Jurnal Penelitian Politeknik Penerbangan Surabaya
Vol 11, No 1, April, 2026
p-ISSN: 1978-6832
e-ISSN: 2622-5948

BEYOND HIGH SCORES: EVALUATING ITEM DIFFICULTY AND MEASUREMENT SENSITIVITY IN BASIC NUMERACY ASSESSMENT

Laela Fauziah[1], Rufi'i[2], Sabariah[3]
[1,2,3]Universitas PGRI Adi Buana Surabaya
corresponding: [3]sabariah@unipasby.ac.id

Abstract

Assessment of basic numeracy is essential for identifying students' foundational mathematical competence and for supporting instructional decisions. However, very high test scores may conceal limited measurement sensitivity when an instrument contains many low-difficulty items. This study examined an existing 25-item basic numeracy assessment by interpreting highly concentrated score distributions and proportion-correct values (p-values) as preliminary indicators of item difficulty. A descriptive quantitative approach was applied to data from 35 participants. Because the available dataset did not support formal reliability estimation, item discrimination indices, distractor analysis, or IRT/CDM modelling, the analysis was limited to total-score distribution, item-level proportion-correct values, and documentary review of item explanations. Results showed strong upper-end score concentration: scores ranged from 19 to 25 (M = 23.6, SD = 1.83), and 51.4% of participants achieved the maximum score. Items 11, 18, and 19 were answered correctly by all participants, whereas Items 24 and 10 produced the lowest correct-response rates but remained easy overall (85.7% and 88.6%). Item 17 was flagged for possible ambiguity in the answer explanation. The findings indicate that the instrument functions adequately as a baseline mastery check but has limited capacity to differentiate higher-performing students. The study contributes practical evidence for revising classroom-based numeracy instruments through more balanced item difficulty, clearer keying logic, and future validation using larger samples and stronger psychometric evidence.

Keywords: basic numeracy assessment; score distribution; item difficulty; proportion-correct p-value; ceiling effect; educational measurement; item analysis

Article History:
Received: 11 december 2025 ; Reviewed : 11 january 2026; Accepted: 11 march 2026; Published : 01 april 2026

How to cite : Fauziah, L., Rufi'i, R., & Sabariah, S. (2026). Beyond High Scores: Evaluating Item Difficulty and Measurement Sensitivity in Basic Numeracy Assessment. ***Jurnal Penelitian*****, *11*(1), 51–62. https://doi.org/10.46491/jp.v20i1.2547**

© 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution Share A like (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/)

INTRODUCTION

Basic numeracy is a foundational competency that supports students' mathematical development, problem-solving, and participation in everyday quantitative situations. In educational practice, numeracy assessment should therefore provide more than a

summary score; it should generate interpretable evidence about whether students have mastered essential concepts and whether further instructional support is required. Prior studies on early numeracy and standardized assessment show that performance may be influenced not only by mathematical knowledge but also by item format, working memory, language demands, and the match between task characteristics and students' developmental level (Howard et al., 2016; Woodcock et al., 2020).

The quality of a numeracy assessment depends on whether its items are aligned with the intended construct and whether the resulting scores can support the decisions for which the test is used. In classroom settings, teachers often use short multiple-choice instruments to check foundational learning. Such instruments can be valuable, but their usefulness depends on the extent to which they provide evidence that is both interpretable and appropriate for the intended purpose. Recent validity frameworks for classroom assessment emphasize that teachers and assessment users need practical evidence for judging the accuracy of proficiency claims and the appropriateness of subsequent instructional uses (Mor & Karatoprak Erşen, 2023; McMillan, 2026).

One recurring problem in educational assessment is the ceiling effect, which occurs when a large proportion of students obtain scores near the maximum possible value. A ceiling effect may indicate that students have achieved the intended baseline competence; however, it may also indicate that the instrument is too easy for the tested population. When this occurs, the assessment provides limited information about differences among higher-performing students, and the total score may appear more precise than the evidence actually permits (Rijmen et al., 2014; Woodcock et al., 2020; Le & Nguyen, 2024). Item-level evidence is therefore essential for interpreting high score distributions responsibly. The proportion-correct index, often called the p-value in classical item analysis, provides an accessible indicator of item difficulty by showing the proportion of examinees who answer an item correctly. Nevertheless, this index is not sufficient on its own. Effective interpretation also requires considering discrimination, distractor quality, item wording, and alignment with the measurement purpose. Studies on item analysis and multiple-choice assessment have shown that difficulty indices, discrimination indices, and distractor functioning are interconnected indicators of item quality (Feskens et al., 2019; Rezigalla et al., 2024).

Advanced psychometric approaches, including multidimensional Item Response Theory (IRT), Cognitive Diagnostic Models (CDMs), and process-oriented assessment, can provide more detailed information about students' mastery profiles and item functioning (Wu et al., 2020; Wu et al., 2025). However, these approaches require sufficiently large samples, complete item-response matrices, and richer contextual information. In many classroom-based assessments, such information is not available. Under these conditions, descriptive evaluation remains a useful first step for identifying

whether an existing instrument is functioning as intended and for determining which items require revision before more rigorous validation is attempted.

The research gap addressed in this study concerns existing classroom-based numeracy instruments that are already used in practice but have limited technical documentation. Much previous work has focused on new instrument development, large-scale assessments, or sophisticated psychometric modelling. Comparatively less attention has been given to small-scale classroom instruments that show high scores but lack sufficient evidence about item difficulty, discrimination, and measurement sensitivity. This gap is important because educators may interpret high scores as evidence of uniformly strong achievement without considering whether the instrument itself is sufficiently challenging or informative.

The novelty of this study lies in its focused descriptive evaluation of an existing 25-item basic numeracy assessment under constrained data conditions. Rather than proposing a new assessment model or making unsupported psychometric claims, the study integrates total-score distribution, proportion-correct item difficulty, and documentary review of item explanations to clarify what can and cannot be inferred from a high-scoring dataset. This orientation provides practical value for teachers, assessment developers, and researchers who need to improve classroom instruments when complete validation evidence is not yet available.

Accordingly, this study aims to (1) describe students' performance on a 25-item basic numeracy assessment, (2) examine the overall difficulty pattern and preliminary quality of the assessment items, (3) evaluate the instrument's ability to differentiate student achievement, and (4) identify items requiring revision because of excessive ease or potential ambiguity. The findings are expected to support more cautious score interpretation and to provide evidence-based recommendations for revising the instrument before future validation with larger samples and more complete psychometric data.

METHODOLOGY

1. Research Design

This study employed a descriptive quantitative design to examine the results of a 25-item basic numeracy test and to provide a preliminary evaluation of the assessment instrument. The design was appropriate because the purpose of the study was not to test causal effects, compare interventions, or estimate latent traits, but to describe score patterns, inspect item difficulty, and identify potential measurement issues within an existing dataset.

The analysis was intentionally limited to evidence supported by the available data. Formal psychometric procedures such as reliability estimation, item discrimination indices, item-total correlations, distractor analysis, Differential Item Functioning (DIF),

IRT, CDM, and process-data modelling were not conducted because the dataset did not contain the necessary information. This methodological boundary was maintained to avoid overstating conclusions beyond the available evidence.

2. Data Sources and Research Objects

The study used two data sources: (1) a basic numeracy assessment instrument consisting of a test blueprint, 25 multiple-choice items, and answer explanations, and (2) the test results of 35 participants. The instrument was reviewed to examine content coverage and to identify items that appeared overly easy, potentially ambiguous, or problematic in their answer explanations. Participant results were used to analyze the distribution of total scores and the proportion of correct responses for selected items. Because demographic information was unavailable, subgroup analyses were not conducted.

3. Data Collection Procedure

The original test administration procedure was not fully documented. Information on the testing date, location, duration, supervision, and participant instructions was unavailable. Therefore, this study treated the dataset as a completed assessment record and conducted a document-based secondary analysis using the instrument and score records. The absence of administration details is acknowledged as a limitation because testing conditions may affect score reliability and interpretation. Consequently, the study does not draw conclusions regarding administration mode, test security, or contextual influences on performance.

4. Data Analysis Techniques

Data were analyzed using descriptive quantitative techniques. Total-score distribution was examined using the score range, mean, standard deviation, frequency, and percentage to describe participant performance and identify possible ceiling effects. Item difficulty was evaluated using the proportion of correct responses, interpreted as a preliminary p-value or proportion-correct index. In addition, item wording, answer keys, and answer explanations were reviewed to identify possible inconsistencies. Because formal item-response data were not available in sufficient detail, the analysis did not estimate reliability coefficients, item-total correlations, discrimination indices, distractor functioning, DIF, IRT parameters, or CDM profiles.

5. Validity, Reliability, and Ethical Considerations

No formal reliability evidence was available, and the dataset did not allow robust analysis of internal consistency. Therefore, this study makes no claims regarding reliability. The score distribution and item-level indicators are interpreted only as preliminary evidence of instrument performance. Although the test blueprint appears to align with basic numeracy content, the absence of evidence for construct validity, response processes, and score-use justification restricts the study to a descriptive evaluation. Participant confidentiality was protected through aggregated reporting, and

all interpretations were limited to the evidence available in the dataset (Lin et al., 2014; Mor & Karatoprak Erşen, 2023).

RESULTS

1. General Profile of Student Scores

The results of the 25-item basic numeracy test indicate generally high participant performance. Among the 35 participants, scores ranged from 19 to 25, with a mean of 23.6 (SD = 1.83). More than half of the participants (51.4%) achieved the maximum score of 25, while all remaining scores were concentrated between 19 and 24. This pattern shows a clear ceiling tendency and indicates limited variability in participants' performance.

Table 1. Distribution of Total Scores

Total Score	Frequency	Percentage
25	18	51.4%
24	4	11.4%
23	4	11.4%
22	4	11.4%
21	1	2.9%
20	3	8.6%
19	1	2.9%
Total	35	100%

Table 1 presents the frequency distribution of total scores. The accumulation of participants at the maximum score, together with the narrow score range, indicates score compression at the upper end of the distribution. Descriptively, the test successfully identified a generally high level of basic numeracy performance across the participant group. However, the same pattern also shows that the test generated limited separation among participants and should not be interpreted as providing fine-grained ability estimates.

2. Test Reliability

No formal reliability evidence was available because the dataset did not include indices such as Cronbach's alpha, KR-20, or split-half reliability. Therefore, the study cannot make statistical claims about the instrument's reliability. The available evidence only indicates that scores were highly concentrated at the upper end of the scale. This concentration may reflect similar participant mastery, limited item challenge, or both. Accordingly, reliability-related interpretations remain cautious.

3. Item Difficulty

Item difficulty was evaluated descriptively using the proportion of correct responses. Overall, the test was very easy for the sampled participants. Items 11, 18, and 19 were answered correctly by all participants (100%), making them the easiest items. In contrast, Item 24 (85.7%) and Item 10 (88.6%) had the lowest correct-response rates and were

therefore the most difficult items relative to the rest of the test. However, because more than 80% of participants answered these items correctly, they were still easy in absolute terms.

Table 2. Selected Item Difficulty Indicators

Item Number	Proportion Correct	Descriptive Category
11	100.0%	Very easy
18	100.0%	Very easy
19	100.0%	Very easy
10	88.6%	Relatively more difficult
24	85.7%	Most difficult in test

Table 2 shows that even the lowest proportion-correct values remained high. The item difficulty profile was therefore strongly skewed toward ease. This pattern suggests that the item set was better suited to confirming baseline numeracy mastery than to differentiating students across a wider range of ability.

4. Item Discrimination

Formal item discrimination indices were not available, so the study could not evaluate statistically how well each item distinguished between higher- and lower-performing participants. Descriptively, however, the narrow score distribution and uniformly high correct-response rates suggest limited discriminatory capacity. Items 11, 18, and 19 provided no observable differentiation because all participants answered them correctly, whereas Items 10 and 24 generated relatively more response variation but remained easy overall.

5. Item-Total Correlation

No item-total correlation data were available. Therefore, the study could not assess how well individual items aligned with the overall test score or identify items that behaved inconsistently with the intended construct. This section is limited to acknowledging the absence of item-total correlation evidence without drawing additional conclusions.

6. Distractor Quality

No distractor-function data were available because the dataset did not include response frequencies for each incorrect option. Consequently, the study could not determine whether the distractors functioned appropriately or whether some options were non-functional. This limitation is important because distractor quality can influence both item difficulty and discrimination in multiple-choice assessment (Rezigalla et al., 2024).

7. Identification of Relatively Strong and Problematic Items

Although formal discrimination, item-total correlation, and distractor statistics were unavailable, the available evidence still permitted descriptive identification of items that appeared relatively more informative and items that required review. This identification

was based on two sources of evidence: response variation and documentary consistency in the answer explanations.

Items 24 and 10 were the relatively strongest items in the current dataset because they produced the lowest correct-response rates, at 85.7% and 88.6%, respectively. Although both items were still easy in absolute terms, they generated more response variation than the rest of the item set. By contrast, Items 11, 18, and 19 were the weakest in terms of observable differentiation because all participants answered them correctly.

A separate issue emerged in relation to Item 17. This item was flagged as potentially problematic not primarily because of its proportion-correct value, but because its answer explanation appeared to create ambiguity in the scoring logic. The documentation suggested that more than one option could be interpreted as correct, or at least that the explanation did not clearly preserve a single best answer. This ambiguity should be addressed before the item is used in future administrations.

Table 3. Descriptive Identification of Stronger and Problematic Items

Item Number	Basis for Identification	Result Category
24	Lowest proportion correct (85.7%)	Relatively strongest descriptively
10	Second-lowest proportion correct (88.6%)	Relatively strongest descriptively
11	100.0% correct	Very easy / non-differentiating
18	100.0% correct	Very easy / non-differentiating
19	100.0% correct	Very easy / non-differentiating
17	Ambiguity in answer explanation	Potentially problematic

Table 3 classifies the items that stood out most clearly in the available evidence. The classification is descriptive rather than psychometric because it is based on observed correctness rates and document review rather than formal parameter estimation. The results indicate a test characterized by high participant performance, low overall item difficulty, limited observable variation, and at least one item that requires substantive review because of ambiguity in the answer explanation.

DISCUSSION

1. Interpreting High Scores as Evidence of Mastery and Limited Sensitivity

The main finding of this study is that the basic numeracy instrument produced a highly concentrated score distribution at the upper end of the scale. More than half of the participants achieved a perfect score, the mean score was 23.6 out of 25, and the lowest score remained relatively high. At the item level, all selected items showed high proportions of correct responses, with three items answered correctly by all participants and the lowest-performing item still answered correctly by 85.7% of the sample. These findings indicate that the instrument was effective in confirming broad mastery of basic numeracy content but less effective in differentiating levels of competence within a high-performing group.

This distinction between mastery confirmation and measurement sensitivity is central to the contribution of the study. In educational measurement, a test can be useful for confirming whether students have achieved a minimum standard while still being weak for diagnosing finer differences among students. IRT-based and large-scale assessment literature shows that when item difficulty is not well matched to the ability range of the tested population, score distributions may become compressed and measurement information may decline in the upper range (Rijmen et al., 2014; Feskens et al., 2019; Opesemowo et al., 2026). The present data fit this pattern: the test appears well suited to a baseline mastery purpose, but insufficiently calibrated for detailed differentiation.

The interpretation should remain cautious because the available dataset did not permit formal parameter estimation. It is possible that the participants genuinely possessed strong numeracy competence. However, the combination of high total scores, very high item p-values, and several non-differentiating items suggests that item ease contributed substantially to the observed ceiling tendency. Thus, the most defensible interpretation is not simply that students performed well, but that the instrument provided limited evidence beyond broad attainment of foundational skills.

2. Educational Implications of Ceiling Effects and Limited Discrimination

The ceiling effect has important implications for educational assessment practice. If a test is used only as a screening tool to confirm whether students possess minimum foundational numeracy competence, the current instrument may still be useful. However, if the same test is used for ranking students, identifying advanced learning needs, comparing relative ability, or making detailed instructional decisions, the interpretation becomes problematic. A score of 25 and a score of 23 may not represent a meaningful difference in ability when the item set lacks sufficient challenge and discrimination.

In classroom practice, this issue can affect instructional decision-making. Teachers may conclude that most students have uniformly strong numeracy skills and may therefore reduce opportunities for enrichment or advanced problem-solving. At the same time, students who require extension tasks may not be identified because the test does not provide enough upper-level items to reveal their relative strengths. Conversely, students who obtained slightly lower scores may be interpreted as weaker even though the test provides limited evidence for fine-grained differentiation.

The findings therefore support a more careful approach to score reporting. Scores from this instrument should be interpreted as evidence of broad foundational attainment rather than as precise indicators of individual ability differences. This recommendation aligns with classroom assessment validity literature, which emphasizes that score interpretation should be linked to the intended use of the assessment and to the strength of supporting evidence (Mor & Karatoprak Erşen, 2023; McMillan, 2026).

3. Item-Level Implications for Instrument Revision

The item-level findings provide practical guidance for revising the instrument. Items 11, 18, and 19 answered correctly by all participants should be reviewed first because they

contributed no observable differentiation. These items may be retained only if the purpose is to confirm very basic competence. If the instrument is intended to support diagnostic or comparative interpretation, these items should be revised to increase cognitive demand or replaced with items requiring more complex reasoning, representation, or multi-step calculation.

Items 10 and 24 should be examined as useful models for future item development because they produced the greatest response variation in the current dataset. Although they were still easy in absolute terms, their relative difficulty suggests that their content, wording, or cognitive demands may be more informative than those of other items. Future revisions should build a more balanced item pool that includes low-, moderate-, and higher-difficulty items so that the test remains accessible while also improving measurement sensitivity across the ability spectrum.

Item 17 requires a different type of revision. The concern is not merely statistical but qualitative: the answer explanation appears to weaken the clarity of the keying logic. A multiple-choice item should have a single best answer supported by a clear rationale. If more than one option can be defended, the item may compromise fairness because students may be penalized for selecting an answer that is conceptually reasonable. Therefore, Item 17 should be rewritten, reviewed by subject-matter experts, and piloted before it is used again.

The absence of distractor data also limits interpretation. Future administrations should record option-level responses so that distractor functioning can be analyzed. This is important because non-functional distractors can make items artificially easy and reduce item discrimination. Prior item-analysis research shows that distractor efficiency is related to item difficulty and discrimination, making it a relevant component of multiple-choice instrument refinement (Rezigalla et al., 2024).

4. Practical Contribution of the Study

The practical contribution of this study is that it demonstrates how limited descriptive data can still be used responsibly to improve an assessment instrument. In many classroom contexts, teachers and assessment developers do not have large samples or complete item-response matrices. Nevertheless, they often need to decide whether an existing test is appropriate for future use. The present study shows that even simple evidence—score distribution, proportion-correct values, and item documentation—can identify important problems such as ceiling effects, excessive item ease, and ambiguity in answer explanations.

The findings also clarify the distinction between high achievement and high measurement quality. High scores are desirable from a learning perspective, but they do not automatically indicate that an instrument functions well. A test may show that students generally know the material while still failing to provide enough information for diagnostic, comparative, or developmental purposes. This distinction is important for educators because it prevents overinterpretation of total scores and encourages evidence-

based instrument revision.

For assessment developers, the study provides a revision pathway: retain items that confirm essential baseline competence, revise or replace items that are universally solved, add items with higher cognitive demand, clarify ambiguous scoring rationales, and collect richer response data in the next administration. This pathway strengthens the practical value of the manuscript by translating descriptive findings into actionable recommendations for improving classroom-based numeracy assessment.

5. Recommendations for Future Validation

Future validation should proceed in stages. First, the instrument should undergo expert review to ensure that all items align with the test blueprint, represent the intended basic numeracy construct, and have unambiguous answer keys. Second, a revised item pool should include a wider range of difficulty levels, especially moderate and upper-difficulty items, to reduce ceiling effects and improve discrimination. Third, the instrument should be administered to a larger and more diverse sample so that reliability, item-total correlations, discrimination indices, distractor functioning, and subgroup patterns can be evaluated.

At a later stage, IRT or CDM analysis may be appropriate if larger samples and complete item-level responses become available. Such analyses would allow researchers to estimate item parameters, examine whether the instrument provides information across the relevant ability range, and explore whether students with similar total scores have different mastery profiles (Wu et al., 2020; Wu et al., 2025). However, these procedures should be treated as future validation steps rather than as conclusions supported by the current dataset.

The recommendations above are consistent with the methodological boundary of the present study. The study does not claim that the instrument is invalid, unreliable, or unsuitable for all purposes. Rather, it clarifies that the current evidence supports use as a baseline mastery check but not as a fine-grained diagnostic or ranking instrument. This distinction allows the instrument to be improved without overstating what the current data can support.

6. Limitations of the Study

Several limitations should be acknowledged. First, the sample included only 35 participants, and no demographic, contextual, or longitudinal information was available. As a result, the findings cannot be generalized beyond the immediate dataset and cannot be used to examine subgroup differences or learning changes over time.

Second, the dataset did not include formal psychometric indicators such as reliability coefficients, item-total correlations, discrimination indices, distractor distributions, DIF evidence, or IRT/CDM parameters. The study was therefore restricted to descriptive interpretation. Third, the administration context was not documented in sufficient detail, which means that potential influences of testing conditions could not be evaluated. These limitations do not undermine the descriptive value of the findings, but they restrict the

strength of claims that can be made about the instrument.

CONCLUSION

This study examined highly concentrated score distributions and the use of proportion-correct values as preliminary indicators of item difficulty in a 25-item basic numeracy assessment. The findings show that participants performed very highly, with scores ranging from 19 to 25 and more than half of the participants achieving the maximum score. At the item level, several items were answered correctly by all participants, while the relatively most difficult items remained easy in absolute terms.

The findings indicate that the instrument is useful for confirming baseline mastery of basic numeracy skills but has limited sensitivity for differentiating among higher-performing students. Therefore, high scores should not be interpreted as evidence that the instrument provides precise or fine-grained measurement. Instead, the score pattern should be understood as evidence of broad attainment within a limited range of item challenge.

The study contributes to educational assessment practice by showing how descriptive evidence can support responsible interpretation and practical instrument refinement when complete psychometric data are unavailable. Future revision should prioritize clearer scoring logic, revision of universally solved items, inclusion of more moderately and highly challenging items, and collection of richer item-level data. Subsequent validation with larger samples should examine reliability, discrimination, distractor functioning, item-total behavior, and, where appropriate, IRT or CDM-based evidence. Overall, the study underscores that high test scores alone are not sufficient indicators of measurement quality; item structure and score-use evidence must also support meaningful and fair interpretation.

AI Declaration

The author acknowledges the use of artificial intelligence (AI) tools in supporting the preparation of this manuscript. The author remains fully responsible for the accuracy, originality, and integrity of the research and its findings.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

No funding was received for this work.

REFERENCES

Feskens, R., Fox, J., & Zwitser, R. (2019). Differential item functioning in PISA due to mode effects. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 231–247).

Springer. https://doi.org/10.1007/978-3-030-18480-3_12

Howard, S. J., Woodcock, S., Ehrich, J., & Bokosmaty, S. (2016). What are standardized literacy and numeracy tests testing? Evidence of the domain-general contributions to students' standardized educational test performance. *British Journal of Educational Psychology*, 87(1), 108–122. https://doi.org/10.1111/bjep.12138

Le, H. V., & Nguyen, L. Q. (2024). A comparative study of critical reading abilities among students in Malaysia and Vietnam: Insights from PISA-based assessment. *Research in Comparative and International Education*, 19(2), 153–174. https://doi.org/10.1177/17454999241242994

Lin, M., Bumgarner, E., & Chatterji, M. (2014). Understanding validity issues in international large-scale assessments. Quality Assurance in Education, 22(1), 31–41. https://doi.org/10.1108/qae-12-2013-0050

McMillan, J. H. (2026). Classroom assessment validation: Proficiency claims and uses. *Educational Measurement: Issues and Practice*, 45(1), Article e70014. https://doi.org/10.1111/emip.70014

Mor, E., & Karatoprak Erşen, R. (2023). Implications of current validity frameworks for classroom assessment. *International Journal of Assessment Tools in Education*, 10(Special Issue), 164–173. https://doi.org/10.21449/ijate.1368458

Opesemowo, O. A. G., Opatunji, K. O., Babatimehin, T., & Opesemowo, T. R. (2026). Analysis of 2022 and 2023 Osun State basic education certificate examination mathematics items using item response theory: Implications for large scale assessment. *Social Sciences & Humanities Open*, 13, Article 102381. https://doi.org/10.1016/j.ssaho.2025.102381

Rezigalla, A. A., Eleragi, A. M. E. S. A., 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, Article 445. https://doi.org/10.1186/s12909-024-05433-y

Rijmen, F., Jeon, M., von Davier, M., & Rabe-Hesketh, S. (2014). A third-order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys. *Journal of Educational and Behavioral Statistics*, 39(4), 235–256. https://doi.org/10.3102/1076998614531045

Woodcock, S., Howard, S. J., & Ehrich, J. (2020). A within-subject experiment of item format effects on early primary students' language, reading, and numeracy assessment results. School Psychology, 35(1), 80–87. https://doi.org/10.1037/spq0000340

Wu, X., Li, N., Wu, R., & Liu, H. (2025). Cognitive analysis and path construction of Chinese students' mathematics cognitive process based on CDA. *Scientific Reports*, 15(1). https://doi.org/10.1038/s41598-025-89000-5

Wu, X., Wu, R., Chang, H.-H., Kong, Q., & Zhang, Y. (2020). International comparative study on PISA mathematics achievement test based on cognitive diagnostic models. *Frontiers in Psychology*, 11. https://doi.org/10.3389/fpsyg.2020.02230

Schneider, J. (2017). *Beyond test scores: A better way to measure school quality*. Harvard University Press.

Kallus, N., & Zhou, A. (2019). The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *Advances in neural information processing systems*, *32*.