus Satrio[3)]

UIN Sulthan Thaha Saifuddin Jambi, Indonesia[1)], Institut Agama Islam Negeri Kerinci, Indonesia[2)], Universitas Jambi, Indonesia[3)]

[*]Corresponding e-mail: hediarizki@uinjambi.ac.id

Abstract. *This study evaluated the quality of an English for Mathematics course administered to 35 third-semester students in the Mathematics Education Study Program at UIN Sulthan Thaha Saifuddin Jambi using Classical Test Theory. The 50-item multiple-choice test covered eight mathematical topics, including numbers, rational numbers, powers and roots, statistics and probability, logic, algebra, geometry, and Measurement. Item analysis examined difficulty indices, discrimination indices, reliability (KR-20), and distractor effectiveness. Results revealed excellent reliability ($r_{11} = 0.911$) and balanced difficulty distribution: 36% easy items ($P > 0.70$), 52% moderate items (0.30-0.70), and 12% difficult items ($P < 0.30$). The majority of items (86%) demonstrated acceptable discrimination capacity, with 56% achieving excellent discrimination ($D \geq 0.40$). However, critical concerns emerged: difficult items clustered in geometry and measurement topics (items 39-44, 48), 14% of items showed poor discrimination, and 27% of distractors proved ineffective. Findings indicate that the instrument possesses strong overall technical quality, while requiring targeted revisions to poorly discriminating items and ineffective distractors. The study recommends enhanced instructional support for geometry and measurement topics, qualitative analysis of student reasoning, and future research employing differential item functioning analysis to distinguish mathematical ability from language proficiency effects in English for Mathematics assessment.*

Keywords: *Classical Test Theory, English for Mathematic, Evaluation*

INTRODUCTION

Mathematics as a universal language employs uniform symbols and notation across cultures; however, comprehension of mathematical concepts depends heavily on linguistic competence, particularly when conveyed in English (Coetzee & Mammen, 2017; Garrison & Kanuka, 2004). In the era of higher education globalization, the ability to understand mathematical terminology in English has become an essential competence for students, especially in science, technology, engineering, and mathematics (STEM) fields. This is reflected in the dominance of English-language academic literature across various international scientific repositories (Buana, A. N. H. et al., 2025; Hyland, 2006).

English for Mathematics (EFM) represents a form of English for Specific Purposes (ESP) that integrates language learning with mathematical content (Dudley-Evans & St. John, 1998). Building on ESP frameworks, Prediger, S. et al. (2016) propose that effective bilingual mathematics instruction requires a purposeful connection between everyday language registers and mathematical registers across languages, suggesting that code-switching and register-switching serve distinct pedagogical functions in multilingual mathematics classrooms. This complexity creates dual challenges for learners: mastering mathematical concepts while simultaneously acquiring the academic language that represents them. Previous research has revealed various linguistic difficulties students face in understanding mathematics in English. Misinterpretation of mathematical texts more frequently stems from limited language comprehension rather than computational errors (Bailey, A. L. et al., 2015).

Furthermore, Barwell (2018) argues that in multilingual mathematics classrooms, language functions not merely as a transparent medium for mathematical communication but also as a fundamental source of meaning-making, emphasizing that linguistic choices shape how mathematical concepts are understood and negotiated. A more comprehensive study by Buana, A. N. H. et al. (2025) identified three major categories of linguistic difficulties, including words such as "root," "function," and "table," which carry different meanings in mathematical versus general contexts. According to Schleppegrell (2007), a systematic review of linguistic challenges in mathematics education identified dense noun phrases, logical connectors, and technical vocabulary as primary obstacles that impede students' mathematical comprehension, particularly in second-language contexts. Similar findings were reported by Mullen (2009), who emphasized that many students fail standardized mathematics tests not because of conceptual inability but because of insufficient reading skills and comprehension of academic vocabulary.

In Indonesia, research by Cesaria, A., & Fitri (2014) demonstrated that Mathematics Education students at STKIP PGRI West Sumatra experienced significant obstacles in English for Mathematics courses due to limited teaching materials and monotonous instructional methods. Research by Kurniawati (2023) found that Indonesian mathematics teachers struggled to master technical vocabulary, sentence structure, and pronunciation when teaching bilingual mathematics, which affected student comprehension. Although various studies have explored linguistic difficulties in English-medium mathematics instruction and developed learning strategies, a significant gap exists in evaluating the quality of assessment instruments used to measure English for Mathematics competence.

Similar challenges have been documented globally (Setati, M. & Adler, 2000). Observations have shown that in multilingual South African mathematics classrooms, tensions arise when teachers must navigate between learners' home languages and the official language of instruction, creating communicative dilemmas that parallel those faced by Indonesian mathematics educators.

No comprehensive research has analyzed whether English for Mathematics test instruments genuinely measure mathematical ability or primarily assess linguistic competence alone. According to Taifour & Alabdulaziz (2025), linguistic demands in context-based mathematics problems often create discrepancies between students' abilities and instructors' expectations. Furthermore, in-depth analysis of the difficulty index, discrimination index, and reliability of the English for Mathematics test instruments remains severely limited. This study combines Classical Test Theory (CTT) and Item Response Theory (IRT) for a comprehensive evaluation of English for Mathematics test instruments. As DeMars (2018) explicates, while Classical Test Theory provides accessible metrics for evaluating test quality at the item and test level, it assumes measurement precision is constant across ability levels—a limitation that Item Response Theory addresses by modeling item characteristics as a function of latent trait levels. The CTT approach analyzes difficulty level, discrimination power, and distractor effects (Ebel, R. L., & Frisbie, 1991), whereas IRT explores item characteristics across various levels of student ability (Hambleton & Swaminathan, 1985). The integration of these two approaches enables a more robust evaluation than previous research, which typically employed only one approach.

RESEARCH METHOD

This study employed a quantitative descriptive approach to evaluate the quality of the English for Mathematics midterm examination through item analysis and to identify students' difficulties across different mathematical topics. The research was conducted during the odd semester of the 2025/2026 academic year d. The participants in this study were 35 third-semester undergraduate students enrolled in the English for Mathematics course.

Kontinu: Jurnal Penelitian Didaktik Matematika
E-ISSN: 2656-5544
P-ISSN: 2715-7326
Vol. 10, No. 1: May 2026
Page 176-190

All participants were selected through census sampling, as all students taking the course during the study period were included in the analysis. The participants had completed approximately 7 weeks of English instruction for Mathematics before taking the midterm examination. The instrument used in this study was a multiple-choice midterm examination consisting of 50 items, administered through Google Forms. The examination was designed to assess students' understanding of mathematical terminology and concepts in English across eight topic areas.

Table 1. Distribution of Questions by Topic

No.	Topic	Subtopic	Number of Questions
1.	Numbers and Computation	Integers, LCM, GCF, absolute value, composite numbers, divisibility, mixed numbers, and decimals	8 items
2.	Rational Numbers	Multiplicative inverse and reading fractions	2 items
3.	Powers and Roots	Base, exponent, squared, square root, and power operations	5 items
4.	Statistics and Probability	Irrational numbers, graphs, independent events, compound events, and mode	5 items
5.	Logic	Mutually exclusive events, statements, truth values, conditional statements, hypotheses, converse, counterexamples, and set operations	8 items
6.	Algebra	Variables, coefficients, like terms, distributive property, quadratic equations, functions, intercepts, slope, and expressions	9 items
7.	Geometry	Points, angles, lines, parallelograms, rectangular prisms, edges, and spheres	7 items
8.	Measurement	Circumference, measurement processes, length, scale, mass, and volume	6 items

Each item was worth 2 points, with a total possible score of 100. The examination was developed based on the course syllabus and learning objectives for English for Mathematics. All items were constructed in multiple-choice format with four options (A, B, C, D) to facilitate objective scoring and statistical analysis.

Kontinu: Jurnal Penelitian Didaktik Matematika
E-ISSN: 2656-5544
P-ISSN: 2715-7326
Vol. 10, No. 1: May 2026
Page 176-190

Data Analysis

Data analysis was conducted in several stages using both Microsoft Excel and statistical software. The difficulty index was calculated using the formula.

$$P = \frac{B}{N}$$

Where:

P = difficulty index

B = number of students who answered correctly

N = total number of students

Interpretation criteria (Arikunto, 2013):

$P > 0.70$ = Easy item

$0.30 \leq P \leq 0.70$ = Moderate item

$P < 0.30$ = Difficult item

The discrimination index was calculated using the upper-lower group method:

$$D = \frac{(BA - BB)}{n}$$

Where:

D = discrimination index

BA = number of correct answers in the upper group

BB = number of correct answers in the lower group

n = number of students in each group

Test reliability was estimated using the Kuder-Richardson Formula 20 (KR-20), which is appropriate for dichotomously scored items:

$$\boldsymbol{r_{11}} = \left(\frac{\boldsymbol{k}}{(\boldsymbol{k}-\mathbf{1})}\right)\left(\frac{(\boldsymbol{\sigma_t^2} - \sum \boldsymbol{pq}}{\boldsymbol{\sigma_t^2}}\right)$$

Where:

- r_{11} = reliability coefficient
- k = number of items (50)
- σ^2_t = variance of total test scores
- p = proportion of students answering correctly
- q = proportion of students answering incorrectly (1-p)
- Σpq = sum of pq for all items

Interpretation criteria (Nunnally, 1978):

- $r_{11} > 0.70$ = High reliability
- $0.40 \le r_{11} \le 0.70$ = Moderate reliability
- $r_{11} < 0.40$ = Low reliability

RESULT AND DISCUSSION

Item Difficulty Analysis

Based on the analysis of 35 students who took the English for Mathematics midterm examination, the difficulty level of each item was calculated using the formula $P = \frac{B}{N}$, where B is the number of students who answered correctly, and N is the total number of students. The analysis results showed that the majority of items (52%) had moderate difficulty levels ($0.30 \le P \le 0.70$), 36% were classified as easy ($P > 0.70$), and 12% as difficult ($P < 0.30$), as presented in Table 2.

Table 2. Difficulty Index

Category	Range	Number of Items	Percentage	Item Numbers
Easy	P > 0.70	18	36%	1, 2, 6, 7, 10, 11, 12, 13, 14, 18, 20, 22, 28, 29, 31, 45, 46, 50
Moderate	0.30-0.70	26	52%	3, 4, 5, 8, 9, 15, 16, 17, 19, 21, 23, 24, 25, 26, 27, 30, 32, 33, 34, 35, 36, 37, 38, 41, 47, 49
Difficult	P < 0.30	6	12%	39, 40, 42, 43, 44, 48

This distribution aligns with the ideal psychometric recommendation that a test should contain predominantly moderate items (40-60%) to maximize information about students' ability levels (Ebel, R. L.). However, the concentration of difficult items ($P < 0.30$) specifically within the Geometry (items 39-44) and Measurement (item 48) topics warrants careful interpretation and aligns with existing literature documenting the particular difficulty of spatial and Measurement vocabulary in second-language mathematics education (Bailey, A. L. et al., 2015; Coetzee & Mammen, 2017).

This finding aligns with (Haag et al., 2013) The potential confounding of linguistic and mathematical competence identified in our study echoes Abedi and Lord's (2001) finding that unnecessary linguistic complexity in mathematics assessments

Kontinu: Jurnal Penelitian Didaktik Matematika
E-ISSN: 2656-5544
P-ISSN: 2715-7326
Vol. 10, No. 1: May 2026
Page 176-190

systematically disadvantages English language learners, suggesting that test modifications reducing linguistic load can improve validity without compromising mathematical content evidence that academic language features particularly in geometrically oriented word problems pose distinct challenges for second-language learners, with linguistic complexity accounting for significant variance in mathematics performance beyond general language proficiency. The potential confounding of linguistic and mathematical competence identified in our study echoes. Abedi's J., & Lord's (2010) finding that unnecessary linguistic complexity in mathematics assessments systematically disadvantages English language learners, suggesting that test modifications reducing linguistic load can improve validity without compromising mathematical content

The clustering of difficult items in geometry and measurement topics suggests these domains present compounded linguistic and conceptual challenges for students learning mathematics in English. This observation is consistent with Buana, A. N. H. et al., who identified geometric terminology as particularly problematic due to polysemous words (e.g., "plane," "volume," "area") that carry different meanings in mathematical versus everyday contexts. Furthermore, measurement concepts introduce unit-specific vocabulary and conversion terminology that may be less familiar to Indonesian students who primarily use the metric system in their native language instruction (Cesaria, A. & Fitri, 2014).

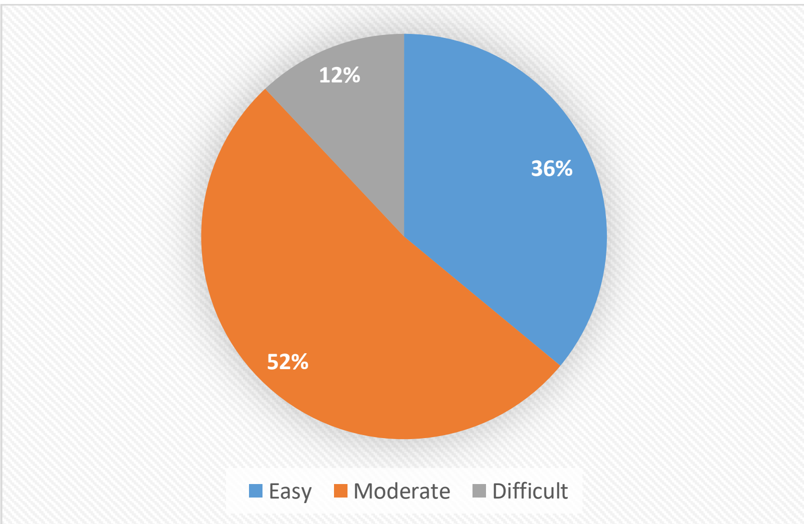

Figure 1. Distribution of Item Difficulty Levels

Kontinu: Jurnal Penelitian Didaktik Matematika
E-ISSN: 2656-5544
P-ISSN: 2715-7326
Vol. 10, No. 1: May 2026
Page 176-190

Item Discrimination Analysis

The discrimination index was calculated by comparing the upper group (top 27%; 9 students with the highest scores) and the lower group (bottom 27%; 9 students with the lowest scores). The discrimination analysis revealed that 56% of items achieved excellent discrimination ($D \geq 0.40$), 30% demonstrated acceptable discrimination ($0.20 \leq D < 0.40$), and 14% showed poor discrimination ($D < 0.20$), as shown in Table 3.

Table 3. Distribution of Item Discrimination

Category	Range	Number of Items	Percentage	Item Quality
Good	$D \geq 0.40$	28	56%	Discriminates very well
Acceptable	0.20-0.39	15	30%	Discriminates adequately
Poor	$D < 0.20$	7	14%	Needs revision or replacement

This strong discrimination profile indicates that the majority of items successfully differentiate between high-achieving and low-achieving students, which is essential for a valid assessment instrument (Arikunto, 2013; Hambleton & Swaminathan, 1985). However, the 14% of items with poor discrimination require critical examination. Interestingly, the poorly discriminating items fell into two distinct categories: items that were too easy ($P > 0.80$; e.g., 7, 22, 28, 46) and those that were excessively difficult ($P < 0.30$; e.g., 39, 42). According to Martiniello (2008), even when mathematical content remains constant, linguistic modifications to problem presentation significantly affect English language learners' performance, suggesting that items with poor discrimination in our study may contain linguistic features that obscure rather than assess mathematical competence.

Table 4. Items with Problematic Discrimination ($D < 0.20$)

Item No.	Topic	D Value	P Value	Identified Problem
7	Fractions	0.11	0.89	Too easy
18	Probability	0.15	0.77	Ineffective distractors
22	Statistics	0.00	0.91	Nearly all students answered correctly.
28	Set Theory	0.11	0.83	Too easy
39	Geometry	0.17	0.23	Too difficult for all groups
42	Solid Geometry	0.11	0.26	Linguistic problems at all levels
46	Measurement	0.11	0.86	Does not discriminate by ability

Kontinu: Jurnal Penelitian Didaktik Matematika
E-ISSN: 2656-5544
P-ISSN: 2715-7326
Vol. 10, No. 1: May 2026
Page 176-190

For overly easy items, the ceiling effect likely prevented discrimination—when nearly all students answer correctly, the item cannot distinguish ability levels (Nunnally, 1978). These items may test basic vocabulary recognition rather than deeper conceptual understanding. Conversely, the extremely difficult items with poor discrimination may contain ambiguous wording, cultural bias, or content that extends beyond the instructional scope. This phenomenon has been documented in multilingual mathematics education research, where linguistic complexity can obscure mathematical competence (Bailey, A. L. et al., 2015; Mullen, 2009; Taifour, A. & Alabdulaziz, 2025)

Reliability Assessment

Test reliability was calculated using the Kuder-Richardson Formula 20 (KR-20), which is appropriate for multiple-choice instruments with dichotomous scoring (correct=1, incorrect=0). The following presents the reliability results.

Table 5. Reliability Calculation Results

Parameter	Value
Number of Items (k)	50
Mean Score	31.03 (62.06%)
Standard Deviation	10.24
Variance (σ^2)	104.86
Σpq	10.37
KR-20 (r_{11})	0.911

The KR-20 reliability coefficient of 0.911 indicates excellent internal consistency, well above the conventional threshold of 0.70 for acceptable reliability (Nunnally, 1978). This high reliability indicates that the test items collectively measure a cohesive construct—in this case, English for Mathematics competence—with minimal measurement error. The reliability is particularly impressive given the relatively small sample size (N = 35). It suggests that if this examination were administered to similar groups of third-semester students in mathematics education, it would yield consistent and dependable results.

The mean score of 31.03 out of 50 (62.06%) and the standard deviation of 10.24 indicate adequate score variability across the student sample—a prerequisite for achieving high reliability (Hambleton & Swaminathan, 1985). However, it is important to acknowledge that high reliability does not guarantee validity. While the test consistently measures something, further investigation is needed to ensure that the construct being measured is indeed the integrated ability to understand mathematics through English rather than primarily English language proficiency alone (Dudley-Evans & St. John, 1998).

Distractor Effectiveness Analysis

A distractor effectiveness analysis was conducted to determine whether incorrect answer choices (distractors) effectively attracted lower-performing students. The analysis revealed that 73% of distractors (110 out of 150) functioned effectively or very effectively, representing strong overall item construction quality (Haladyna, T. M., & Rodriguez, 2013).

Table 6. Categories of Distractor Effectiveness

Category	Criteria	Number of Distractors	Percentage
Highly Effective	Selected >15%, more by lower group	42	28%
Effective	Selected 5-15%, more by the lower group	68	45%
Minimally Effective	Selected <5% or does not discriminate between groups	28	19%
Non-Functional	Not selected at all	12	8%

Effective distractors are plausible incorrect options that attract students with incomplete understanding without misleading knowledgeable students—a hallmark of well-crafted multiple-choice items (Ebel, R. L., 1991). The finding that 28% of distractors were highly effective indicates that these options successfully represented common misconceptions or partial knowledge states.

However, the 27% of distractors that proved ineffective or non-functional require examination and revision. The 8% of distractors that attracted zero responses suggest these options were implausible even to struggling students, possibly due to obvious errors, extreme incorrectness, or grammatical inconsistencies. Such non-functional distractors effectively reduce a four-option item to a three-option or even two-option item, which decreases guessing penalty and potentially inflates scores artificially (Gronlund, N. E., & Waugh, 2009).

In the context of English for Mathematics, ineffective distractors could arise from several sources: translation artifacts from Indonesian mathematical pedagogy, culturally specific mathematical conventions, or poorly conceived foils that do not align with students' actual knowledge gaps (Kurniawati, 2023). The significant performance variation across mathematical topics observed in our study may reflect developmental patterns documented by Swanson et al. (2015), who found that English language learners' mathematics achievement is mediated by working memory capacity and literacy development, with complex linguistic demands creating differential cognitive load across content areas. Revising these distractors should involve qualitative analysis of student reasoning, possibly through cognitive interviews or think-aloud protocols, to identify authentic alternative conceptions that merit representation in revised options (Garrison & Kanuka, 2004).

Performance by Mathematical Topics

To provide a more comprehensive understanding of students' proficiency in English mathematical terminology, the examination items were categorized into eight major mathematical topics. The performance comparison across these topics is presented in Table 7.

Table 7. Classification of Item Quality and Recommendations

Category	Criteria	Number	Percentage	Recommendation
Excellent	P: 0.30-0.70, D ≥ 0.40, Effective distractors	22	44%	Retain
Good	P: 0.30-0.70, D: 0.20-0.39	12	24%	Retain with minor revision to distractors
Minor Revision Needed	Too easy (P > 0.70) but adequate D	8	16%	Increase difficulty or use as warm-up items
Major Revision Needed	D < 0.20 or ineffective distractors	5	10%	Complete revision: improve distractors, increase linguistic complexity
Replacement Needed	P < 0.30, low D, linguistic bias	3	6%	Replace or significantly modify: simplify linguistic structure, add scaffolding.

The topic-level analysis reveals significant variation in student performance, with Rational Numbers achieving the highest mastery (80.0%) and Geometry demonstrating

Kontinu: Jurnal Penelitian Didaktik Matematika
E-ISSN: 2656-5544
P-ISSN: 2715-7326
Vol. 10, No. 1: May 2026
Page 176-190

the lowest performance (47.4%). This finding provides empirical evidence directly addressing the research objective of evaluating test quality and identifying areas requiring instructional intervention.

The notably poor performance in Geometry (47.4%) aligns with the difficulty index analysis, which shows a concentration of difficult items in this domain. This convergent evidence from multiple psychometric indicators strengthens the validity of our findings and confirms that geometry terminology presents particular challenges for Indonesian students learning mathematics in English (Buana, A. N. H. et al., 2025; Coetzee & Mammen, 2017). The spatial reasoning required in geometry, combined with specialized vocabulary, creates a compounded cognitive load for students learning English as a foreign language (Hyland, 2006).

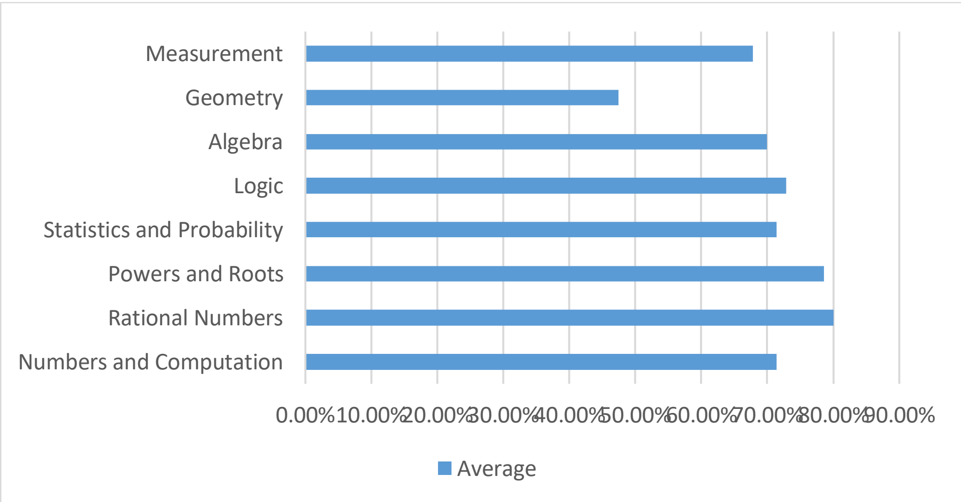

Figure 2. Performance Comparison Across Topics

The comprehensive item analysis of the English for Mathematics midterm examination reveals several critical insights into the instrument's quality and students' performance patterns. The discussion is organized around four key dimensions: difficulty distribution, discrimination capacity, reliability, and distractor effectiveness.

CONCLUSION

This study evaluated the quality of an English for Mathematics midterm examination administered to 35 third-semester students using Classical Test Theory item analysis. Addressing the research objectives, the findings demonstrate that the 50-item test possesses strong overall psychometric quality while revealing specific areas requiring improvement. First, the examination achieved balanced difficulty distribution (52% moderate, 36% easy, 12% difficult items), though difficult items clustered in geometry and measurement topics, indicating compounded linguistic and conceptual challenges (Buana et al., 2025; Coetzee & Mammen, 2017). Second, 86% of items demonstrated acceptable discrimination capacity, with 56% achieving excellent discrimination ($D \geq 0.40$), confirming effective differentiation of student ability levels. Third, the instrument achieved excellent reliability (KR-20 = 0.911), substantially exceeding acceptable thresholds and demonstrating consistent Measurement of English for Mathematics competence (Nunnally, 1978). Fourth, 73% of distractors functioned effectively, though 27% require revision to enhance item quality. Fifth, topic-specific analysis revealed significant performance variation, with Rational Numbers achieving the highest mastery (80.0%) and Geometry demonstrating the poorest performance (47.4%), providing clear targets for instructional intervention.

REFERENCES

Abedi, J., & Lord, C. (2010). The Language Factor in Mathematics Tests. *Applied Measurement in Education*, *14*(3), 219–234. https://doi.org/https://doi.org/10.1207/S15324818AME1403_2

Arikunto, S. (2013). *Dasar-dasar evaluasi pendidikan (2nd ed.)*. Bumi Aksara.

Bailey, A. L., Blackstock-Bernstein, A., & Heritage, M. (2015). At the intersection of mathematics and language: Examining mathematical strategies and explanations by grade and English learner status. *The Journal of Mathematical Behavior*, *40(Part A)*, 60–28.

Barwell, R. (2018). From Language as a Resource to Sources of Meaning in Multilingual Mathematics Classrooms. *The Journal of Mathematical Behavior*, *50*, 155–168. https://doi.org/https://doi.org/10.1016/j.jmathb.2018.02.007

Buana, A. N. H., Waru, M. V., & Aditia, R. (2025). Analisis Kesulitan Linguistik dalam Pemahaman Istilah Matematika Berbahasa Inggris oleh Mahasiswa Teknik Informatika. *EQUALS: Jurnal Ilmiah Pendidikan Matematika*, *8*(2), 122–134.

Cesaria, A., & Fitri, D. Y. (2014). Efektivitas pengembangan lembar kerja mahasiswa pada perkuliahan Bahasa Inggris untuk Matematika. *Jurnal Pelangi*, *6*(2), 85–91.

Coetzee, J., & Mammen, K. J. (2017). Language of Learning and Teaching as one of the Causes for Difficulties in Learning Fractions. *Science Journal of Education*, *5*(2), 66–70. https://doi.org/10.11648/j.sjedu.20170502.15

DeMars, C. E. (2018). Classical test theory and item response theory. In P. Irwing, T. Booth, & D. J. Hughes (Eds.). In *Classical Test Theory and Item Response Theory* (pp. 49–73). https://doi.org/https://doi.org/10.1002/9781118489772.ch2

Dudley-Evans, T., & St. John, M. J. (1998). *Developments in English for specific purposes: A multi-disciplinary approach*. Cambridge University Press.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational Measurement (5th ed.)*. Prentice-Hall.

Garrison, D. R., & Kanuka, H. (2004). Blended learning: Uncovering its transformative potential in higher education. *Internet and Higher Education*, *7*(2), 95–105. https://doi.org/10.1016/j.iheduc.2004.02.001

Gronlund, N. E., & Waugh, C. K. (2009). *Assessment of Student Achievement (9th ed.)*. Pearson.

Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction*, *28*, 24–34. https://doi.org/https://doi.org/10.1016/j.learninstruc.2013.04.001

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and Validating Test Items*. Routledge.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Springer Science & Business Media.

Hyland, K. (2006). *English for academic purposes: An advanced resource book*. Routledge.

Kurniawati, D. (2023). Teachers' linguistic challenges in teaching mathematics in English: A case study in Indonesian bilingual classrooms. *Indonesian Journal of Language and Education Studies*, *5*(2), 55–68.

Martiniello, M. (2008). Language and the Performance of English-Language Learners in Math Word Problems. *Harvard Educational Review*, *78*(2), 333–368. https://doi.org/https://doi.org/10.17763/haer.78.2.7893w88255765631

Mullen, J. (2009). *Enhancing Mathematical Literacy: How has open access to Fisher Digital Publications benefited you ?*

Nunnally, J. C. (1978). *Psychometric theory (2nd ed.)*. McGraw-Hill.

Prediger, S., Clarkson, P., & Bose, A. (2016). Purposefully Relating Multilingual Registers: Building Theory and Teaching Strategies for Bilingual Learners Based on an Integration of Three Traditions. *International Journal of Science and Mathematics Education*, *14*(1), 143–164. https://doi.org/https://doi.org/10.1007/s10763-016-9720-2

Schleppegrell, M. J. (2007). The Linguistic Challenges of Mathematics Teaching and Learning: A research review. *Reading & Writing Quarterly*, *23*(2), 139–159. https://doi.org/https://doi.org/10.1080/10573560601158461

Setati, M., & Adler, J. (2000). Between Languages and Discourses: Language Practices in Primary Multilingual Mathematics Classrooms in South Africa. *Educational Studies in Mathematics*, *43*, 243–269. https://doi.org/https://doi.org/10.1023/A:1011996002062

Swanson, H. L., Orosco, M. J., & Lussier, C. M. (2015). Journal of Experimental Child Growth in literacy, cognition, and working memory in English language learners. *Journal of Experimental Child Psychology*, *132*, 155–188. https://doi.org/10.1016/j.jecp.2015.01.001

Taifour, A., & Alabdulaziz, M. (2025). Challenges of English mathematical literacy among EFL learners in higher education. *International Journal of Language and Education*, *9*(1), 22–35.