Jurnal Penelitian Politeknik Penerbangan Surabaya
Vol 11, No 2, Agustus, 2026
p-ISSN: 1978-6832
e-ISSN: 2622-5948

Psychometric Quality Of An Elementary School Mathematics Test: A Classical Test Theory Analysis Of Item Difficulty, Discrimination, And Distractor Functioning

Muhammad Lis Haryanto[1], Rufi'i[2], Sabariah[3], Adi[4]
[1,2,3]Universitas PGRI Adi Buana Surabaya
e-mail: [1]harysaprint@gmail.com, [2]rufi,i@unipasby.ac.id, [3]sabariah@unipasby.ac.id, adibandono@unipasby.ac.id
corresponding: [1]harysaprint@gmail.com

DOI : https://doi.org/10.46491/jp.v11i2.2550

Abstract

This study evaluated the psychometric functioning of a 25-item multiple-choice mathematics test on perimeter and area administered to 35 fifth-grade students. Although item analysis is widely used, evidence from routine elementary classroom assessments remains limited, particularly studies that integrate difficulty, discrimination, and option-level response patterns into actionable item-revision decisions. Using a descriptive Classical Test Theory design, all 35 students represented in the archived item summaries were included. Item difficulty (P) was calculated from the proportion of correct responses, discrimination (D) from the difference between the upper and lower approximately 27% groups (n = 10 per group), and distractor functioning from option-selection frequencies. Twenty-three items were easy (P = 0.71-0.91), while only two were moderately difficult (P = 0.69). Discrimination ranged from -0.10 to 0.60; six items showed good or very good discrimination, three had D = 0.00, and two had negative values. Distractor functioning was uneven, especially for items with concentrated response patterns. The findings indicate that the test was suitable for checking basic mastery but insufficiently balanced for differentiating students across a broader ability range. The study contributes a decision-oriented framework for retaining, revising, or reviewing classroom test items. Because the archived dataset did not contain the complete person-by-item matrix, internal-consistency reliability could not be estimated. Larger multi-school samples, repeated administrations, and Item Response Theory analyses are recommended.

Keywords: *classical test theory, item analysis, elementary mathematics, item difficulty*

Article History:
Received: May 10, 2026 ; Revised: July 23, 2026 ; Accepted: August 6, 2026

INTRODUCTION

High-quality classroom assessment depends on test items that represent the intended content and produce interpretable score differences. In elementary mathematics, inaccurate item functioning can distort judgments about foundational knowledge and may lead teachers to provide remediation or enrichment on the basis of misleading evidence. Accordingly, content alignment alone is insufficient; the empirical behavior of items after administration should also be examined (Hrnjičić et al., 2022; Yurtyapan & Çirkinoğlu Şekercioğlu, 2023).

Within Classical Test Theory (CTT), post-administration item analysis commonly examines item difficulty, discrimination, and distractor functioning. The difficulty index (P) is the proportion of examinees who answer an item correctly and is therefore more precisely interpreted as an item-facility indicator: a higher P value denotes an easier item for the observed group. The discrimination

index (D) indicates whether an item differentiates higher- and lower-performing examinees, while option-level frequencies show whether distractors attract plausible misconceptions rather than being ignored. These indicators are most informative when interpreted jointly rather than as isolated statistics (Kim et al., 2023; Rezigalla et al., 2024).

Previous studies consistently show that an examination dominated by very easy items may provide insufficient information about differences among students, particularly at the upper end of the ability distribution. Likewise, non-functioning distractors reduce the effective number of response options, increase the probability of successful guessing, and can weaken item discrimination (Bhat & Prasad, 2021; Kheyami et al., 2018; Rezigalla et al., 2024). Negative discrimination is especially important because it may indicate an incorrect answer key, ambiguous wording, multiple defensible answers, or a mismatch between the item and the intended learning outcome.

Recent mathematics-assessment research has emphasized systematic test development, expert review, reliability estimation, and the construction of larger item banks (Hrnjičić et al., 2022; Mata et al., 2026). By contrast, much of the detailed empirical evidence on distractor efficiency and repeated post-validation has been generated in professional and licensing examinations (Rezigalla et al., 2024; Zubairi et al., 2025). These studies provide valuable psychometric guidance, but routine elementary classroom tests present a distinct applied context because cohorts are smaller, item banks are less formalized, and teachers often need to make immediate revision decisions with limited data.

This study addresses that applied gap by integrating three sources of evidence - difficulty, upper-lower discrimination, and option-level response patterns - into a transparent item-level framework for deciding which questions should be retained, revised, or subjected to urgent review. Its novelty does not lie in proposing a new psychometric model; rather, it lies in translating established CTT indicators into a decision-oriented quality audit for a fifth-grade mathematics test on perimeter and area. This practical contribution is intended to connect blueprint-based test design with empirical post-validation in a resource-constrained classroom setting.

CTT was selected because its indices are transparent and feasible for a small classroom dataset. However, CTT parameters are sample dependent: an item may appear easy or discriminating in one cohort and function differently in another. Item Response Theory (IRT) and Rasch-family models can provide item and person estimates on a common latent scale and richer information about measurement precision, but they generally require larger datasets, explicit model-fit evaluation, and complete person-by-item response records (Volk et al., 2022). CTT is therefore used here as a first diagnostic audit rather than as evidence of parameter invariance.

Accordingly, this study aimed to evaluate the quality of a 25-item fifth-grade mathematics test on perimeter and area. Three questions guided the analysis: (1) how were the items distributed by difficulty level; (2) how effectively did the items discriminate between higher- and lower-performing students; and (3) which distractors and items should be retained, revised, or reviewed on the basis of their empirical response patterns? The study also derives practical recommendations for teachers and assessment developers while explicitly identifying the limits of conclusions that can be drawn from a single small-sample administration.

METHODS

Research design

This study employed a quantitative descriptive, post-administration item-analysis design within the CTT framework. The unit of analysis was the test item, and the purpose was to describe the empirical functioning of the instrument in one observed administration rather than to test causal

relationships or estimate population parameters. Consequently, all results are interpreted as administration-specific evidence.

Instrument quality was evaluated through three complementary indicators: item difficulty (P), item discrimination (D), and distractor functioning. The indicators were integrated to support practical decisions. Items with coherent difficulty, positive discrimination, and plausible option distributions were considered candidates for retention; items that were overly easy, weakly discriminating, or supported by weak distractors were marked for revision; and items with zero or negative discrimination were prioritized for answer-key and content review.

Data sources, participants, and sampling

The analyzed instrument comprised 25 four-option multiple-choice items covering perimeter and area of plane figures. The accompanying blueprint mapped the items to the intended learning indicators and cognitive levels C1-C4. Each item contained one keyed answer and three distractors. The blueprint was used as evidence of intended content coverage, while the empirical analysis examined whether the items functioned as intended after administration.

The analytic cohort consisted of the complete set of 35 fifth-grade response records available from the test administration. A total-coverage approach was used: no probabilistic sampling or sample-size calculation was undertaken, and no available record was excluded. This approach was appropriate for a diagnostic audit of the specific classroom test, but the cohort should not be treated as representative of other classes, schools, or grade levels.

Instrument administration and data preparation

The test was administered under a common classroom protocol. During the original scoring process, each response was coded dichotomously (1 = correct; 0 = incorrect), and total scores were used to rank examinees. For the discrimination analysis, the upper and lower groups were defined as approximately the highest and lowest 27% of the cohort, resulting in 10 students per group; the middle 15 students were not included in the upper-lower calculation. This grouping procedure is commonly used to maximize contrast between performance groups (Kim et al., 2023; Rezigalla et al., 2024).

For the present secondary analysis, the archived analysis file contained the keyed-response counts, option-selection frequencies, and upper- and lower-group summaries required to calculate P and D. It did not retain a complete 35 × 25 person-by-item binary response matrix in a form suitable for reconstructing inter-item covariances. Before analysis, the records were checked for item-number consistency, a total frequency of 35 responses per item, and agreement between keyed-response counts and the reported P values.

The analytical sequence consisted of four steps: calculating P, calculating D, examining the distribution of responses across the keyed answer and distractors, and integrating the three indicators into an item-level decision. Negative D values triggered priority review for possible answer-key error, ambiguity, or content mismatch. Zero D values indicated that an item provided no separation between the upper and lower groups. Low positive values and highly concentrated response patterns were treated as evidence that revision and re-piloting were needed.

Data analysis

The difficulty index was calculated as $P = R/N$, where R is the number of correct responses and N is the total number of examinees (N = 35). Values of 0.31-0.70 were interpreted as moderate and values above 0.70 as easy; no difficult items were observed. Discrimination was calculated as $D = P_U - P_L$, where P_U and P_L are the proportions of correct responses in the upper and lower groups. For descriptive consistency with the source analysis, D values were summarized as very good (≥ 0.50), good (0.40-0.49), low (0.01-0.39), zero (0.00), or negative (< 0.00).

Distractor functioning was evaluated descriptively from option-selection frequencies. A distractor was flagged when it was not selected or attracted only a negligible number of responses, because such options provide little evidence of plausible misunderstanding. Given the small cohort, distractor judgments were not treated as stable population characteristics. The final item classification therefore relied on convergence among P, D, and the response-distribution pattern rather than on any single cutoff.

Validity, reliability, and ethical considerations

Evidence related to content representation was provided by the blueprint, which linked items to learning indicators and cognitive levels. However, no formal expert-judgment coefficient, construct-validity analysis, or criterion-related validity study was available. The findings should therefore be interpreted as evidence about item functioning and content alignment, not as a complete validation of the test or the inferences made from its total scores.

Internal-consistency reliability was not calculated because KR-20 and coefficient alpha require the complete person-by-item score matrix to estimate inter-item covariances. The archived data contained item-level summaries and group counts but not the full matrix; therefore, reconstructing a reliability coefficient would have been statistically invalid. This absence is reported as a data limitation rather than as evidence that the test is either reliable or unreliable. From an ethical perspective, the secondary dataset contained no student names, identification numbers, or demographic details. The analysis involved no experimental manipulation, results were reported only in aggregate, and no attempt was made to identify or profile individual students.

RESULTS AND DISCUSSION

The analysis covered all 35 response records and all 25 items. Because the archived file did not contain the complete person-by-item matrix, participant-level score distributions and internal-consistency reliability are not reported. The results therefore focus on item-level evidence: P values, upper-lower D values, option-selection patterns, and the resulting revision priorities.

Observed P values ranged from 0.69 to 0.91. This high-facility pattern indicates that most items were answered correctly by a large proportion of the cohort. The result may reflect strong student mastery, an insufficiently challenging item set, or both; item analysis from a single administration cannot separate these explanations. What can be concluded is that the test provided limited variation in challenge for this group.

The item-difficulty distribution was markedly unbalanced. Only items 1 and 11 were in the moderate category (P = 0.69), whereas the remaining 23 items were easy (P = 0.71-0.91). Item 5 had the highest facility value (P = 0.91). Thus, 92% of the test consisted of easy items and 8% consisted of moderately difficult items.

The absence of difficult items means that the instrument offered little opportunity to distinguish students whose mastery exceeded the level targeted by the easier questions. Although this distribution may be acceptable for a minimum-mastery check, it is less appropriate when the assessment is intended to represent a broad range of cognitive demand or differentiate performance across C1-C4.

Table 1. Distribution of Item Difficulty

Category	Observed P value(s)	Number of items	Item numbers
Moderate	0.69	2	1, 11
Easy	0.71-0.91	23	2-25 (except 11)

Note: P is the proportion of students who answered an item correctly; higher values indicate easier items for this cohort.

Item discrimination showed substantial variation, ranging from -0.10 to 0.60. Six items had good or very good discrimination, indicating that they were answered correctly more often by the upper group. Fourteen items had low positive discrimination, three had D = 0.00, and two had negative values. Because these are descriptive indices from a small cohort, the classifications should be treated as diagnostic signals rather than fixed item properties.

Items 11 and 21 showed the highest discrimination (D = 0.60), followed by item 18 (D = 0.50). Items 10, 16, and 20 each had D = 0.40. In contrast, items 13 and 19 had D = -0.10, while items 4, 23, and 24 had D = 0.00. The negative values warrant immediate verification of the keyed answers and item wording, whereas the zero values indicate no observed separation between the upper and lower groups.

Table 2. Distribution of Item Discrimination

Category	Range of D	Number of items	Item numbers
Very good	≥ 0.50	3	11, 18, 21
Good	0.40-0.49	3	10, 16, 20
Low positive	0.01-0.39	14	1, 2, 3, 5, 6, 7, 8, 9, 12, 14, 15, 17, 22, 25
Zero	0.00	3	4, 23, 24
Negative	< 0.00	2	13, 19

Note: D is the difference between the proportions correct in the upper and lower groups; classifications are descriptive and administration specific.

Analysis of distractor functioning

Option-selection frequencies showed that distractor functioning was uneven. Item 11, for example, was answered correctly by 24 students, while the 11 incorrect responses were distributed across the distractors. This pattern suggests that the incorrect options were sufficiently plausible to attract some examinees and is consistent with the item's strong discrimination (D = 0.60).

Item 5 showed the opposite pattern: 32 students selected the keyed answer, only three responses were distributed among the distractors, and one distractor was not selected. Together with P = 0.91, this pattern indicates that the item offered little challenge and that at least one distractor was non-functioning. Such distractors should be replaced with alternatives derived from common computational or conceptual errors.

Item-level decision classification

The integrated classification identified items 1, 10, 11, 12, 16, 18, 20, and 21 as the strongest candidates for retention with routine review. These items showed the most favorable combination of difficulty, discrimination, and response distribution. Items 13 and 19 require urgent review because negative discrimination may indicate miskeying, ambiguity, or a mismatch with the intended learning indicator.

Items 4, 23, and 24 should be revised or replaced because D = 0.00. Items 2, 3, 5, 6, 7, 8, 9, 14, 15, 17, 22, and 25 should be revised and re-piloted because they were overly easy, weakly discriminating, supported by weak distractors, or affected by a combination of these conditions. This complete classification accounts for all 25 items.

Discussion

From a CTT perspective, the dominance of easy items narrows observed score variation and creates a ceiling-prone instrument. This is not inherently problematic when the sole purpose is to verify minimum mastery; however, it is misaligned with an assessment intended to differentiate students across a broad range of performance or cognitive demand. Because P is jointly influenced by item characteristics and cohort ability, the findings do not prove that the mathematical content

itself was easy. They show that this particular item set provided limited information about higher levels of performance in this cohort (Hrnjičić et al., 2022; Mata et al., 2026).

The discrimination results provide a more diagnostic interpretation than a simple good-versus-bad ranking. A positive D value indicates that the keyed answer was more common among higher-performing students. A zero value indicates no observed group separation, while a negative value suggests that lower-performing students answered correctly more often than higher-performing students. Negative discrimination may arise from an incorrect key, ambiguous language, multiple defensible answers, excessive reading demands, or content that was not taught or interpreted as intended. Items 13 and 19 should therefore undergo qualitative review before any decision to delete them, whereas items 11, 18, and 21 can be used as internal models for clearer item construction.

Distractor functioning helps explain why some items were both easy and weakly discriminating. Plausible distractors preserve the diagnostic value of a multiple-choice item by attracting examinees who hold predictable misconceptions. In contrast, an option that no student selects reduces the effective number of alternatives and makes successful guessing easier. The contrast between item 11 and item 5 illustrates this mechanism: the former distributed incorrect responses across alternatives and discriminated strongly, whereas the latter concentrated responses on the keyed answer and contained an unused distractor. This interpretation is consistent with evidence that distractor efficiency is related to both item facility and discrimination (Bhat & Prasad, 2021; Kheyami et al., 2018; Rezigalla et al., 2024).

The study's theoretical contribution is deliberately modest but distinct. It demonstrates that blueprint coverage does not guarantee empirical functioning and that difficulty, discrimination, and distractor behavior should be interpreted as a connected system. In a small classroom assessment, revision decisions should be based on converging evidence rather than a single threshold. This decision-oriented use of CTT complements larger-scale test-development research and operationalizes post-validation for elementary mathematics practice (Yahia, 2022; Zubairi et al., 2025).

For teachers, the findings support a cyclical quality-assurance process: develop a blueprint, conduct peer review, administer or pilot the items, analyze item performance, revise weak items, and re-administer the improved form. For the present test, the eight strongest items may be retained with routine monitoring; negative- and zero-discrimination items should be checked against the answer key and learning indicators; and easy, low-discrimination items should be rewritten using more plausible distractors and a broader range of cognitive demand. Distractors should be based on common student errors in perimeter and area, such as confusing formulas, omitting units, or applying the correct formula to the wrong dimension. Sustained item-writing training and peer moderation can improve discrimination and reduce non-functioning distractors (Shaikh et al., 2020).

Table 3. Illustrative Response Distributions and Distractor Functioning

Item no.	Observed response pattern	Interpretation
11	24 correct; 11 responses distributed across distractors	Relatively functional distractors
5	32 correct; three incorrect responses; one distractor unselected	At least one non-functioning distractor

Note: Option distributions are illustrative examples; distractor judgments were interpreted descriptively because of the small cohort.

Practical implications, limitations, and future research

For assessment developers, data management is as important as item writing. Future administrations should retain the complete person-by-item response matrix, the answer-key version, cognitive-level labels, option frequencies, administration conditions, and item-revision history.

These records would permit calculation of KR-20 or coefficient alpha, point-biserial correlations, uncertainty estimates, longitudinal monitoring, and direct analysis of whether C1-C4 classifications correspond to empirical item behavior (Lahza et al., 2022; Zubairi et al., 2025).

Several limitations constrain the interpretation of the findings. First, the cohort was small (N = 35) and came from a single test administration, so CTT indices may be unstable and should not be generalized beyond this context. Second, the absence of the complete response matrix prevented internal-consistency analysis and participant-level score analysis. Third, content evidence was limited to the blueprint; no formal expert-validity coefficient, construct-validity analysis, criterion comparison, or cognitive-interview evidence was available. Fourth, distractor functioning was interpreted descriptively, and no confidence intervals or resampling procedures were applied.

Future research should replicate the analysis with larger, multi-school datasets and repeated administrations. With an adequate sample and complete response data, IRT or Rasch modeling could estimate item difficulty and discrimination on a latent scale, produce item-information functions and person-item maps, evaluate model fit, and investigate differential item functioning across relevant student groups. A comparison of CTT and IRT results would show whether the identified revision priorities are robust across psychometric frameworks (Volk et al., 2022). Qualitative cognitive interviews should also be used to determine why students select particular distractors, especially for items with negative discrimination.

Accordingly, the present findings should be treated as a diagnostic baseline for item revision rather than as final validation of the instrument. The value of the analysis lies in identifying specific, testable improvements that can be evaluated in the next administration.

Table 4. Item Classification and Revision Priority

Revision priority	Item numbers
Retain with routine review	1, 10, 11, 12, 16, 18, 20, 21
Urgent key/content review (negative D)	13, 19
Revise or replace (D = 0.00)	4, 23, 24
Revise and re-pilot	2, 3, 5, 6, 7, 8, 9, 14, 15, 17, 22, 25

Note: The classification integrates P, D, and option-distribution patterns and accounts for all 25 items.

CONCLUSION

This study found that the 25-item elementary mathematics test had an unbalanced difficulty distribution and uneven item functioning. Twenty-three items were easy, only two were moderately difficult, discrimination ranged from -0.10 to 0.60, three items had zero discrimination, and two had negative discrimination. The distractor analysis further showed that several options attracted few or no responses.

The results indicate that the instrument may be useful for checking basic mastery in this cohort, but it is not yet sufficiently balanced to differentiate students across a broader range of performance. The practical contribution of the study is an item-level revision framework: retain the strongest items with routine monitoring, verify negative-discrimination items for key or wording problems, revise zero-discrimination items, and reconstruct weak distractors from common mathematical misconceptions.

The conclusions are limited by the small cohort, the use of a single administration, and the absence of a complete person-by-item response matrix. Consequently, internal-consistency reliability, participant-level score distributions, and more comprehensive validity evidence could not be evaluated. The reported P and D values should therefore be regarded as sample-dependent diagnostic evidence rather than invariant psychometric parameters.

Future studies should use larger multi-school samples, preserve complete response matrices, repeat the analysis across administrations, and apply advanced approaches such as IRT or Rasch modeling. Such work would allow more stable calibration, item-information analysis, person-item mapping, model-fit evaluation, and differential item functioning. Until those data are available, systematic CTT-based post-validation remains a practical basis for improving the clarity, balance, and instructional usefulness of classroom mathematics assessments.

REFERENCES

Bhat, S. K., & Prasad, K. H. (2021). Item analysis and optimizing multiple-choice questions for a viable question bank in ophthalmology. *Indian Journal of Ophthalmology, 69*(2), 343-346. https://doi.org/10.4103/ijo.IJO_1610_20

Hrnjičić, A., Alihodžić, A., Čunjalo, F., & Kamber Hamzić, D. (2022). Development of an item bank for measuring students' conceptual understanding of real functions. *European Journal of Science and Mathematics Education, 10*(4), 455-470. https://doi.org/10.30935/scimath/12222

Kheyami, D., Jaradat, A., Al-Shibani, T., & Ali, F. A. (2018). Item analysis of multiple-choice questions at the Department of Paediatrics, Arabian Gulf University, Manama, Bahrain. *Sultan Qaboos University Medical Journal, 18*(1), e68-e74. https://doi.org/10.18295/squmj.2018.18.01.011

Kim, Y. H., Kim, B. H., Kim, J., Jung, B., & Bae, S. (2023). Item difficulty index, discrimination index, and reliability of the 26 health professions licensing examinations in 2022, Korea: A psychometric study. *Journal of Educational Evaluation for Health Professions, 20*, 31. https://doi.org/10.3352/jeehp.2023.20.31

Lahza, H., Smith, T. G., & Khosravi, H. (2022). Beyond item analysis: Connecting student behaviour and performance using e-assessment logs. *British Journal of Educational Technology, 54*(1), 335-354. https://doi.org/10.1111/bjet.13270

Mata, P. V., Abao, Q. R. H., Bihasa, R. K. S., Kiamco, G. N., & Neiz, A. M. C. (2026). Development and validation of curriculum-based problem-solving MCQ classroom tests in mathematics. *International Journal of Evaluation and Research in Education, 15*(1), 286-302. https://doi.org/10.11591/ijere.v15i1.35698

Rezigalla, A. A., Eleragi, A. M. S., Elhussein, A. B., Alfaifi, J., Alghamdi, M. A., Al Ameer, A. Y., Yahia, A. I. O., Mohammed, O. A., & Adam, M. I. E. (2024). Item analysis: The impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education, 24*, 445. https://doi.org/10.1186/s12909-024-05433-y

Shaikh, S., Kannan, S. K., Naqvi, Z. A., Pasha, Z., & Ahamad, M. (2020). The role of faculty development in improving the quality of multiple-choice questions in dental education. *Journal of Dental Education, 84*(3), 316-322. https://doi.org/10.21815/JDE.019.189

Volk, C., Rosenstiel, S., Demetriou, Y., Sudeck, G., Thiel, A., Wagner, W., & Höner, O. (2022). Health-related fitness knowledge in adolescence: Evaluation of a new test considering different psychometric approaches (CTT and IRT). *German Journal of Exercise and Sport Research, 52*(1), 11-23. https://doi.org/10.1007/s12662-021-00735-5

Yahia, A. I. O. (2022). Post-validation item analysis to assess the validity and reliability of multiple-choice questions at a medical college with an innovative curriculum. *The National Medical Journal of India, 34*, 359-362. https://doi.org/10.25259/NMJI_414_20

Yurtyapan, E., & Çirkinoğlu Şekercioğlu, A. G. (2023). Development of an achievement test for the "Seasons and Climate" unit of the eighth-grade science course in secondary school. *International Journal of Contemporary Educational Research, 10*(4), 893-916. https://doi.org/10.52380/ijcer.2023.10.4.501

Zubairi, N. A., AlAhmadi, T. S., Ibrahim, M. H., Hegazi, M. A., & Gadi, F. U. (2025). Effective use of item analysis to improve the reliability and validity of undergraduate medical examinations: Evaluating the same exam over many years. *Pakistan Journal of Medical Sciences, 41*(3), 810-815. https://doi.org/10.12669/pjms.41.3.10693