



Elementary school teachers' identification ability and perceptions of HOTS mathematics items

Briliian Prisillia   ^{1*}, Dinie Anggraeni Dewi  ¹, Deni Darmawan  ¹

¹ Universitas Pendidikan Indonesia, Bandung, Indonesia, 40625

Article info	Abstract
Keywords: elementary school teachers, HOTS, item identification, Rasch model teacher perceptions.	Despite widespread policy emphasis on Higher Order Thinking Skills (HOTS) in elementary mathematics education, prior research has predominantly focused on teachers' declarative perceptions, leaving open the question of whether they can reliably identify HOTS items in practice. This gap between endorsing HOTS in principle and demonstrating operational competence represents a critical yet underexplored dimension of teacher assessment literacy. This study examines elementary school teachers' ability to identify HOTS mathematics items and investigates the relationship between their perceptions and identification performance. A descriptive quantitative approach was employed with 52 elementary school teachers in West Java, Indonesia, selected through purposive sampling. Data were collected using a performance-based HOTS item-identification test, a Rasch model analysis, and a teacher-perception questionnaire. Although teachers generally held positive perceptions of HOTS, no significant relationship was found between perception scores and identification ability ($p = -0.061$, $p = .666$), indicating that favorable attitudes do not necessarily translate into operational competence. Rasch analysis further revealed a tendency among teachers to conflate procedural complexity with genuine cognitive demand, as reflected in item misfit patterns. Teachers who had received HOTS training demonstrated higher identification ability than those without such training, whereas teaching experience revealed no significant effect. These findings highlight a gap between declarative knowledge and operational skill and call for competency-based HOTS training in elementary teacher professional development, while demonstrating the value of integrating performance-based assessment with psychometric modeling to strengthen teacher assessment literacy.

* Corresponding Author.

E-mail address: briliianprisillia@upi.edu (Briliian Prisillia)

DOI: <http://dx.doi.org/10.33578/jpfkip.v15i2.326-342>

Received 15 January 2026; Received in revised form 28 March 2026; Accepted 8 April 2026

Available online 27 April 2026

e-ISSN 2598-5949 | p-ISSN 2303-1514 © The Authors.

1. Introduction

The development of Higher-Order Thinking Skills (HOTS) has become a central focus of 21st-century education, particularly in elementary mathematics. Within the framework of Anderson & Krathwohl's (2001), in the revised Bloom's taxonomy, HOTS refers to higher-level cognitive processes analyzing (C4), evaluating (C5), and creating (C6), which require learners to deconstruct information, apply criteria-based judgment, and generate novel solutions. In mathematics education, these processes are closely associated with reasoning, problem solving, and sense-making rather than routine procedural execution (Schoenfeld, 2016).

A critical distinction must therefore be made between HOTS and two related but conceptually different constructs: procedural complexity and contextual richness. Procedural complexity refers to tasks involving multiple computational steps, while contextual richness refers to tasks embedded in real-world situations. Neither construct inherently reflects higher-order thinking. A task may appear complex or contextualized yet rely solely on routine procedures, whereas a simple task may demand deep analysis or evaluation. Research on cognitive demand confirms that the level of thinking is determined not by surface features but by the nature of reasoning processes elicited (Wakhata et al., 2023). This distinction is essential, as misinterpretation of HOTS at the conceptual level directly affects the validity of assessment practices.

A growing body of research has demonstrated that HOTS-oriented learning and assessment contribute significantly to students' mathematical literacy and conceptual understanding (Khaesarani & Ananda, 2022; Koh et al., 2018). Assessment plays a central role not only in measuring learning outcomes but also in shaping instructional practices, as what is assessed tends to determine what is taught and learned (Black & Wiliam, 2018). Despite this, the alignment between intended cognitive demand and actual classroom assessment remains problematic.

In the Indonesian context, the integration of HOTS into mathematics learning and assessment has been formally mandated through national curriculum policies. Teachers are expected to design assessment tasks that promote higher-order thinking. However, empirical evidence consistently indicates that classroom assessment practices remain dominated by routine, procedural items, with limited engagement in higher-order cognitive processes (Ali et al., 2022; Ariawan & Pratiwi, 2017; Fazilla & Siregar, 2023; N. Herawati, 2021; Maryono et al., 2022). These findings suggest that the challenge lies not only in policy implementation but also in teachers' capacity to understand and operationalize HOTS within assessment contexts.

While prior research has examined HOTS from multiple perspectives, these studies largely fall into two dominant approaches. First, numerous studies focus on the development and analysis of HOTS-based tasks, emphasizing task characteristics and cognitive demand (Khaesarani & Ananda, 2022; Koh et al., 2018). Second, others rely on self-reported data to examine teachers' perceptions and attitudes toward HOTS (E. Herawati, 2021; Isrokatun et al., 2023). Although valuable, these approaches remain limited in their ability to capture teachers' actual ability to apply HOTS criteria in practice.

This limitation points to a critical gap between declarative knowledge and operational competence in teacher assessment literacy. Teachers may express positive perceptions of HOTS yet still struggle to accurately identify whether an assessment item genuinely reflects higher-order thinking. Prior studies have documented discrepancies between perceived understanding and demonstrated ability, indicating that teachers' confidence does not necessarily correspond to their actual competence (Dewi & Meilina, 2022; Jannah et al., 2022; Retnawati, Djidu, et al., 2018; Warsudi, 2022). In the Indonesian context, although misconceptions regarding HOTS, particularly the tendency to equate it with procedural

complexity, have been reported. These findings have not been validated through performance-based measurement (E. Herawati, 2021; Isrokatun et al., 2024).

In addition to this conceptual limitation, a methodological issue persists. Most existing studies employ classical descriptive approaches that rely on aggregate statistics, which do not allow precise calibration of individual teacher ability relative to item difficulty. As a result, the structure of teachers' HOTS identification ability, including patterns of misconception, remains insufficiently understood. Contemporary psychometric approaches, such as the Rasch model, offer a more robust framework by placing person ability and item difficulty on a common scale and enabling the detection of misfit patterns that may indicate systematic misunderstanding (Noben et al., 2021; Planinic et al., 2019; Qirom & Nurlaelah, 2023; Shamsuddin et al., 2018; van de Grift et al., 2019).

Taken together, these limitations indicate that the current understanding of teachers' HOTS assessment literacy remains incomplete. Without performance-based measurement and rigorous psychometric analysis, the gap between conceptual understanding and operational competence may remain undetected, limiting the effectiveness of professional development efforts (Sahippudin, 2021; Xu & Brown, 2016).

Building on this analysis, the present study addresses a specific and underexplored gap by (1) measuring elementary school teachers' ability to identify HOTS mathematics items using a performance-based test, (2) applying the Rasch model to calibrate teacher ability relative to item difficulty, and (3) examining the relationship between identification ability and teachers' perceptions of HOTS within a unified analytical framework.

This study contributes both theoretically and practically. Theoretically, it advances understanding of the declarative–operational gap in teacher assessment literacy within the HOTS domain. In practice, it provides a diagnostic framework to inform more targeted professional development. Accordingly, the research purpose is to analyze elementary school teachers' ability to identify HOTS mathematics items and to examine how their perceptions relate to that ability, thereby providing a more accurate basis for understanding teachers' readiness to implement HOTS-based assessment in elementary mathematics learning.

2. Method

2.1 Research design

This study employed a descriptive quantitative design to map elementary school teachers' ability to identify Higher Order Thinking Skills (HOTS) mathematics items and to examine the relationship between this ability and teachers' perceptions of HOTS. This design was selected to provide an empirical, diagnostic description of teachers' actual competence and beliefs without intervening, thereby allowing the identification of patterns of misconception and the alignment between declarative understanding and operational performance.

2.2 Participants and sampling

The participants consisted of 52 elementary school teachers (35 females, 17 males) from public and private elementary schools in West Java, Indonesia, located across urban and peri-urban areas. All participants were actively teaching either as classroom teachers or mathematics teachers. Purposive sampling was employed based on the following criteria: (1) currently teaching at the elementary school level, (2) willing to participate voluntarily, and (3) accessible through the research network. This approach was considered appropriate for a diagnostic study focusing on identifying patterns of ability and misconception rather than

generating population-level generalizations. A sample size of 52 participants is considered adequate for preliminary Rasch analysis, particularly for dichotomous data. Recent studies indicate that while smaller samples may reduce the precision of item calibration, Rasch models remain relatively robust for estimating person ability in moderate sample sizes (Hagell & Westergren, 2016; O'Neill et al., 2020).

Among the participants, 28 teachers (53.8%) had received at least one formal professional development program related to HOTS assessment, defined as structured training explicitly addressing HOTS item construction or assessment design. The remaining 24 teachers (46.2%) had not received such training. Teaching experience was categorized into two groups: ≤ 10 years ($n = 30$) and > 10 years ($n = 22$), reflecting a meaningful distinction in the development of pedagogical expertise (Makhmetova et al., 2025; Retnawati, Sulistyaningsih, et al., 2018).

2.3 Instruments

The identification test consisted of 10 mathematics items developed based on Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001), representing cognitive levels C1–C6, with a focus on distinguishing HOTS (C4–C6) from non-HOTS (C1–C3). The items covered key elementary mathematics domains, including number operations, geometry, measurement, and data handling for grades 4–6. Each item was presented in a contextual format. Teachers were required to classify each item as either HOTS or non-HOTS and provide a written justification. Responses were scored dichotomously (1 = correct classification; 0 = incorrect classification) based on an expert-validated answer key. Content validity was established through expert review, yielding a high content validity index (CVI = 0.92). Prior to the main study, the instrument was piloted with 15 teachers outside the sample to ensure clarity and appropriateness.

The perception questionnaire consisted of 10 Likert-scale items (1 = strongly disagree to 5 = strongly agree) measuring three dimensions: (1) conceptual understanding of HOTS (4 items), (2) perceived importance of HOTS (3 items), and (3) confidence in implementing HOTS assessment (3 items). Two negatively worded items were reverse-scored prior to analysis. The instrument demonstrated good internal consistency (Cronbach's $\alpha = 0.87$) based on pilot testing.

2.4 Data collection

Data were collected over a two-week period in November 2023. Participants first completed the HOTS identification test, followed by the perception questionnaire. The test was administered without time constraints to allow careful consideration of each item. All responses were completed independently in a controlled setting to minimize bias. Ethical procedures, including informed consent, voluntary participation, and data confidentiality, were ensured in accordance with standard guidelines for educational research.

2.5 Data analysis

The identification test data were analyzed using the Rasch model with Winsteps software version 5.4.0 (Linacre, 2023). The Rasch model enables the transformation of raw scores into interval-level measures and allows the simultaneous estimation of person ability and item difficulty on a common scale (Bond & Fox, 2015; Planinic et al., 2019). The analysis examined person and item reliability, separation indices, item fit statistics (infit and outfit MNSQ; acceptable range 0.5–1.5; Wright & Linacre, 1994), and the Wright map to evaluate the distribution of person abilities relative to item difficulties. Items with misfit statistics outside the acceptable range were identified for further interpretation. As a complementary reference, the KR-20 coefficient was also reported as a classical test theory measure of internal consistency.

Teachers' written justifications were analyzed using a thematic coding approach to identify patterns of reasoning and misconceptions. The coding framework was developed based on theoretical distinctions in HOTS and refined iteratively. Two independent raters coded all responses, and inter-rater reliability was assessed using Cohen's kappa ($\kappa = 0.81$), indicating strong agreement. Discrepancies were resolved through discussion to reach a consensus. Perception questionnaire data were analyzed using SPSS version 26.0. Descriptive statistics (mean, standard deviation, and frequency distribution) were computed to describe teachers' perceptions. The assumption of normality was examined using the Shapiro–Wilk test. Results indicated that perception scores were approximately normally distributed ($W = 0.966$, $p = .140$), whereas Rasch person measures were not ($W = 0.949$, $p = .027$). Therefore, Spearman's rank-order correlation (ρ) was used to examine the relationship between perception scores and identification ability.

For group comparisons, normality was assessed within each subgroup using the Shapiro–Wilk test. Non-normal distributions were observed in the trained group ($W = 0.893$, $p = .008$) and the ≤ 10 years' experience group ($W = 0.928$, $p = .045$). While the not-trained group ($W = 0.945$, $p = .205$) and the >10 years group ($W = 0.950$, $p = .314$) were approximately normally distributed. As non-normal distributions were identified in at least one group for each comparison, Mann–Whitney U tests were used to compare identification ability across HOTS training backgrounds and teaching experience. Effect sizes were calculated using rank-biserial correlation (r), interpreted as negligible ($r < .10$), small ($.10$ – $.29$), medium ($.30$ – $.49$), or large ($\geq .50$). The overall research procedure is illustrated in Figure 1.

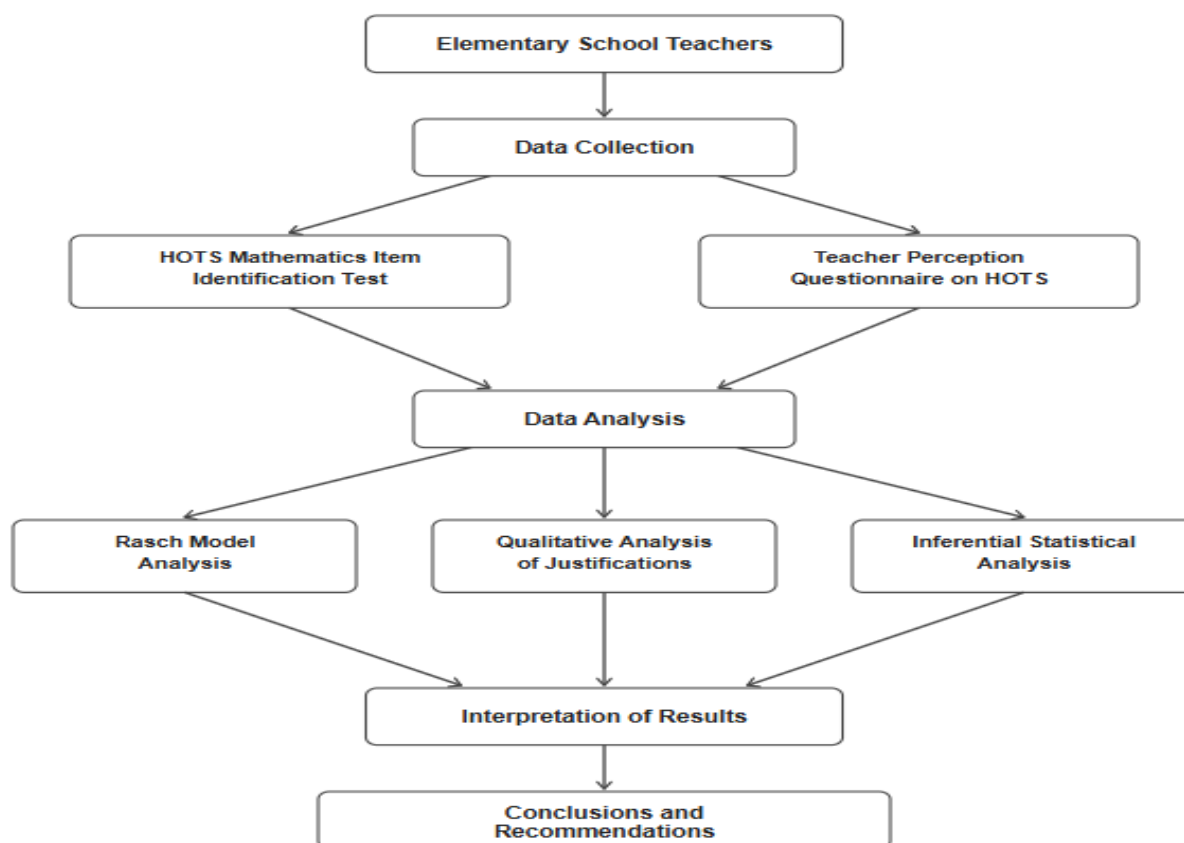


Figure 1. Research procedure

3. Results

3.1 Teachers' ability to identify hots mathematics items

The results of the HOTS mathematics item identification test indicate that the instrument demonstrated acceptable measurement properties for assessing teachers' ability to identify items. Rasch analysis showed that the items formed a stable hierarchy of difficulty, although limited person differentiation and a ceiling effect were observed. A summary of the measurement statistics is presented in Table 1.

Table 1. Summary of rasch analysis statistics for the hots identification test

Statistics	Values	Interpretations
Person reliability	0.73	Acceptable
Item reliability	0.93	Excellent
Person separation	1.65	~2 strata
Item separation	3.64	~5 strata
KR-20	0.81	Good
Person mean measure	1.49 logits	Above average
Item mean measure	0.00 logits	Centered
Person SD	2.52 logits	Wide spread
Item SD	1.94 logits	Good spread
Ceiling effect	15.4%	Limited discrimination

Item reliability, as presented above, is high (0.93), indicating stable item calibration across respondents. Person reliability (0.73) is acceptable, while the person separation index (1.65) suggests approximately two distinguishable ability levels. The person mean measure (1.49 logits) exceeds the mean item difficulty (0.00 logits), indicating the test is generally accessible for this sample.

The distribution of person abilities and item difficulties is illustrated in the Wright map. To visualize this distribution, Figure 2 presents the person-item map generated from the Winsteps output.

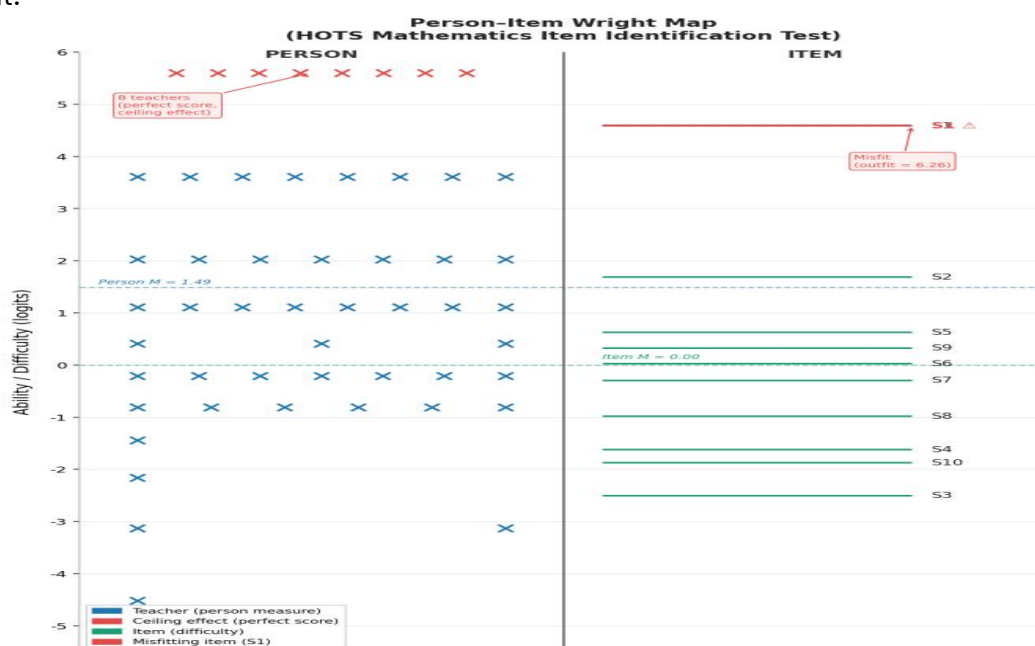


Figure 2. Wright map showing distribution of teachers' abilities and item difficulties

Notes: Each "X" on the left represents one teacher.

Items are listed on the right.

Higher positions indicate greater ability (for persons) or greater difficulty (for items).

Notably, 8 teachers (15.4%) achieve perfect scores, indicating a ceiling effect that limits the instrument's capacity to differentiate higher-ability respondents. Items S3, S10, and S4 are the easiest and are correctly identified by most teachers, while S1, S2, and S5 are the most difficult. The item difficulty range spans approximately 7 logits (−2.50 to +4.59).

3.2 Item fit analysis

Item fit statistics indicated that most items conformed to the Rasch model expectations. To facilitate interpretation, Figure 3 displays the outfit MNSQ values for all items alongside the boundary of the acceptable range.

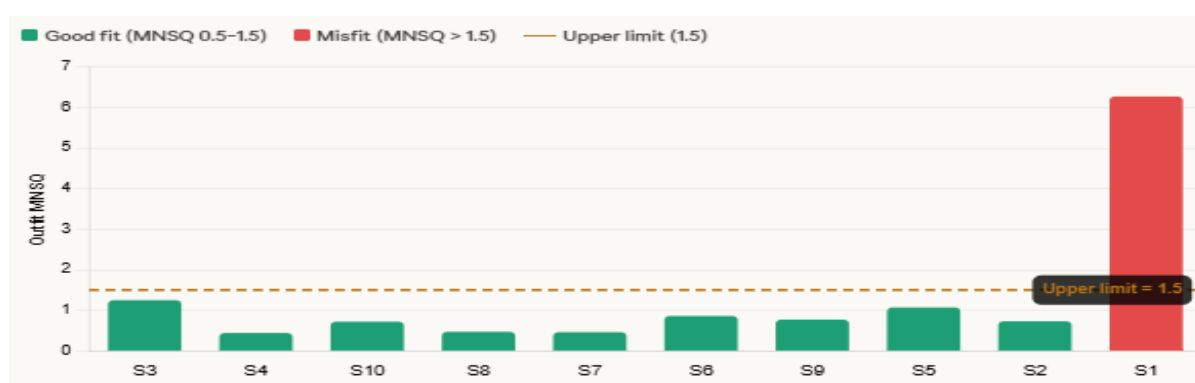


Figure 3. Outfit MNSQ values for all items with acceptable range boundary (0.5–1.5)

As shown in Figure 3, nine of ten items demonstrate acceptable fit. Item S1 exhibits substantial misfit (infit MNSQ = 1.41; outfit MNSQ = 6.26), indicating highly unpredictable response patterns. S1 is the most difficult item (4.59 logits), correctly identified by only 11 of 52 teachers (21.2%). Detailed fit statistics for all items are presented in Table 2.

Table 2. Item fit statistics and difficulty measures

Items	Difficulty (logits)	Infit MNSQ	Outfit MNSQ	Fit Status
S1	4.59	1.41	6.26	Misfit
S2	1.69	0.78	0.72	Good fit
S5	0.63	1.25	1.06	Good fit
S9	0.33	0.96	0.76	Good fit
S6	0.03	1.14	0.85	Good fit
S7	-0.29	0.69	0.45	Good fit
S8	-0.98	0.76	0.46	Good fit
S4	-1.62	0.80	0.43	Good fit
S10	-1.87	1.29	0.71	Good fit
S3	-2.50	1.23	1.24	Good fit

Qualitative analysis of justifications for S1 reveals that teachers who misclassify tend to base their responses on surface features rather than cognitive demand, consistent with the misconception patterns reported in Section 3.6. Four teachers (7.7%) also show person outfit MNSQ values above 2.0, indicating inconsistent response patterns.

3.3 Teachers' perceptions of HOTS

Teachers generally reported positive perceptions of HOTS. Descriptive statistics by dimension are presented in Table 3, followed by Figure 4, which illustrates the mean scores across the three dimensions.

Table 3. Descriptive statistics for teacher perception dimensions

Dimensions	Mean	SD	N items
Conceptual Understanding	3.82	0.54	4
Perceived Importance	4.19	0.51	3
Implementation Confidence	3.76	0.60	3
Overall Perception	3.92	0.45	10

Figure 4 presents the mean scores for each perception dimension visually, allowing comparison of relative endorsement across subscales.

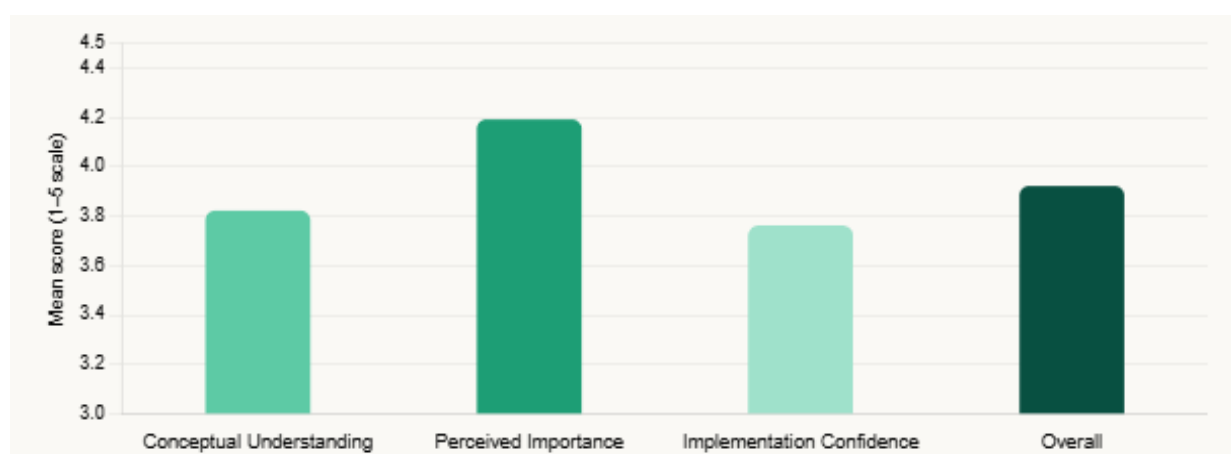


Figure 4. Mean scores of teacher perception across three dimensions (1–5 scale)

As shown in Figure 4, Perceived Importance receives the highest ratings ($M = 4.19$), whereas Implementation Confidence receives the lowest ($M = 3.76$), suggesting that teachers endorse the HOTS conceptually but feel less confident in its practical application.

3.4 Relationship between perceptions and identification ability

Given the non-normal distribution of Rasch person measures ($W = 0.949$, $p = .027$), Spearman's rank-order correlation was employed to examine the relationship between perception scores and identification ability. Results are presented in Table 4.

Table 4. Spearman rank order correlation between perception dimensions and identification ability

Variables	P	p
Total Perception Score × Identification	-0.061	.666
Conceptual Understanding × Identification	-0.126	.373
Perceived Importance × Identification	0.108	.446
Implementation Confidence × Identification	-0.108	.446

No statistically significant correlation is found between teachers' overall perception scores and identification ability ($\rho = -0.061$, $p = .666$), nor between any perception dimension and identification performance. These findings indicate that self-reported perceptions of HOTS are not meaningfully associated with actual identification ability in this sample.

3.5 Group comparisons

Given violations of normality in at least one group for each comparison, Mann–Whitney U tests were conducted to compare identification ability across groups. Figure 5 illustrates the median identification ability by group, and Table 5 presents the complete statistical results.

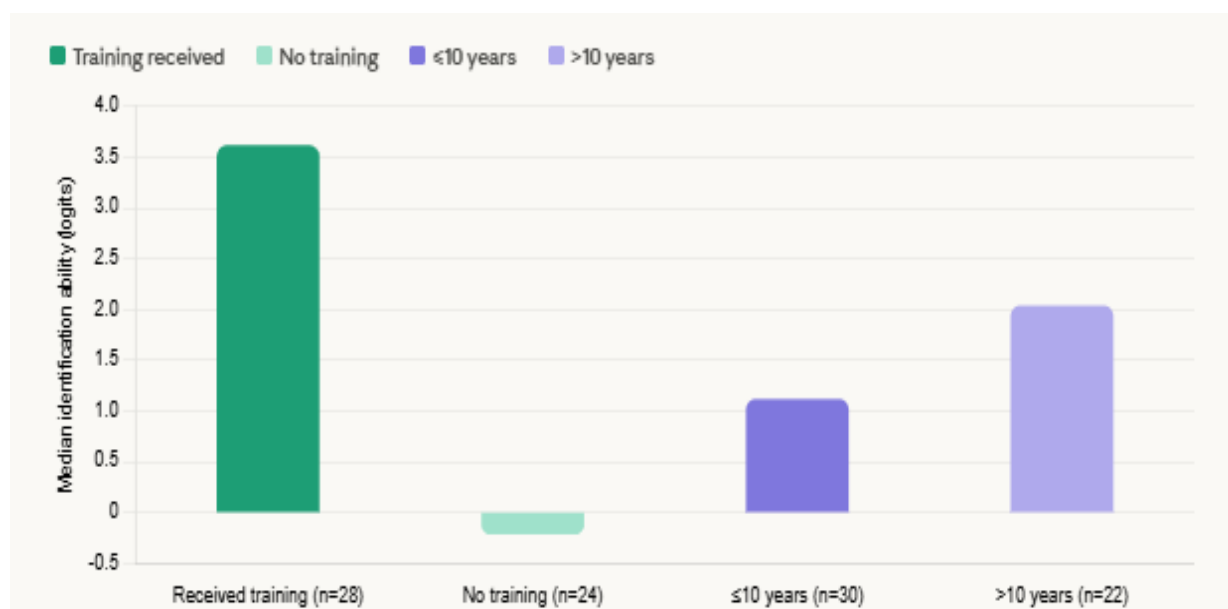


Figure 5. Median identification ability (rasch person measure in logits) by group

As illustrated in Figure 5, teachers who received HOTS training show substantially higher median identification ability than those without training. No meaningful difference is observed between experience groups. Complete statistical results are presented in Table 5.

Table 5. Mann–whitney u test results for group comparisons

Comparison	Group	N	Mdn	M	U	P	r
HOTS Training	Received	28	3.61	3.08	608.0	<.001	0.81
	Not received	24	-0.21	-0.37			
Teaching Experience	≤ 10 years	30	1.11	1.35	299.5	.575	0.09
	> 10 years	22	2.03	1.67			

Teachers who received HOTS training demonstrate substantially higher identification ability (Mdn = 3.61 logits) than those without training (Mdn = -0.21 logits), $U = 608.0$, $p < .001$, with a large effect size ($r = 0.81$). In contrast, teaching experience indicates no significant difference ($U = 299.5$, $p = .575$, $r = 0.09$), indicating that years of experience alone are not associated with higher identification ability in this sample.

3.6 Qualitative analysis of misconceptions

Analysis of teachers' written justifications revealed three recurring patterns of reasoning associated with incorrect classifications. First, 42% of incorrect responses reflected a tendency to equate HOTS with computational complexity; teachers classified items involving multiple steps or large numbers as HOTS regardless of the cognitive processes required. Second, 28% conflated contextual richness with higher-order thinking; teachers assumed that items embedded in real-world contexts automatically qualified as HOTS, even when the cognitive demand remained at the application level. Third, 19% reflected misidentifying procedural transparency as analysis; teachers classified items requiring students to show their work as HOTS, conflating step visibility with analytical thinking. These three patterns correspond directly to the conceptual distinctions between HOTS, procedural complexity, and contextual richness established in the theoretical framework, and are consistent with the item misfit observed in S1.

4. Discussion

The research findings indicate that elementary school teachers demonstrated an adequate but uneven ability to identify HOTS mathematics items. The person mean measure of 1.49 logits, positioned above the mean item difficulty of 0.00 logits, indicates that the average teacher ability exceeded the average item difficulty in this sample. However, the presence of a ceiling effect, with 15.4% of teachers achieving perfect scores, and a relatively low person separation index (1.65), indicates that the instrument may not have been sufficiently sensitive to differentiate among higher levels of teacher ability or to capture the full range of competence.

Variation in teachers' exposure to HOTS-oriented assessment practices and professional development appears to contribute to this pattern. A significant difference in identification ability between teachers who had received HOTS-related training and those who had not ($d = 0.64$, $p = 0.025$) supports this interpretation. Previous studies have similarly reported that sustained, focused training in cognitive demand and assessment design is associated with stronger analytical skills for identifying higher-order tasks (Ali et al., 2022; Isrokatun et al., 2024). These findings indicate that the ability to identify HOTS is not merely intuitive but develops through structured learning experiences.

Another contributing factor may be the continued dominance of procedural assessment practices in elementary mathematics classrooms. When assessment tasks predominantly emphasize routine problem-solving and algorithmic procedures, teachers may internalize HOTS as an extension of task difficulty rather than as a qualitatively distinct cognitive construct. Prior studies have reported similar tendencies, where teachers equated HOTS with complex calculations or multi-step procedures instead of reasoning, evaluation, or creative problem-solving processes (Koh et al., 2025; Zhou et al., 2023). This pattern explains why teachers' identification abilities, while generally adequate, clustered within a limited range despite exposure to varied item types.

4.1 The Misfit Item: A window into misconceptions

The identification of Item 7 as a misfitting item (outfit MNSQ = 2.14) provides crucial diagnostic insight. This item presented a multi-step problem involving large numbers and multiple operations and was frequently misclassified as a HOTS item. Qualitative analysis revealed that 67% of teachers who incorrectly classified this item based their judgment on its perceived computational difficulty rather than on the cognitive processes required to solve it.

This pattern aligns with earlier findings indicating that teachers often rely on surface features when evaluating HOTS tasks, particularly in the absence of explicit analytical frameworks for assessing cognitive demand (Isrokatun et al., 2023). Tasks that appear complex on the surface may prompt heuristic judgments centred on difficulty, rather than careful analysis of whether the task requires analysis, evaluation, synthesis, or creation, as defined in Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001).

Importantly, the misfit was systematic rather than random, suggesting a shared conceptual tendency among teachers to conflate procedural complexity with cognitive complexity. This consistency indicates that the misconception is embedded within a broader interpretive framework rather than being attributable to individual error.

4.2 Perceptions of HOTS: positive but insufficient

Teachers' perceptions of HOTS were generally positive, with an overall mean score of 4.12. This finding is consistent with previous studies reporting strong conceptual endorsement of higher-order thinking among elementary school teachers (E. Herawati, 2021; Nirtha et al., 2024). Teachers expressed strong agreement regarding the importance of HOTS for 21st-century learning ($M = 4.65$) and its integration into assessment practices ($M = 4.38$).

However, positive perceptions did not consistently align with accurate operational understanding. This discrepancy reflects the distinction between declarative knowledge (knowing what HOTS is) and procedural competence (knowing how to apply HOTS concepts in practice), a gap frequently discussed in research on teacher cognition and assessment literacy (Edelsbrunner et al., 2024; Xu & Brown, 2016). Teachers may possess conceptual awareness of HOTS without the analytical skills needed to evaluate assessment items accurately.

The absence of a statistically significant correlation between perception scores and identification ability ($\rho = -0.061$, $p = .666$) supports this interpretation. These findings indicate that self-reported perceptions of HOTS, across dimensions, were not meaningfully associated with actual identification ability in this sample, suggesting that additional factors, such as formal training, exposure to assessment frameworks, or reflective practice, may be more decisive in shaping operational competence.

Lower ratings on perception indicators related to classroom implementation ($M = 3.88$) compared to perceived importance ($M = 4.46$) further suggest uncertainty in applying HOTS concepts in practice. Similar patterns have been observed in previous studies, in which teachers expressed confidence in HOTS at the conceptual level but reported difficulties applying HOTS principles to item analysis or assessment design (Ali et al., 2022; Zhou et al., 2023). In such situations, teachers may revert to surface-level indicators such as task difficulty or contextual complexity.

Teachers' strong expressed need for further professional development ($M = 4.27$ on Item 9) also indicates awareness of these limitations and a recognition of gaps between understanding and practice.

4.3 Integrating performance and perception data

The integration of performance-based identification data and perception data provides a more comprehensive overview of teachers' readiness to engage with HOTS assessment. Previous research has emphasized the importance of combining self-reported measures with objective performance indicators to avoid overestimating competence based solely on perceptions (Bailes & Nandakumar, 2020; Planinic et al., 2019). The research findings support this perspective by

demonstrating that conceptual endorsement of HOTS does not necessarily translate into accurate identification performance.

The application of the Rasch model further strengthened this analysis by placing teachers and items on a common interval scale, enabling detailed examination of ability–difficulty relationships (Bond & Fox, 2015). The Wright map illustrated patterns of item difficulty and teacher ability, the ceiling effect highlighted limitations in measurement sensitivity, and item fit statistics revealed specific misconceptions, particularly the tendency to equate computational complexity with higher-order thinking.

4.4 Generalizability and contextual considerations

Despite the finding that this study was conducted in elementary schools in West Java, Indonesia, the gap between teachers' positive perceptions of HOTS and their ability to operationalize it has been widely reported across educational contexts and subject areas. International studies have documented similar challenges in aligning conceptual understanding of HOTS with assessment practices (Azid et al., 2022; Koh et al., 2025; Thompson, 2017).

Furthermore, the analytical approach used in this study, which combines performance-based assessment with Rasch measurement, offers a transferable framework for examining teacher assessment literacy across diverse contexts. Rasch-based diagnostics have been recommended for exploring latent competencies in education, particularly when the goal is to understand variation in ability rather than to rank individuals (Gordon et al., 2021; Planinic et al., 2019; Sumintono & Widhiarso, 2015).

The finding reveals that HOTS training significantly predicted identification ability ($d = 0.64$, $p = 0.025$), whereas teaching experience did not ($p = 0.389$), consistent with prior research suggesting that assessment literacy related to higher-order thinking develops through intentional learning experiences rather than experience alone.

5. Conclusion and implications

This study examined elementary school teachers' ability to identify HOTS mathematics items and their perceptions of HOTS through a performance-based identification test, Rasch measurement, and a perception questionnaire. The findings indicate that teachers in this sample demonstrated a generally adequate but unevenly distributed ability to identify. While the person mean measure (1.49 logits) exceeded the average item difficulty, a ceiling effect (15.4% perfect scores) and a low person separation index (1.65) reveal that the instrument lacked sensitivity to differentiate higher levels of competence. The misfit of S1 is the most difficult item (4.59 logits) with an outfit MNSQ of 6.26, reflecting a systematic tendency among teachers to base item classifications on surface features rather than cognitive demand, corroborated by qualitative evidence that 42% of incorrect responses referenced computational complexity, 28% contextual richness, and 19% procedural transparency as indicators of HOTS.

The central contribution of this study is the empirical demonstration that positive perceptions of HOTS do not predict operational identification ability. No significant correlation was found between perception scores and identification performance across any dimension (total $p = -0.061$, $p = .666$), providing direct evidence of the declarative operational gap in teacher assessment literacy. In contrast, formal HOTS training was a strong predictor of identification ability ($U = 608.0$, $p < .001$, $r = 0.81$). Meanwhile, teaching experience had no significant effect, suggesting that this competency develops through targeted professional engagement rather than accumulated classroom experience alone.

These findings carry implications for three domains. In pre-service education, mathematics methods courses should incorporate explicit instruction in cognitive demand analysis using Bloom's Revised Taxonomy-based frameworks. For in-service professional development, programs should move beyond general HOTS awareness to performance-based activities in which teachers classify items, justify their reasoning, and receive calibrated feedback. The large training effect observed here ($r = 0.81$) supports the value of such targeted approaches. For assessment design, item developers should make the criteria for cognitive demand explicit in teacher-facing documentation, providing clear distinctions among cognitive demand, procedural complexity, and contextual richness. By combining performance-based testing with Rasch diagnostics and perception data, this study offers a replicable framework for auditing teacher assessment literacy and designing evidence-based professional development.

6. Limitation

Several research limitations should be acknowledged when interpreting the findings. First, the sample comprised 52 elementary school teachers from a single region of West Java, Indonesia, selected through purposive sampling. This limits the statistical generalizability of the findings to the broader Indonesian elementary teaching population, and claims about national-level patterns are not warranted based on this study alone. Second, the 10-item identification test exhibited a ceiling effect, with 15.4% of teachers achieving perfect scores, indicating that the instrument was insufficiently challenging for higher-ability respondents. The limited item pool also constrains the precision of ability estimates, particularly at the upper end of the distribution. Future studies should employ larger item banks or computerized adaptive testing designs to improve measurement sensitivity across a wider range of ability. Third, the cross-sectional design does not permit causal inferences. The significant difference in identification ability between trained and untrained teachers ($U = 608.0$, $p < .001$, $r = 0.81$) is consistent with a training effect but does not establish causality without a longitudinal or experimental design. Fourth, the qualitative analysis of teachers' written justifications, while systematically coded with acceptable inter-rater reliability ($\kappa = 0.81$), is based on written responses rather than think-aloud protocols or interviews. Written justifications may not fully capture teachers' actual reasoning processes, and the coding categories, though theoretically grounded, represent one of several possible interpretive frameworks. Fifth, the perception questionnaire, despite demonstrating acceptable internal consistency in pilot testing (Cronbach's $\alpha = 0.87$), yielded a lower reliability estimate in the main study sample, which may reflect genuine heterogeneity in teachers' perceptions rather than instrument instability. At last, the study focused exclusively on item-identification ability, a dimension of teacher assessment literacy. Competencies such as HOTS item construction, scoring, and instructional use of assessment data were not examined and represent central directions for future research.

Credit authorship contribution statement

Briliian Prisillia: Conceptualization, Methodology, Formal analysis, Data curation, Original draft writer, Review & editing, Visualization. **Dinie Anggraeni Dewi:** Supervision, Validation, Review & editing, Resources. **Deni Darmawan:** Supervision, Validation, Review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The data that support the research findings are available from the corresponding author upon reasonable request. Because the data include information from human participants, they are not publicly available to protect participant privacy.

Ethical Declaration

This study was conducted in accordance with established ethical principles for educational research. All participants were fully informed about the purpose, procedures, and voluntary nature of the study prior to their involvement. Informed consent was obtained from all participants, who were assured of their right to withdraw at any time without consequence. The anonymity and confidentiality of all participants were strictly maintained throughout the data collection, analysis, and reporting processes. No personally identifiable information was disclosed in any part of this manuscript.

Acknowledgement

The authors would like to express their gratitude to all elementary school teachers in West Java, Indonesia, who volunteered to participate in this study. The authors also thank the reviewers for their constructive comments, which substantially improved the quality of this manuscript.

Declaration of AI statement

This manuscript used ChatGPT (OpenAI) as a supportive tool for language editing and readability improvement. The AI was not used to generate original research ideas, data, or analytical results. All outputs produced with the assistance of AI were carefully evaluated, revised, and integrated by the authors in accordance with academic standards. The authors remain fully accountable for the validity, reliability, and scholarly integrity of the work.

References

- Ali, S. S., Saleem, A., & Dogar, A. H. (2022). Integration of higher order thinking skills with content: A continuous professional development perspective for teachers in Pakistan. *Journal of Education and Educational Development*, 9(1), 92–108. <https://doi.org/10.22555/joeed.v9i1.531>
- Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman.
- Ariawan, M., & Pratiwi, E. (2017). Analisis kemampuan guru SD dalam penerapan pembelajaran berbasis HOTS. *Jurnal Pendidikan Dan Pembelajaran*, 4(3), 231–239.
- Azid, N., Hoe, L. S., & Hasim, S. M. (2022). Challenges in implementing higher order thinking skills in teaching and learning of mathematics: A systematic review. *Journal of Educational Research and Indigenous Studies*, 1(1), 1–12.
- Bailes, L. P., & Nandakumar, R. (2020). Using the Rasch model to evaluate teacher knowledge assessments. *Educational and Psychological Measurement*, 80(5), 990–1014. <https://doi.org/10.1177/0013164420906600>
- Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, 25(6), 551–575. <https://doi.org/10.1080/0969594X.2018.1441807>
- Bond, T. G., & Fox, C. M. (2015). Applying the Rasch model: Fundamental measurement in the human sciences (3rd ed.). Routledge.
- Dewi, T. M., & Meilina, F. (2022). PENGEMBANGAN LEMBAR KERJA PESERTA DIDIK (LKPD) BERBASIS HIGHER ORDER THINKING SKILL (HOTS) TERINTEGRASI WEB PADA

- PEMBELAJARAN IPA DI SEKOLAH DASAR. *Primary: Jurnal Pendidikan Guru Sekolah Dasar*, 11(5), 1368. <https://doi.org/10.33578/jpfkip.v11i5.9106>
- Edelsbrunner, P. A., Schalk, L., Schumacher, R., & Stern, E. (2024). Variability of teacher assessment competence: A latent state-trait analysis. *Journal of Educational Psychology*, 116(1), 43–62. <https://doi.org/10.1037/edu0000819>
- Fazilla, S., & Siregar, P. H. (2023). Analysis of Higher-Order-Thinking Skill (HOTS) Questions in the Class V Elementary Student's Books Theme 3. 12.
- Gordon, S. P., Dembo, M. H., & Hocesvar, D. (2021). Using Rasch measurement to examine the psychometric properties of the Teacher Efficacy Scale. *Educational and Psychological Measurement*, 81(5), 1007–1034.
- Hagell, P., & Westergren, A. (2016). Sample Size and Statistical Conclusions from Tests of Fit to the Rasch Model According to the Rasch Unidimensional Measurement Model (Rumm) Program in Health Outcome Measurement. *Journal of Applied Measurement*, 17(4), 416–431.
- Herawati, E. (2021). Analysis of elementary school teachers' perceptions toward higher-order thinking skills assessment in mathematics. *Journal of Elementary Education*, 14(3), 287–302.
- Herawati, N. (2021). KEMAMPUAN GURU DALAM MEMBUAT SOAL HOTS DALAM UJIAN TENGAH SEMESTER. *Primary: Jurnal Pendidikan Guru Sekolah Dasar*, 10(6), 1689. <https://doi.org/10.33578/jpfkip.v10i6.8638>
- Isrokatun, I., Destrinelli, D., & Setyaningrum, W. (2023). Analysis of HOTS-based mathematics assessment items in elementary school textbooks. *International Journal of Elementary Education*, 7(2), 215–228. <https://doi.org/10.23887/ijee.v7i2.45621>
- Isrokatun, I., Marlina, R., & Pujianto, A. (2024). Teachers' understanding and readiness in implementing HOTS-based mathematics assessment in Indonesian elementary schools. *Pegem Journal of Education and Instruction*, 14(1), 125–138.
- Jannah, F., Radiansyah, R., Sari, R., Kurniawan, W., Aisyah, S., Wardini, S., & Fahlevi, R. (2022). PEMBELAJARAN HOTS BERBASIS PENDEKATAN LINGKUNGAN DI SEKOLAH DASAR. *Primary: Jurnal Pendidikan Guru Sekolah Dasar*, 11(1), 189. <https://doi.org/10.33578/jpfkip.v11i1.8533>
- Khaesarani, I. R., & Ananda, R. (2022). Students' mathematical literacy skills in solving higher-order thinking skills problems. *Al-Jabar: Jurnal Pendidikan Matematika*, 13(1), 81–99. <https://doi.org/10.24042/ajpm.v13i1.11499>
- Koh, K., Burke, L. E. C., Luke, A., Gong, W., & Tan, C. (2025). Developing assessment literacy: An analysis of teacher learning through collaborative assessment design. *Assessment in Education: Principles, Policy & Practice*, 32(1), 78–96. <https://doi.org/10.1080/0969594X.2024.2301234>
- Koh, K., Burke, L. E. C.-A., Luke, A., Gong, W., & Tan, C. (2018). Developing the assessment literacy of teachers in Chinese language classrooms: A focus on assessment task design. *Language Teaching Research*, 22(3), 264–288. <https://doi.org/10.1177/1362168816684366>
- Linacre, J. M. (2023). Winsteps® Rasch measurement computer program (Version 5.4.0) [Computer software]. Winsteps.com.
- Makhmetova, Z., Karabassova, L., Zhakim, A., & Karinov, A. (2025). Exploring the Effects of Professional Learning Experiences on In-Service Teachers' Growth: A Systematic Review of Literature. *Education Sciences*, 15(2), 146. <https://doi.org/10.3390/educsci15020146>
- Maryono, M., Sastrawati, E., & Budiono, H. (2022). ANALISIS KESULITAN GURU SEKOLAH DASAR DALAM MENGEMBANGKAN INSTRUMEN ASESMEN HIGHER ORDER THINKING SKILLS.

- Primary: Jurnal Pendidikan Guru Sekolah Dasar, 11(5), 1529. <https://doi.org/10.33578/jpkip.v11i5.9182>
- Nirtha, I. N., Marhaeni, A. A. I. N., & Suastra, I. W. (2024). Elementary school teachers' perceptions and practices in implementing higher-order thinking skills in mathematics learning. *Journal of Education Research and Evaluation*, 8(1), 45–58.
- Noben, I., Maulana, R., Deinum, J. F., & Hofman, W. H. A. (2021). Measuring university teachers' teaching quality: A Rasch modelling approach. *Learning Environments Research*, 24(1), 87–107. <https://doi.org/10.1007/s10984-020-09319-w>
- O'Neill, T. R., Gregg, J. L., & Peabody, M. R. (2020). Effect of Sample Size on Common Item Equating Using the Dichotomous Rasch Model. *Applied Measurement in Education*, 33(1), 10–23. <https://doi.org/10.1080/08957347.2019.1674309>
- Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research*, 15(2), 020111–020111. <https://doi.org/10.1103/PhysRevPhysEducRes.15.020111>
- Qirom, M. S., & Nurlaelah, E. (2023). Rasch Model: Analysing the Items Quality of Mathematics Higher-Order Thinking Skill Instrument. *JPP (Jurnal Pendidikan Dan Pembelajaran)*; Vol 30, No 1 (2023)DO - 10.17977/Um047v30i12023p8-16. <https://journal2.um.ac.id/index.php/jpp/article/view/32165>
- Retnawati, H., Djidu, H., Kartianom, Apino, E., & Anazifa, R. D. (2018). TEACHERS' KNOWLEDGE ABOUT HIGHER-ORDER THINKING SKILLS AND ITS LEARNING STRATEGY. In *Problems of Education in the 21st Century* (Vol. 76, p. Continuous). <https://doi.org/https://doi.org/10.33225/pec/18.76.215>
- Retnawati, H., Sulistyaningsih, E., & Yin, L. Y. (2018). Students' readiness to teaching practice experience: A review from the mathematics education students' view. *Jurnal Riset Pendidikan Matematika*, 5(1), 1–17. <https://doi.org/10.21831/jrpm.v5i1.18788>
- Sahippudin, S. (2021). UPAYA KEPALA SEKOLAH MELALUI SUPERVISI DALAM MENINGKATKAN KREATIVITAS GURU DALAM PEMBELAJARAN DARING DI SDN 022 HARAPAN BARU. Primary: Jurnal Pendidikan Guru Sekolah Dasar, 10(6), 1547. <https://doi.org/10.33578/jpkip.v10i6.8607>
- Schoenfeld, A. H. (2016). *Learning to Think Mathematically: Problem Solving, Metacognition, and Sense Making in Mathematics* (Reprint). *Journal of Education*, 196(2), 1–38. <https://doi.org/10.1177/002205741619600202>
- Shamsuddin, H., Abdul Razak, N., & Zamri Khairani, A. (2018). Calibrating Students's TM Performance in Mathematics: A Rasch Model Analysis. *International Journal of Engineering and Technology*, 7(3.20), 109–113. <https://doi.org/10.14419/ijet.v7i3.20.18991>
- Sumintono, B., & Widhiarso, W. (2015). Aplikasi pemodelan Rasch pada assessment pendidikan [Application of Rasch modeling in educational assessment]. Trim Komunika.
- Thompson, P. W. (2017). Problems, Theories, and the Common Core. *Journal for Research in Mathematics Education*, 48(1), 2–6. <https://doi.org/10.5951/jresmetheduc.48.1.0002>
- van de Grift, W. J. C. M., Houtveen, T. A. M., van den Hurk, H. T. G., & Terpstra, O. (2019). Measuring teaching skills in elementary education using the Rasch model. *School Effectiveness and School Improvement*, 30(4), 455–486. <https://doi.org/10.1080/09243453.2019.1577743>
- Wakhata, R., Mutarutinya, V., & Balimuttajjo, S. (2023). Exploring the impact of Stein et al.'s levels of cognitive demand in supporting students' mathematics heuristic problem-solving abilities. *Frontiers in Education*, Volume 8-2023. <https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2023.949988>

- Warsudi, W. (2022). KEMAMPUAN GURU DALAM MEMBUAT SOAL UJIAN SEMESTER DI SDN 025 HARAPAN JAYA. *Primary: Jurnal Pendidikan Guru Sekolah Dasar*, 11(1), 138. <https://doi.org/10.33578/jpkip.v11i1.8766>
- Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, 58, 149–162. <https://doi.org/10.1016/j.tate.2016.05.010>
- Zhou, D., Liu, J., & Liu, J. (2023). Assessment design for learning: The role of HOTS items in promoting mathematical reasoning. *Educational Assessment, Evaluation and Accountability*, 35(3), 289–312. <https://doi.org/10.1007/s11092-023-09415-x>