MIKAILALSYS
Journal of Multidisciplinary Sciences

e-ISSN : 2987-2286
p-ISSN : 2987-3924
Index: Harvard, Boston, Sydney, Copernicus, Sinta 4, Scilit, Semantic, Google, etc

https://doi.org/10.58578/mikailalsys.v4i3.12457

Peer Assessment of Item Relevance in Mathematics Ability Test Instruments Designed by Mathematics Education Students: A Gregory Coefficient Approach

Nur Ismiyati & Intan Buhati Asfyra
Universitas Negeri Makassar, Indonesia
nurismiyati@unm.ac.id; intan.buhati@unm.ac.id

Submitted:	Revised:	Accepted:	Published:
Sep 18, 2026	Oct 17, 2026	Oct 28, 2026	Nov 2, 2026

Abstract

Designing test items that are relevant to their intended indicators is a core competency for prospective researchers and mathematics educators, yet empirical evidence from peer assessments of student-designed instruments remains limited. This study examined item–indicator relevance in mathematics ability test instruments developed by mathematics education students using peer assessment and the Gregory coefficient. A descriptive quantitative approach with a document-analysis design was employed, involving 15 constructed-response test instruments comprising 150 items assessed by two peers. Relevance ratings on a four-point scale were analyzed using the Gregory technique, descriptive statistics, and categorical agreement percentages in Microsoft Excel. The mean Gregory coefficient across instruments was 0.91; ten instruments achieved coefficients of 1.00, and twelve had coefficients ≥ 0.80. For the nine instruments with traceable item-level scores, the analysis covered 180 ratings, of which 77% were assigned the maximum score of 4. The mean rating was 3.76, and categorical agreement was 96.7%. For the remaining six instruments, coefficients were available only in student summary tables, with a mean of 0.83. The analysis also identified discrepancies between rating sheets and summaries, identical summaries, and misinterpretations of the coefficient. Although peer assessment provided useful feedback and indicated generally high item relevance, ceiling effects and reporting inconsistencies limited confidence in the evidence. The findings highlight the need for clear

Mikailalsys is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

relevance rubrics, rater calibration, verification of summary calculations, and complementary expert assessment to strengthen instrument development and evaluation.

Keywords: Item–Indicator Relevance; Gregory Coefficient; Constructed-Response Tests; Peer Assessment; Mathematics Education

INTRODUCTION

The quality of research and assessment practices in mathematics education depends heavily on the quality of the instruments used. Validity is understood as evidence supporting the interpretation and use of test scores (Raymond et al., 2025; Sireci & Benítez, 2023), and content validity is one source of such evidence, concerning the extent to which test items represent and are relevant to the domain being measured (Sireci & Benítez, 2023). For Mathematics Education students, who will eventually design instruments for thesis research and classroom assessment, the ability to construct tests and examine their relevance is part of assessment literacy that supports learning (Hull & Vígh, 2026; Oo et al., 2022; Pastore, 2023).

The student-designed mathematics ability test instruments included the targeted abilities, a test blueprint, and test items. Two peers were then asked to assess the relevance of each item to its corresponding indicator. Peer assessment has substantial pedagogical value because it trains both assessors and those being assessed to provide and use feedback (Nicol & McCallum, 2022; Ozyer, 2025; Topping, 2021). However, peer assessment should be interpreted cautiously when used as evidence of content validity, because recommended practice involves an expert panel with an adequate number of assessors (Hurtado-Arenas et al., 2025; Jeldres et al., 2023). The Gregory technique combines ratings from two assessors in a 2×2 matrix and produces a single coefficient (Gregory, 2015). It is widely used in this context (Larasati & Syamsurizal, 2022) because it is straightforward, although its simplicity can conceal the underlying score patterns and agreement.

Studies on mathematics instrument development in Indonesia generally report on a single instrument that has been validated by experts and, in some cases, empirically tested. Examples include instruments measuring mathematical reflective thinking (Anggraini & Muntazhimah, 2021), junior high school mathematics tests (Pandra & Aswarliansyah, 2023),

mathematics learning independence (Prananto et al., 2022), and mathematical connections (Putri et al., 2025). Meanwhile, methodological literature on content-validity indices discusses assessor criteria, coefficient thresholds, and chance agreement (Hurtado-Arenas et al., 2025; Jeldres et al., 2023). In the authors' limited literature search, no study was identified that examined a collection of student-designed instruments as a snapshot of competency while also analyzing score patterns, agreement, and the consistency of peer-assessment reporting. This study addresses that gap.

The novelty of this study lies in combining analyses that go beyond the Gregory coefficient to include peer-rating patterns, categorical agreement, sensitivity analysis, examination of documentation anomalies, and pedagogical implications for research-methods courses. The theoretical foundation is the concept of content validity as the relevance and representativeness of items with respect to the measurement domain (Sireci & Benítez, 2023), with the Gregory technique used as the quantification procedure (Gregory, 2015).

Accordingly, this study aimed to describe the relevance of items in mathematics ability test instruments designed by Mathematics Education students, based on peer assessment using the Gregory coefficient; describe relevance-score patterns and categorical agreement between assessors; and identify discrepancies and data anomalies that should be considered when interpreting and reporting the results.

METHODS

This study used a descriptive quantitative approach. It aimed to describe the content relevance of the instruments and their rating patterns, rather than test relationships or effects between variables. The study employed an evaluative, retrospective document-analysis design. The documents analyzed were previously developed test instruments and item-to-indicator relevance assessments completed by two peers.

The population comprised 31 instruments developed by students in Class C of the 2024 cohort of the Mathematics Education program. Purposive sampling was conducted using three inclusion criteria: (a) the instrument was a constructed-response test with clearly specified target abilities and a test blueprint; (b) item-relevance assessments by two peers were traceable; and (c) the instrument measured mathematical ability. Sixteen instruments met criteria (a) and (b) and were coded I1–I16. One instrument (I4) did not meet criterion

(c), because it assessed understanding of learning motivation based on Self-Determination Theory; it was excluded from the main analysis and included only in the sensitivity analysis. Thus, the main analysis comprised 15 instruments (150 items, with 10 items per instrument). Four other instruments were excluded because only one assessor's sheet was available or the documentation was incomplete; the remaining instruments had no traceable assessment documents at the time of analysis.

The main-analysis dataset was divided into two groups according to source traceability. The first group consisted of nine instruments (I1-I3 and I5-I10; 90 items) for which item-level scores from both assessors were available: eight from rating sheets and one (I2) from a student summary table. The second group consisted of six instruments (I11-I16; 60 items) for which only the Gregory coefficients and categories reported by students in summary tables were available. Source-selection rules were established before analysis: readable original rating sheets served as the primary source, and summary tables were used only when the sheets were unavailable. Any discrepancies between the two sources were recorded as findings.

Data were collected using item–indicator relevance rating sheets. The assessors were two peers-students in the same class-hereafter referred to as raters. Each rater evaluated the alignment of every item with its indicator on a scale of 1 (not relevant), 2 (slightly relevant), 3 (relevant), and 4 (highly relevant), and could provide suggestions for improvement. Students then compiled both raters' scores in a Gregory table. The researchers collected all documents, read the scores from rating sheets and summary tables, and checked the two sources against each other when both were available. Because the study focused on item relevance as one source of content-validity evidence, instrument reliability was not calculated. Data trustworthiness was supported by cross-checking rating sheets and summaries, visually inspecting original image-based or scanned sheets, and documenting every discrepancy as a finding.

Scores of 1-2 were categorized as low relevance, while scores of 3-4 were categorized as high relevance. The two raters' assessments were arranged in a 2 × 2 matrix: cell A (both low), B (Rater 1 low and Rater 2 high), C (Rater 1 high and Rater 2 low), and D (both high). The Gregory coefficient was calculated using Equation (1). The mean coefficient across instruments was calculated as the arithmetic mean using Equation (2), where n is the number of instruments and Vi is the coefficient for instrument i; this value was used solely as a

descriptive statistic. Categorical agreement, defined as the proportion of items assigned the same category by both raters (both relevant or both not relevant), was calculated using Equation (3).

$$V = \frac{D}{\mathrm{A} + \mathrm{B} + \mathrm{C} + \mathrm{D}} \quad (1)$$

$$R = \frac{\mathrm{V_1} + \mathrm{V_2} + \ldots + \mathrm{V_n}}{n} \quad (2)$$

$$Agreement = \frac{A+D}{A+B+C+D} \times 100\% \quad (3)$$

The researchers established the following interpretation criteria for each instrument: a coefficient of 1.00 meant that both raters considered all items relevant; coefficients of 0.80-0.99 were classified as adequate; and coefficients below 0.80 were classified as requiring revision. No threshold was applied to the mean across instruments because the mean Gregory coefficient across instruments is not a scale-level validity index in the usual methodological sense and has no established cutoff. Descriptive statistics (frequency, percentage, mean, and standard deviation) were calculated in Microsoft Excel. Categorical agreement was not interpreted as inter-rater reliability. Cohen's kappa was not reported because nearly all items were rated as relevant, making kappa uninformative under such a highly prevalent category (Tan et al., 2024; Vach & Gerke, 2023). Sensitivity analysis recalculated the mean coefficient across several dataset compositions. Student and rater identities were anonymized using codes, and the analysis focused on instruments rather than individuals.

RESULTS

Instrument characteristics are presented in Table 1. Of the 15 instruments in the main analysis, 10 (66.7%) measured problem-solving ability; the others measured reasoning (two instruments), communication and representation (one instrument each), while one instrument did not explicitly state the ability being measured. The content areas included numbers, algebra, matrices, sequences and series, geometry, statistics, and probability. I4 is listed in Table 1 but was excluded from the main analysis because it assessed understanding of learning motivation rather than mathematical ability.

Table 1. Characteristics of the Instruments Analyzed

Code	Ability Measured	Topic
I1	Problem solving and probabilistic thinking	Probability
I2	Mathematical communication	Statistics
I3	Problem solving	Matrices
I4	Critical thinking and conceptual understanding	Learning motivation (SDT); excluded from the main analysis
I5	Conceptual understanding and reasoning	Arithmetic sequences and series
I6	Mathematical representation	Geometric transformations
I7	Problem solving	Numbers
I8	Problem solving	Numbers
I9	Problem solving	Statistics
I10	Not explicitly stated	Basic algebraic operations
I11	Problem solving	Integers
I12	Problem solving	Algebra
I13	Mathematical reasoning	Statistics
I14	Problem solving	Sequences and series
I15	Problem solving	Numbers
I16	Problem solving	Counting rules (permutations and combinations)

Gregory coefficients are presented in Table 2 and Figure 1, distinguishing results with traceable item-level scores from those based on summary tables. Among the nine instruments with traceable item-level scores (90 items), 87 items (96.7%) were rated relevant by both raters, with a mean coefficient of 0.97; eight instruments had a coefficient of 1.00, and one instrument (I7) had a coefficient of 0.70. Among the six instruments based on summaries (60 items), 50 items (83.3%) were reported as relevant by both raters, with a mean coefficient of 0.83; two instruments had a coefficient of 1.00, and the remaining coefficients were 0.90, 0.80, 0.70, and 0.60. For this second group, the raw scores could not be independently verified.

Across all 15 instruments in the main analysis, the mean coefficient was 0.913, ranging from 0.60 to 1.00. Ten instruments (66.7%) had a coefficient of 1.00; two (13.3%) had coefficients of 0.80-0.99 (I11 and I14); two (13.3%) had coefficients of 0.70 (I7 and I12); and one (6.7%) had a coefficient of 0.60 (I13). Twelve instruments (80.0%) achieved a coefficient of at least 0.80. This combined figure is descriptive and combines two groups of data with different levels of traceability.

Table 2. Gregory Coefficients (Item Relevance) by Instrument

Kode	Items	D	Coefficient	Mean Score	Score Source	Category
I1	10	10	1.00	3.90	Sheet	Very adequate
I2	10	10	1.00	3.60	Summary	Very adequate
I3	10	10	1.00	3.60	Sheet	Very adequate
I5	10	10	1.00	3.95	Sheet	Very adequate
I6	10	10	1.00	3.75	Sheet	Very adequate
I7	10	7	0.70	3.45	Sheet	Needs revision
I8	10	10	1.00	4.00	Sheet	Very adequate
I9	10	10	1.00	3.80	Sheet	Very adequate
I10	10	10	1.00	3.75	Sheet	Very adequate
I11	10	9	0.90	-	Summary	Adequate
I12	10	7	0.70	-	Summary	Needs revision
I13	10	6	0.60	-	Summary	Needs revision
I14	10	8	0.80	-	Summary	Adequate
I15	10	10	1.00	-	Summary	Very adequate
I16	10	10	1.00	-	Summary	Very adequate
Subtotal I1-I3, I5-I10	90	87	0.97	3.76	-	-
Subtotal I11-I16 (Summary)	60	50	0.83	-	-	-
Total (15 instrumens)	150	137	0.91	-	-	-

Note: D = number of items rated relevant (score 3-4) by both raters; Rating sheet = scores from the original rating sheet; Summary = coefficient from the student summary table (item-level scores for I2 also came from the summary); mean score was calculated from both raters. The overall coefficient is a descriptive arithmetic mean.

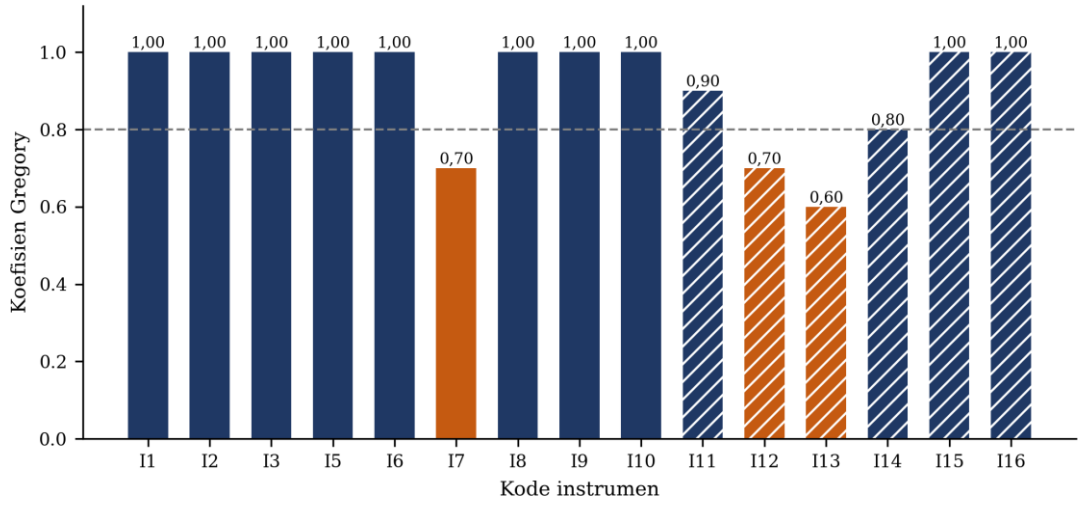


Figure 1. Gregory coefficients by instrument (main analysis, excluding I4); hatched bars = coefficients from student summaries; dashed line = 0.80.

The score distribution for the nine instruments with traceable item-level scores (180 ratings) is shown in Table 3. There were no ratings of 1, three ratings of 2 (1.7%), 38 ratings of 3 (21.1%), and 139 ratings of 4 (77.2%); the mean was 3.76 and the standard deviation was 0.47. Mean scores by instrument ranged from 3.45 (I7) to 4.00 (I8). Of the 90 pairs of item ratings, both raters assigned the same category (both relevant) to 87 items (categorical agreement of 96.7%). For the other three items, all from I7, Rater 1 considered the items relevant while Rater 2 rated them as slightly relevant. This agreement is categorical and should not be equated with inter-rater reliability.

Table 3. Distribution of Relevance Scores for 9 Instruments with Item-Level Scores (180 Ratings)

Skor	Frequency	Percentage
1 (Not Relevant)	0	0.0%
2 (Less Relevant)	3	1.7%
3 (Relevant)	38	21.1%
4 (Highly Relevant)	139	77.2%
Total	180	100.0%

Note: mean score = 3.76; standard deviation = 0.47.

Written comments that could be examined for I5 and I9, as well as for I4 which was excluded from the main analysis and is used here only as a qualitative illustration showed that a score of 3 was generally accompanied by comments about misalignment between indicators and items. For I4, raters noted that an item addressed only one aspect of an indicator that contained two aspects; an item referred to teachers although the indicator covered both teachers and parents; and a personal-reflection item was potentially subjective and should be accompanied by an alternative scenario. For I9, comments concerned items that did not precisely represent their indicators (for example, an item directly presented data although the indicator referred to a word problem) and ambiguous wording. For I5, raters suggested aligning the indicators and items for items 7 and 9.

The main findings are summarized in Tables 1-4 and Figure 1. Table 2 and Figure 1 present coefficients by instrument and identify the score source; Table 3 presents the score distribution; and Table 4 presents a sensitivity analysis of the mean coefficient under several dataset compositions. The mean coefficient ranged from 0.900 to 0.919 in compositions combining both data groups, whereas the two groups analyzed separately yielded 0.967 (item-level scores) and 0.833 (summaries). This difference indicates that the combined figure is strongly influenced by the composition of the data sources.

Table 4. Sensitivity Analysis of the Mean Gregory Coefficient Across Instruments

Dataset Composition	Number of Instrumens	Mean Coefficient
Main analysis (excluding I4)	15	0.913
All instruments, including I4	16	0.919
Main analysis excluding I15 and I16 (identical summary data)	13	0.900
Main analysis incorporating student summary scores for I1 (0.90) and I13 (0.70)	15	0.913
Only instruments with item-level scores (I1-I3, I5-I10)	9	0.967
Only instruments with summary coefficients (I11-I16)	6	0.833

Data anomalies identified in this study included the following. First, three items were rated 2 by Rater 2, all in I7 (items 1, 2, and 9), which had a coefficient of 0.70. Second, uniform scores of 4 across all items occurred in four of the 18 rating sheets (22.2%): Rater 2 for I1 and I5, and both raters for I8; these cases were concentrated in three instruments. Third, discrepancies were found between rating sheets and summaries. For I1, the summary table assigned a score of 2 to item 9 by Rater 2 (coefficient 0.90), whereas Rater 2's rating sheet assigned a score of 4 to every item. For I13, item 4 received a score of 2 on Rater 1's sheet but 3 in the summary table. In accordance with the source-selection rule established before analysis, the original sheets were used; thus, I1's coefficient became 1.00 and I13's became 0.60. The summary-based values (0.90 and 0.70) are reported in Table 4 for comparison. The two adjustments were in opposite directions and cancelled each other out in the mean across instruments (0.913 under either source). For I13, Rater 2's scores were available only in the summary, so the coefficient of 0.60 remains uncertain. Fourth, the coefficient summaries for I15 and I16 were identical across all 20 paired ratings, even though their content areas differed (numbers versus permutations and combinations) and some raters were different; Table 4 therefore also reports the result excluding both instruments (0.900). Fifth, two summaries (I12 and I13) interpreted the Gregory coefficient using terms such as "high" or "very high correlation," although the coefficient represents the proportion of items rated relevant by both raters. In one summary (I13), the definitions of cells B and C were identical.

DISCUSSION

Peer assessment in this dataset was characterized by a predominance of high scores: 77.2% of the 180 ratings were 4, 21.1% were 3, and only 1.7% were 2. This pattern indicates limited discriminatory power, particularly because approximately one-fifth of the rating

sheets assigned a score of 4 to every item. A tendency for peer assessors to overrate work and to be less sensitive to lower-quality work has also been reported elsewhere (Van Meenen et al., 2022). These findings are consistent with literature indicating that peers may assign higher scores than expert assessors, especially to lower-quality work, and may favor upper-range or uniform scores (Istikharoh, 2025; Power & Tanner, 2023).

This study cannot establish the causes of this pattern. It may reflect items that are genuinely relevant, shared understanding among students, reluctance to criticize peers, or differences in understanding the relevance criteria. Without expert ratings for comparison, any interpretation remains speculative.

The high categorical agreement (96.7%) should be interpreted cautiously because nearly all items fell into the relevant category, leaving very limited variation in the data. A high percentage of agreement does not necessarily indicate high inter-rater reliability (Derksen et al., 2024; Tan et al., 2024). This resembles the paradox of high agreement with low kappa discussed by Tan et al. (2024) and Vach & Gerke (2023). Nevertheless, peer assessment did distinguish some items: I7 identified three items requiring revision, and written comments on I5 and I9 showed that raters could identify misalignment between indicators and items even when assigning a score of 3. In this course context, the main value of peer assessment appears to lie in qualitative feedback, consistent with literature emphasizing the learning benefits of peer assessment (Nicol & McCallum, 2022; Ozyer, 2025; Topping, 2021).

The clearest methodological finding concerns reporting problems rather than the coefficient values themselves. Discrepancies between rating sheets and summaries for I1 and I13 show that coefficients can change depending on the score source used. Identical summaries for I15 and I16 raise questions about data provenance, while the combined value (0.913) obscures the difference between the traceable group (0.967) and the summary-based group (0.833). Therefore, the combined value is better interpreted as a descriptive statistic than as a validity index.

Interpretation is also problematic. The Gregory coefficient represents the proportion of items rated relevant by both raters (Gregory, 2015); therefore, labeling it as a correlation coefficient, as occurred in two summaries, is incorrect. This error illustrates how the simplicity of a technique can lead to conceptual misunderstandings when its underlying

principles are not understood, underscoring the need to strengthen assessment literacy among prospective teachers (Hull & Vígh, 2026; Oo et al., 2022; Pastore, 2023).

Studies on mathematics instrument development have reported instruments as suitable following expert review, sometimes supplemented by empirical testing (Anggraini & Muntazhimah, 2021; Pandra & Aswarliansyah, 2023; Prananto et al., 2022). In contrast, those studies involved experts and, in some cases, empirical student data, whereas the present study relied on peer assessment without pilot testing. The use of only two raters also falls short of recommended practice: content-validity literature relates critical index values to panel size and raises concerns about the discriminatory power of small panels (Jeldres et al., 2023). Chance-corrected agreement using a modified kappa also requires at least three experts (Hurtado-Arenas et al., 2025), so it could not be applied to two raters. Given these limitations, the coefficients in this study should be interpreted as preliminary indications of item relevance according to peers, not as definitive evidence of content validity.

For course instructors, these findings support providing a clear item-relevance rubric, calibrating raters before assessment, requiring written comments for every item scored below 4, cross-checking rating sheets against summaries, and using a standardized summary template to reduce calculation and interpretation errors. Expert assessment should be considered as a follow-up so that peer assessment is not treated as a substitute for expert validation. Methodologically, reports should include score distributions, agreement, data sources, and sensitivity analyses, while recognizing that content relevance is only one source of validity evidence among several (Sireci & Benítez, 2023).

This study has several limitations. The raters were classmates rather than experts, and there were only two; reciprocal rating between students may also have occurred and was not controlled. The main analysis included 15 of the 31 instruments, selected according to document availability, so the findings cannot be generalized to all instruments produced by the class. For six instruments, only student-reported coefficient summaries were available; one instrument (I2) used item-level scores from a student summary; and one rating sheet (I3, Rater 2) was read using automated extraction without visual verification. The coefficient for I13 remains uncertain because Rater 2's scores were available only in the summary. This study did not assess reliability or other sources of validity evidence.

Future studies should involve an adequately sized expert panel, calculate item- and instrument-level content-validity indices and chance-corrected agreement, compare peer and

expert ratings of the same instruments, and add empirical pilot testing to assess reliability and other validity evidence. Longitudinal research could also examine whether assessment-literacy training improves the quality of instrument design and assessment.

CONCLUSION

Across the 15 student-designed mathematics ability test instruments, the mean Gregory coefficient across instruments was 0.91, ranging from 0.60 to 1.00; 10 instruments had a coefficient of 1.00, and 12 had coefficients of at least 0.80. Among the nine instruments with traceable item-level scores, 77% of ratings were 4 and categorical agreement was 96.7%, whereas the six instruments based on summaries had a mean coefficient of 0.83. Peer assessment produced useful feedback, but the concentration of scores at the upper end, discrepancies between rating sheets and summaries, identical summaries, and incorrect interpretations of the coefficient limited the strength of the evidence. This study provides an empirical snapshot of peer assessment of item relevance in student-designed instruments and shows that the Gregory coefficient should be interpreted alongside data sources, score distributions, categorical agreement, and sensitivity analysis. Its contribution is methodological and pedagogical, particularly for research-methods courses, and it is not intended to provide definitive evidence of instrument validity. Future research should involve expert raters, use content-validity indices and chance-corrected agreement measures, compare peer ratings with expert ratings, and empirically test instruments to obtain evidence of reliability and other forms of validity.

REFERENCES

Anggraini, A., & Muntazhimah. (2021). Pengembangan Instrumen Kemampuan Berpikir Reflektif Matematis Siswa Madrasah Aliyah. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika, 10*(4), 2465–2476. https://doi.org/10.24127/ajpm.v10i4.4223

Derksen, B. M., Bruinsma, W., Goslings, J. C., & Schep, N. W. L. (2024). The kappa paradox explained. *The Journal of Hand Surgery, 49*(5), 482–485. https://doi.org/10.1016/j.jhsa.2024.01.006

Gregory, R. J. (2015). *Psychological testing: History, principles, and applications* (7th ed.). Pearson Education Limited. https://www.pearson.com/en-au/subject-catalog/p/psychological-testing-history-principles-and-applications-global-edition/P200000004343/9781292067551

Hull, P., & Vígh, T. (2026). Preservice teachers' assessment literacy: A systematic review. *Studies in Educational Evaluation, 89*, Article 101600. https://doi.org/10.1016/j.stueduc.2026.101600

Hurtado-Arenas, P., Guevara, M. R., & González-Chordá, V. M. (2025). Comparison of content validity indices for clinical nursing research: A practical case. *Enfermería Clínica (English Edition), 35*(3), Article 502214. https://doi.org/10.1016/j.enfcle.2025.502214

Istikharoh, L. (2025). Exploring EFL students' peer evaluation practices in collaborative projects: Patterns, biases, and pedagogical implications. *Journal of Educational Analytics, 4*(3), 517–530. https://doi.org/10.55927/jeda.v4i3.242

Larasati, F., & Syamsurizal, S. (2022). Validitas Instrumen Tes Keterampilan Berpikir Kritis Peserta Didik Kelas XII SMA/MA tentang Materi Mutasi. *Journal on Teacher Education, 4*(1), 365–372.

Nicol, D., & McCallum, S. (2022). Making internal feedback explicit: Exploiting the multiple comparisons that occur during peer review. *Assessment & Evaluation in Higher Education, 47*(3), 424–443. https://doi.org/10.1080/02602938.2021.1924620

Oo, C. Z., Alonzo, D., & Asih, R. A. (2022). Acquisition of teacher assessment literacy by pre-service teachers: A review of practices and program designs. *Issues in Educational Research, 32*(1), 352–373. https://www.iier.org.au/iier32/oo-abs.html

Özyer, K. K. (2025). The impact of online peer assessment on academic performance in higher education: A meta-analytic review. *Journal of Educators Online, 22*(3). https://doi.org/10.9743/JEO.2025.22.3.11

Pandra, V., & Aswarliansyah. (2023). Validitas Isi Instrumen Tes Matematika Siswa Kelas VII SMP Semester Ganjil. *Jurnal Pendidikan Matematika: Judika Education, 6*(2), 121–128. https://doi.org/10.31539/judika.v6i2.7631

Pastore, S. (2023). Teacher assessment literacy: A systematic review. *Frontiers in Education, 8*, Article 1217167. https://doi.org/10.3389/feduc.2023.1217167

Power, J. R., & Tanner, D. (2023). Peer assessment, self-assessment, and resultant feedback: An examination of feasibility and reliability. *European Journal of Engineering Education, 48*(4), 615–628. https://doi.org/10.1080/03043797.2023.2185769

Prananto, I. W., Rakhmawati, Y., & Pamungkas, T. (2022). Content validity ratio (CVR), content validity index (CVI) and confirmatory factor analysis (CFA) in mathematics learning independence instruments. *Kontinu: Jurnal Penelitian Didaktik Matematika, 6*(2), 116–132. https://doi.org/10.30659/kontinu.6.2.116-132

Putri, H. E., Alam, B. C., Rahayu, P., Alindra, A. L., Hikmatunisa, N. P., Nuraeni, F., & Krisnawati, N. (2025). Pengembangan Instrumen Penelitian untuk Mengukur Kemampuan Koneksi Matematis Peserta Didik di Sekolah Dasar. *Pendas: Jurnal Ilmiah Pendidikan Dasar, 10*(1), 1007–1018. https://journal.unpas.ac.id/index.php/pendas/article/view/21239

Raymond, J., Dai, D. W., & McAllister, S. (2025). The interpretation-use argument–The essential ingredient for high quality assessment design and validation. *Advances in Health Sciences Education, 30*(4), 1313–1332. https://doi.org/10.1007/s10459-024-10392-6

Romero Jeldres, M., Díaz Costa, E., & Faouzi Nadim, T. (2023). A review of Lawshe's method for calculating content validity in the social sciences. *Frontiers in Education, 8*, Article 1271335. https://doi.org/10.3389/feduc.2023.1271335

Sireci, S., & Benítez, I. (2023). Evidence for test validation: A guide for practitioners. *Psicothema, 35*(3), 217–226. https://doi.org/10.7334/psicothema2022.477

Tan, K. S., Yeh, Y.-C., Adusumilli, P. S., & Travis, W. D. (2024). Quantifying interrater agreement and reliability between thoracic pathologists: Paradoxical behavior of Cohen's kappa in the presence of a high prevalence of the histopathologic feature in lung cancer. *JTO Clinical and Research Reports, 5*(1), Article 100618. https://doi.org/10.1016/j.jtocrr.2023.100618

Topping, K. (2021). Peer assessment: Channels of operation. *Education Sciences, 11*(3), Article 91. https://doi.org/10.3390/educsci11030091

Vach, W., & Gerke, O. (2023). Gwet's AC1 is not a substitute for Cohen's kappa–A comparison of basic properties. *MethodsX, 10*, Article 102212. https://doi.org/10.1016/j.mex.2023.102212

Van Meenen, F., Coertjens, L., Van Nes, M.-C., & Verschuren, F. (2022). Peer overmarking and insufficient diagnosticity: The impact of the rating method for peer assessment. *Advances in Health Sciences Education, 27*(4), 1049–1066. https://doi.org/10.1007/s10459-022-10130-w