PROSIDING DISKUSI PANEL NASIONAL PENDIDIKAN MATEMATIKA

Original Research

Validating the Assessment of Mathematics Metacognitive and Self Regulation Learning: a Rasch Model Survey

Avhiesta Mahadryanie[1], Auliya Zahrah[2], Nabilah Anaswa[3], Ayu Faradillah[4*], & Isnaini Handayani[5]

[1,2,3,4,5]Universitas Muhammadiyah Prof. DR. HAMKA

INFO ARTICLES

Key Words:

mathematical metacognition; self-regulated learning; Rasch model; instrument validation; psychometric

Abstract: *Metacognitive ability and self-regulation in mathematics learning are essential as they help students monitor their thinking process and manage their learning strategies effectively. This study used a quantitative survey approach with 292 participants to investigate the psychometric properties of an integrated Mathematics Metacognitive and Self-Regulation Learning tool. Data were analyzed using Winsteps software through Rasch-based procedures, including item and person reliability, rating scale effectiveness, Wright Map analysis, and Differential Item Functioning (DIF) across groups defined by gender, grade level, and school type. The results demonstrated satisfactory psychometric properties, with excellent item reliability (.99 and .98), adequate internal consistency (Cronbach Alpha .89 and .93), confirmed unidimensionality (raw variance explained: 46.8% and 40.1%), and item separation indices exceeding the minimum criterion. DIF analysis revealed that a small number of items functioned differently across respondent groups, while the majority demonstrated measurement equivalence. These findings confirm that the instrument provides sound validity, reliability, and measurement fit, with targeted revision of biased items recommended to further strengthen measurement quality.*

Correspondence Address: Jln Tanah Merdeka No.20, RT.11/RW.2, Rambutan, Kec. Ciracas, Kota Jakarta Timur, Daerah Khusus Ibukota Jakarta 13830, Indonesia; e-mail: ayufaradillah@uhamka.ac.id

How to Cite (APA 6th Style): Mahadryanie, A., Zahrah, A., Anaswa, N., Faradillah, A., & Handayani, I. (2026). Validating the assessment of mathematics metacognitive and self regulation learning: A Rasch model survey.*Prosiding Diskusi Panel Nasional Pendidikan Matematika*, 741-756.

INTRODUCTION

Mathematics metacognition refers to learners' awareness of their cognitive strengths and limitations as well as their ability to plan, monitor, regulate, and evaluate cognitive processes during mathematical problem solving activities (Daher, 2022; Muncer et al., 2022). A growing body of research has highlighted its contribution to mathematical problem solving, reasoning, and academic achievement (Muncer et al., 2022; Thi-nga, 2024). These developments indicate that mathematical metacognition has become an increasingly important area of inquiry within mathematics learning research.

Accordingly, understanding how metacognitive ability influences students' learning experience has become a major concern in contemporary mathematics learning. Within contemporary mathematics learning, metacognitive ability has been widely recognized as an important factor associated with mathematical problem solving and learning achievement (Daher, 2022; Muncer et al., 2022). However, many students continue to demonstrate limited metacognitive awareness during mathematical problem solving activities, particularly when tasks require deliberate monitoring and strategic regulation of thinking (Muncer et al., 2022). As interest in mathematical metacognition continues to grow, the development and validation of assessment instruments have become increasingly important for supporting both research and educational practice (Mahamde et al., 2021; Thi-nga, 2024). Therefore, these studies indicate that student success in mathematics is influenced not only by what they know, but also by how they manage their thinking throughout the learning process.

However, effective learning in mathematics is shaped not only by student metacognitive awareness but also by their capacity to control and maintain their own learning patterns over time. The concept of self-regulated learning (SRL) conceptualized as a proactive, cyclical process through which learners actively to maximize the efficiency of their academic learning by keeping an eye on, controlling, and regulating their behavioral, motivational, and cognitive processes (Gaeta et al., 2024). This perspective highlights that effective learning depends not only on students' intellectual abilities but also on their capacity to manage and direct their own learning processes. The capacity of SRL is widely regarded as an important contributor to academic success, as it enables students to develop adaptive learning strategies, regulate their motivation, and persist when facing academic challenges (Lim et al., 2022). These characteristics indicate that students who are able to regulate their learning tend to be better prepared to respond to academic demands and learning difficulties.

Accordingly, growing attention has been directed toward understanding how SRL operates across different educational contexts and influences students' learning outcomes. Within contemporary educational research, SRL has been widely recognized as an important factor associated with student academic performance and learning effectiveness across various educational settings (Lim et al., 2022; Panadero, 2017). Nevertheless, students do not always demonstrate effective self-regulatory behaviors, particularly in sustaining motivation and managing their learning independently when facing academic challenges ((Lim et al., 2022; Panadero, 2017). Additionally, students who possess good self-regulation skills are more likely to use efficient learning techniques, maintain their motivation, and modify their learning habits in response to difficulties and demands in the classroom (Lim et al., 2022; Panadero, 2017). This suggests that SRL contributes not only to academic achievement but also to learner ability to engage actively and persistently throughout the learning process.

Given the critical role of metacognition and self-regulated learning (SRL) in mathematics learning, a rigorous and formal examination of the data processing procedures is deemed necessary. Given the importance of accurately understanding students' metacognitive and self-regulatory characteristics, the availability of sound measurement approaches becomes essential to ensure that these constructs are assessed appropriately. The Rasch model is a psychometric measurement approach that examines the interaction between respondents and assessment items to evaluate the quality of measurement instruments through indicators such as summary, rating scale, right map.item fit, person fit, and reliability (Ayu Faradillah, 2022). Hence, the Rasch model has been employed in educational research for instrument development and validation in various mathematics-related learning contexts (Yunika et al., 2021). Its widespread application reflects the increasing demand for rigorous and evidence measurement practices in educational assessment.

The use of Rasch analysis in educational research has expanded considerably, particularly in studies focused on instrument validation and psychometric evaluation (Febryana et al., 2024). This trend reflects the increasing need for assessment instruments that can provide dependable evidence about the construct being measured. Despite this expansion, concerns remain regarding the extent to which educational instruments consistently demonstrate satisfactory psychometric properties and provide dependable measurement evidence (Febryana et al., 2024; Yunika et al., 2021). Studies employing Rasch-based validation in education have illustrated its capacity to evaluate item fit, reliability, and differential item functioning, thereby providing valuable evidence regarding the quality of educational measurement instruments (Febryana et al., 2024; Widyaningrum et al., 2024). These findings demonstrate the contribution of Rasch analysis to strengthening the psychometric quality of educational measurement instruments. Nevertheless, the existing literature has largely focused on examining mathematical metacognition and SRL as independent constructs, while studies investigating the validation of instruments that incorporate both constructs within a single assessment framework remain relatively limited.

Despite these methodological advances, several important issues remain insufficiently addressed in the existing literature. Previous research has provided substantial evidence regarding the importance of mathematical metacognition and SRL in supporting students' learning achievement and problem-solving performance (Lim et al., 2022; Panadero, 2017). Similarly, Rasch analysis has been widely used to assess educational instruments' psychometric quality in a variety of domains and learning environments (Yunika et al., 2021). Nevertheless, the existing literature has largely focused on examining mathematical metacognition and SRL as independent constructs, while studies investigating the validation of instruments that incorporate both constructs within a single assessment framework remain relatively limited.

The gap of this study is that although several studies have reported the successful application of Rasch measurement models for instrument validation, evidence concerning the psychometric evaluation of instruments specifically designed to assess mathematical metacognition alongside SRL remains limited (Ayu Faradillah, 2022; Gaeta et al., 2024). Consequently, further investigation is needed to establish whether such instruments demonstrate satisfactory measurement properties and can provide reliable information about students' cognitive and self-regulatory processes in mathematics learning. The novelty of this study is that it presents the psychometric validation of an integrated mathematical metacognition and SRL instrument using Rasch model analysis, providing empirical evidence of its validity, reliability, and measurement fit as a practically applicable assessment tool in mathematics learning (Eda Yılmazer, 2024; Ramadhani et al., 2024)

Given the need for further psychometric evidence in this area, the present study aims to validate an assessment instrument designed to measure students' mathematical metacognition and self-regulated learning using the Rasch measurement model. In particular, this study looks at the instrument's psychometric quality using metrics like item fit, person fit, reliability, and differential item functioning to see if it offers adequate measurement evidence for evaluating the intended constructs.

METHOD

Research design, site, and participants

The study employed a quantitative design using a survey approach. A survey approach is a research method that systematically collects data from a group of participants through structured instruments such as questionnaires, enabling researchers to describe characteristics, attitudes, or behaviour of a target population in a standardized and quantifiable manner (Zimba, 2023). This design was considered appropriate given that the primary objective was to examine the psychometric properties of a mathematics metacognitive and self-regulation learning instrument across a diverse group of participants. The Rasch measurement model was applied as the main analytical framework.

Compared to classical test theory approaches, the Rasch model produces person ability and item difficulty estimates that are invariant across different sample characteristics, thereby enabling a more objective and generalizable evaluation of instrument quality (Widhiarso, 2013).

The study was conducted at four schools located in East Jakarta. These schools were selected based on accessibility and the suitability of the target population's characteristics. Purposive sampling was used in the recruitment of participants. Researchers can concentrate on people who have the most instructive and suitable qualities for the study by using purposeful sampling, a non-probability sampling technique in which participants are chosen based on particular, predefined criteria that are directly related to the research objectives (Mumtaz Ali Memona, Ramayah Thurasamyd, Hiram Tingf, 2025). This restriction was based on the consideration that students at these grade levels are in a developmental stage where metacognitive skills and self-regulated learning can be meaningfully observed and measured (Setyaningsih & Rahmawati, 2022).

The complete distribution of participants by gender, grade level, and school type is presented in Figure 1

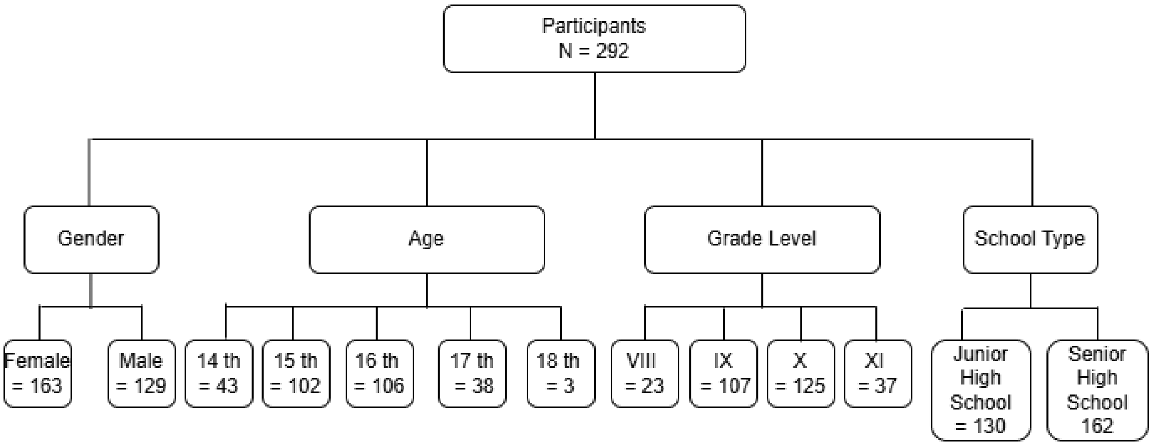

Figure 1. Demographic tabulation of participants

Figure 1 reveals that the ratio of female students slightly outnumbering their male counterparts. Regarding age, the majority of participants were 16 years old, accounting for 109 out of 292 students. Regarding grade level, grade X had the highest percentage of the total subjects at 42.8%. In terms of school type, senior high school students accounted for the majority of participants.

Instrument

This study used two data collection instruments, namely self-regulation questionnaire and mathematical metacognitive essay test. Both instruments were developed through an adaptation process drawing on internationally validated scales and rubrics, subsequently adjusted to fit the context of mathematics learning at the secondary school level in Indonesia.

The first instrument was a self-regulation questionnaire adapted from the Academic Self-Regulated Learning Questionnaire (ASLQ) developed by (Deepika Nambiar, 2022). The ASLQ was developed using Zimmerman's cycle model of self-regulated learning, which outlines three successive stages: self-reflection, performance control, and foresight. (Panadero, 2017). A five-point Likert scale is used to score each item, with 1 denoting "strongly disagree" and 5 denoting "strongly agree."

The subscale structure and descriptions are presented as follows. The first subscale is forethought consists of 10 items measuring students' preparatory activities prior to engaging in learning tasks, encompassing goal setting, strategy selection, time management, and motivational orientation. The second subscale is performance control, contains the largest number of items at 19, reflecting the complexity of the learning execution phase, which involves the application of learning strategies, distraction management, environmental regulation, and self-motivational techniques during the learning process. The third subscale is self-reflection, comprises 7 items that capture students' evaluative processes following the completion of learning or examinations, including self-evaluation, goal review, and assessment of strategy effectiveness. The distribution of items across subscales indicates that the Performance Control phase is weighted most heavily in the instrument, which aligns with its multifaceted nature as the phase in which active learning regulation occurs.

The self-regulation questionnaire comprises three subscales with a total of 36 items. The first subscale, Forethought (F), consists of 10 items. The second subscale, Performance Control (PC), contains the largest number of items at 19, reflecting the complexity of the learning execution phase. The third subscale, Self-Reflection (SR), comprises 7 items.

The essay test comprises three metacognitive indicators, each scored on a scale of 0 to 4, namely Planning (P), Monitoring (M), and Evaluating (E). The scoring rubric outlines three indicators, namely planning, monitoring, and evaluating, each scored on a scale of 0 to 4. The planning indicator assesses students' ability to organize problem-solving prior to execution, including identifying known and unknown information, constructing a mathematical model, and selecting an appropriate solution strategy. The monitoring indicator evaluates the quality and systematicity of students' problem-solving process, focusing on the clarity of calculation steps and the ability to explain each step undertaken. The evaluating indicator measures students' capacity to review their completed work, justify the correctness of their answers, and formulate conclusions aligned with the problem. Across all three indicators, the scoring criteria are designed to capture both the presence and quality of metacognitive regulation, enabling a nuanced assessment of students' metacognitive abilities in mathematics problem-solving..

Translation into Bahasa Indonesia

The questionnaire items, originally developed in English, were translated into Bahasa Indonesia to ensure that all participants could comprehend and respond to each item accurately. The translation process involved independent forward translation, reconciliation of translated versions, and back-translation to ensure conceptual equivalence across languages. Both researchers held backgrounds in mathematics education and were familiar with the Indonesian secondary school learning context.

Data analysis

The analysis was done in two steps that followed one another. A preliminary screening was carried out in the first stage to weed out responses that showed abnormal replying patterns or were incomplete. Following Rasch measurement guidelines, responses with Mean Square (MNSQ) values greater than 2.0 were considered indicative of substantial misfit and were excluded from subsequent analyses (Linacre, 2020). Data that passed the screening stage were then subjected to comprehensive psychometric analysis.

The second stage involved a series of Rasch measurement analyses conducted using Winsteps software (Linacre, 2020). The Rasch model was selected for its capacity to yield invariant person ability and item difficulty estimates that are independent of the specific sample distribution. This enables a more generalizable assessment of instrument quality compared to classical test theory approaches (Widhiarso, 2013). The analyses performed included: (a) unidimensionality, examined through Rasch Principal Component Analysis (PCA) of residuals; (b) item and person reliability and separation indices; (c) rating scale effectiveness; (d) item and person mapping via the Wright Map; and (e) item bias detection through Differential Item Functioning (DIF) analysis based on gender, grade level, and school type (Ningsih et al., 2021).

Unidimensionality was evaluated using the percentage of raw variance explained by the Rasch measures and the PCA of residuals. Higher percentages of variance explained indicate stronger evidence that the instrument measures a single underlying construct, while lower unexplained variance in the residual contrasts supports unidimensionality (Widhiarso, 2013). Item and person reliability were assessed using reliability and separation indices, which indicate the consistency of measurement and the ability of the instrument to distinguish among different levels of respondent ability and item difficulty (Widhiarso, 2013). Item fit was evaluated using infit and outfit MNSQ statistics, with acceptable values generally ranging between 0.5 and 1.5 (Mulyono et al., 2023). Rating scale effectiveness was examined through category functioning, including category frequencies, threshold ordering, and average measures across response categories (Linacre, 2020). Finally, DIF analysis was conducted to identify potential item bias across demographic groups, including gender, grade level, and school type (Mulyono et al., 2023).

RESULT AND DISCUSSION

The Analysis of the Unidimensionality of the Items

A core assumption examined through Rasch analysis is unidimensionality, which statistically evaluates whether every item within the instrument measures one and the same underlying construct, free from dimensional interference (Khotimah et al., 2022). Unidimensionality was evaluated using the percentage of raw variance explained by the Rasch measures, with values above 20% indicating acceptable evidence of a unidimensional construct. The overall scale demonstrated a raw variance explained by measures of 66.9%, while the Evaluating, Monitoring, and Planning subscales accounted for 66.9%, 50.8%, and 55.3% of the variance, respectively. The eigenvalues of the first contrast were 1.90 logits for the Evaluating subscale, 1.78 logits for the Monitoring subscale, and 1.66 logits for the Planning subscale. These results indicate that all three subscales satisfy the criterion for unidimensionality, as the variance explained by the Rasch measures substantially exceeded the minimum recommended threshold and the first contrast eigenvalues remained below 2.0 logits. The Evaluating, Monitoring, and Planning subscales may therefore be interpreted as converging upon a single dominant construct, providing evidence of unidimensionality across the instrument's core dimensions.

Table 1. Summary of Rasch Measurement Model for the Mathematics Metacognitive and Self-Regulated Learning

Mathematical Metacognitive				
Parameter (with quality criteria)	**Global Essay Scale (12 items)**	**P (4 items)**	**M (4 items)**	**E (4 items)**
Model fit: Summary of items				
Item mean in logit (criteria 0.0 logits)	.00, SD = 0.36	.00, SD = 0.40	.00, SD = 0.59	.00, SD = 0.17
Item reliability	0.99	0.98	0.99	0.96
Item separation reliability (criteria: good, 0.81-0.90; very good, 0.91-0.94; excellent, >0.94)	0.99 (Excellent)	0.98	0.99	0.96
Item model fit MNSQ range extremes (criteria: good, 0.5-1.5)	Infit: 0.44 - 2.03	Infit: 0.75 - 1.16	Infit: 0.52 - 1.38	Infit: 0.69 - 1.44
	Outfit: 0.43 - 2.20	Outfit: 0.67 - 1.13	Outfit: 0.51 - 1.36	Outfit: 0.62 - 1.35
Item separation index (criteria >3)	8.89	6.75	12.72	4.77
Separate item strata == [(4 \times \text{separation index}) + 1] / 3 (criteria: fair, 2-3; good, 3-4; very good, 4-5; excellent, >5)	12.18 (12 levels)	9.33 (9 levels)	17.29 (17 levels)	6.69 (7 levels)
Model fit: Summary of persons				
Person means in logits (criteria: 0.0 logits)	-0.65, SD = 0.70	-2.00, SD = 1.15	-0.67, SD = 0.74	-0.28, SD = 0.49
Person reliability	0.78	0.66	0.61	0.50
Person separation reliability (criteria: fair, 0.67-0.80; good, 0.81-0.90)	0.78 (fair)	0.66	0.61	0.50
Person separation index (criteria >2)	1.89	1.38	1.26	1.00
Separate Person strata = [(4 \times \text{separation index}) + 1] / 3 (criteria: fair, 2-3; good, 3-4; very good, 4-5; excellent, >5)	2.85 (3 levels)	2.17 (2 levels)	2.01 (2 levels)	1.67 (2 levels)
Cronbach Alpha (KR-20) Person Raw Score "Test" Reliability	0.89 (Very Good)	0.70	0.60	0.66
Dimensionality				
Raw variance in data explained by measure	46.8%	54.0%	48.6%	40.4%
PCA eigenvalue for first contrast	2.12 logits	1.77 logits	1.94 logits	1.81 logits

Self-Regulated Learning				
Parameter (with quality criteria)	**Global Essay Scale (36 items)**	**F (10 items)**	**PC (19 items)**	**SR (7 items)**
Model fit: Summary of items				
Item mean in logit (criteria 0.0 logits)	.00, SD = 0.67	.00, SD = 0.81	.00, SD = 0.72	.00, SD = 0.17
Item reliability	0.98	0.99	0.99	0.60
Item separation reliability (criteria: good, 0.81-0.90; very good, 0.91-0.94; excellent, >0.94)	0.98 (Excellent)	0.99	0.99	0.60
Item model fit MNSQ range extremes (criteria: good, 0.5-1.5)	Infit: 0.62 - 2.70	Infit: 0.64 - 2.15	Infit: 0.67 - 2.64	Infit: 0.84 - 1.28
	Outfit: 0.61 - 2.82	Outfit: 0.63 - 2.11	Outfit: 0.65 - 2.77	Outfit: 0.86 - 1.30
Item separation index (criteria >3)	7.83	8.87	8.29	1.21
Separate item strata == [(4 \times \text{separation index}) + 1] / 3 (criteria: fair, 2-3; good, 3-4; very good, 4-5; excellent, >5)	10.77 (11 levels)	12.16 (12 levels)	11.39 (11 levels)	1.95 (2 levels)
Model fit: Summary of persons				
Person means in logits (criteria: 0.0 logits)	1.46, SD = 0.91	1.39, SD = 0.93	1.45, SD = 0.90	1.87, SD = 1.29
Person reliability	0.90	0.69	0.81	0.67
Person separation reliability (criteria: fair, 0.67-0.80; good, 0.81-0.90)	0.90 (good)	0.69 (Fair)	0.81 (Good)	0.67 (Fair)
Person separation index (criteria >2)	2.97	1.48	2.05	1.44
Separate Person strata = [(4 \times \text{separation index}) + 1] / 3 (criteria: fair, 2-3; good, 3-4; very good, 4-5; excellent, >5)	4.29 (4 levels)	2.31 (2 levels)	3.07 (3 levels)	2.25 (2 levels)
Cronbach Alpha (KR-20) Person Raw Score "Test" Reliability	0.93 (Excellent)	0.75 (Good)	0.86 (Good)	0.81 (Good)
Dimensionality				
Raw variance in data explained by measure	40.1%	42.3%	40.1%	38.8%
PCA eigenvalue for first contrast	2.59 logits	1.64 logits	2.59 logits	1.68 logits

Item and Person Separation Reliability

As presented in Table 3, the mathematics item separation reliability the global scale of metacognition was .99 (excellent), with subscales ranging from .96 (E) to .99 (M), all exceeding the threshold of .90, indicating that the items consistently reproduced stable difficulty ordering across comparable samples (Sofroniou & Judijanto, 2024). The item separation index for the global scale was 8.89, well above the minimum criterion of 3.0, demonstrating that the instrument effectively distinguishes items of varying difficulty levels. Regarding person separation reliability, the global scale yielded a value of .78 (fair), supported by a Cronbach Alpha of .89 (very good). At the subscale level, person reliability ranged from .50 (E) to .66 (P), indicating that while the instrument demonstrates acceptable internal consistency at the global level, the subscales show limited capacity to differentiate individual respondents into distinct ability groups. A similar pattern of reliability was further examined in the Self-Regulated Learning instrument.

For the Self-Regulated Learning instrument, reliability of item separation for mathematics Metacognitive worldwide scale was .98 (excellent), with subscales F (.99) and PC (.99) also reaching excellent levels, while SR (.60) fell below the acceptable threshold, suggesting that items in the SR subscale require further refinement. The item separation index for the global scale was 7.83, with subscales F (8.87) and PC (8.29) also surpassing the minimum criterion of 3.0; however, SR yielded a notably low index of 1.21, confirming its limited discriminative capacity. Reliability of person separation on a worldwide scale was .90 (good), with a Cronbach Alpha of .93 (excellent), indicating the instrument reliably differentiates respondents at the global level. At the subscale level, PC demonstrated good reliability (.81), while F (.69) and SR (.67) were categorized as fair, suggesting these subscales are more suitable for group-level rather than individual-level measurement decisions.

Effectiveness of the Rating Scale

The rating scale's efficacy was assessed in order to ascertain whether participants could discriminate between the response categories in each instrument. This study included two instruments: an essay test with a 9-point scoring rubric ranging from 0 to 8 and a questionnaire with a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). There has to be more than ten observed individuals in each response category. Furthermore, the overall MNSQ value must be less than 2.0 logits, and the adjacent threshold distance must be between 1.4 and 5 logits. Table 4 present the summary of category structure for both instruments.

Table 2. Summary of Category Structure of Rating Scales Mathematics Metacognitive and Self-Regulated Learning

Rating scale step number	Observed person	Average calibration	Outfit MNSQ	Threshold	Threshold distance
Mathematics Metacognitive					
1	135 (1%)	0.31	2.01	NONE	**(-3.11)**
2	494 (5%)	0.34	1.26	-1.67	**1.50**
3	2948 (28%)	0.71	0.92	-1.32	**0.35**
4	3720 (35%)	1.45	0.85	0.95	**2.27**
5	3214 (31%)	2.41	0.90	2.04	**1.09**
Self-Regulated Learning					
0	1264 (34%)	-1.15	1.06	NONE	**(-1.88)**
1	435 (12%)	-0.85	0.65	0.04	**0.88**
2	539 (15%)	-0.55	0.66	-0.95	**-0.99**
3	327 (9%)	-0.43	1.31	-0.01	**0.94**
4	658 (18%)	-0.20	1.06	-0.99	**-0.98**
5	82 (2%)	-0.02	1.26	1.98	**2.97**
6	97 (3%)	0.19	0.83	-0.08	**-2.06**
7	104 (3%)	0.25	1.34	0.19	**0.27**
8	190 (5%)	0.50	1.07	-0.18	**(-1.63)**

As shown in Table 4, the questionnaire rating scale demonstrated that all five categories met the minimum threshold of 10 observed persons per category. The outfit MNSQ values for categories 3 through 5 were below 2.0 logits, indicating adequate fit; however, category 1 (outfit MNSQ = 2.01) slightly exceeded the acceptable threshold, suggesting that respondents had some difficulty distinguishing the lowest category. The average calibration increased monotonically from 0.31 to 2.41, indicating that higher categories were consistently selected by persons with higher ability measures. Regarding threshold distance, categories 4 and 5 met the ideal range of 1.4–5.0 logits (2.27 and 1.09 logits respectively), while categories 2 and 3 fell below the ideal threshold distance, suggesting that participants may have had difficulty discriminating between adjacent lower categories.

With regard to the essay test rating scale (0–8), the results presented a more complex pattern. All categories met the minimum of 10 observed persons. The outfit MNSQ values across all categories remained below 2.0 logits, reflecting adequate fit. However, several threshold values did not increase monotonically, particularly at categories 2, 4, 6, and 8, indicating disordered thresholds. This suggests that respondents experienced difficulty distinguishing between adjacent score levels, and some scoring categories may not be functioning as distinctly intended.

Item and Person Mapping

The Wright Map analysis was used to match the test or survey variables' degree of difficulty (item difficulty) on a single logit scale with the respondents' degree of proficiency or agreement (person measure) (Mulyono et al., 2023). This approach enables a simultaneous visual comparison between person ability and item difficulty on a single measurement continuum, facilitating a more comprehensive interpretation of how well the instrument targets the respondent sample. Based on the hierarchical structure of the Wright Map, the distribution of these parameters is visually separated by a vertical dotted line; the left-hand side illustrates the location density of the respondents, while the right-hand side represents the latent difficulty parameters of the items (Mulyono et al., 2023). Within the framework of the Rasch measurement model, items situated at the highest logit level are identified as the most challenging variables for respondents to solve or endorse, whereas variables positioned at the bottom baseline represent the easiest elements to master or affirm (Mulyono et al., 2023). Accordingly, the relative positioning of items and persons on the Wright Map provides direct empirical evidence regarding the degree of alignment between instrument difficulty and respondent ability, which is essential for evaluating the overall targeting quality of the assessment. In this study, the evaluation was comprehensively executed across two distinct measuring instruments: an open-ended essay examination (u) and a Likert-scale questionnaire (z).

The empirical mapping of the open-ended essay instrument (u), as displayed in Figure 1, demonstrates a well-distributed range of item difficulty parameters spanning across different logit thresholds

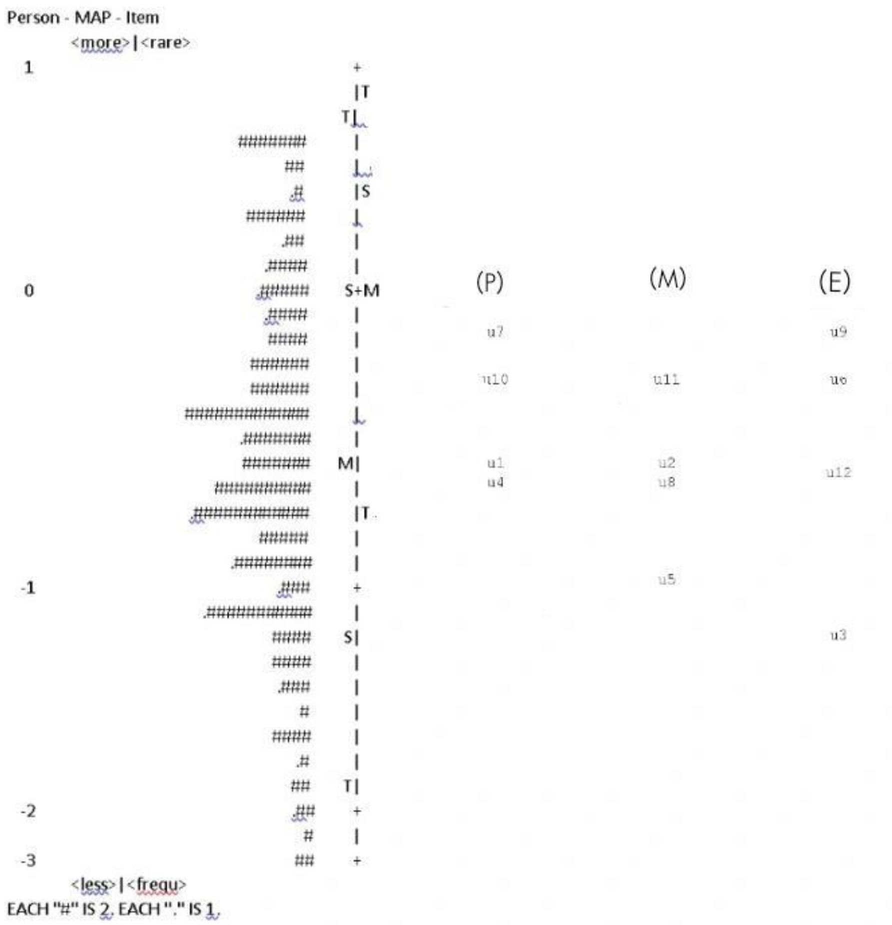

Figure 1. **Wright Map of the Mathematical Metacognitive Instrument**

As observed in Figure 1, item u7 and item u9 occupy the absolute apex of the hierarchy at the +0.5 logit level. This vertical standing marks them as the most difficult tasks within the essay assessment, which were rarely completed successfully by the standard student sample. The elevated position of u7 and u9 suggests that the participants faced profound conceptual or procedural boundaries when attempting to address the specific performance indicators evaluated in these two prompts. Positioned slightly beneath this highest tier, items u10, u11, and u6 represent a moderate-to-high difficulty bracket localized at the +0.3 logit mark. Meanwhile, the dense cluster of items concentrated precisely at the item mean line (0.0 logit) contains u1, u2, u12, u4, and u8. The accumulation of these five items at the baseline mean indicates that their underlying difficulty parameters are perfectly calibrated to match the average cognitive capacity of the student population under evaluation. Conversely, the essay variables extending into the negative spectrum of the map consist of item u5 at -0.3 logit and item u3 positioned at the absolute base of -1.0 logit. The positioning of u3 at the lowest logit tier confirms that this prompt represents the least demanding task, which nearly all participants possessed the requisite academic capacity to answer accurately without significant constraints.

In terms of the self-reported survey data, the psychological positioning of the questionnaire items (z), as illustrated in Figure 2, reveals a distinct structural distribution of 36 Likert-scale items stratified across four primary logit clusters.

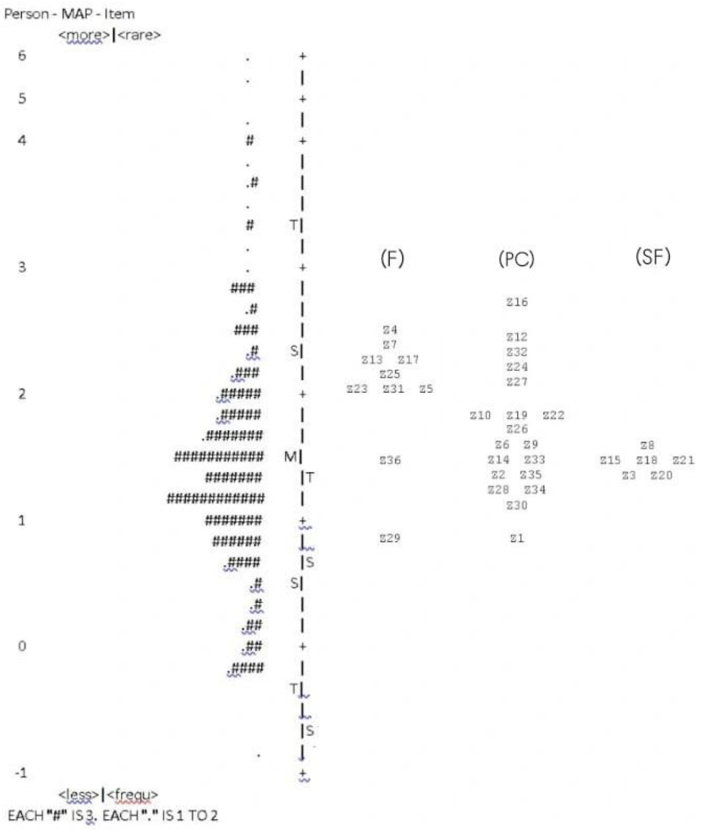

Figure 2. Wright Map of the SLR Questionnaire

The Wright Map distribution of the Self-Regulated Learning questionnaire instrument (z), as illustrated in Figure 2, reveals a well-structured spread of item difficulty estimates positioned across varying logit thresholds. At the uppermost tier of the hierarchy, item Z16 from the Performance Control subscale is positioned at the +2 logit level, identifying it as the most challenging item for respondents to affirm, as it captures habitual note-taking behavior demanding consistent self-disciplined practice. Immediately beneath this apex, a secondary grouping of items is concentrated within the +1 logit range, encompassing Forethought items Z4, Z7, Z13, Z17, Z23, Z25, and Z31, which reflect motivational orientation toward challenging content, pre-study scheduling, target-setting, systematic material organization, content outlining, difficulty-based time allocation, and daily study goal establishment, alongside Performance Control items Z12, Z24, Z27, and Z32, which capture peer-assisted learning strategies, selective concentration on demanding material, social support utilization for missed lessons, and the use of visual organizational tools during study sessions. Collectively, these items capture sophisticated planning and strategic learning behaviors, serving as a psychometric threshold that effectively differentiates respondents whose self-regulatory capacity places them moderately above the general student population mean.

The vast majority of survey indicators are concentrated around the mean value of 0.0 logit, encompassing Performance Control items Z2, Z6, Z9, Z10, Z14, Z19, Z22, Z26, Z28, Z30, Z33, Z34, and Z35, which reflect behaviors associated with active reading review, careful material comprehension, enjoyable study habits, keyword utilization, distraction elimination, note simplification, concept memorization, task completion, assignment organization, multi-source information use, help-seeking, conceptual understanding, and class note reference, alongside Self-Reflection items Z3, Z8, Z15, Z18, Z20, and Z21, which capture study plan adjustment, strategy reinforcement, progress monitoring, post-examination reflection, error-based learning, and effort self-assessment, as well as Forethought item Z36, which reflects attendance regularity commitment. At the lowest segment of -1 logit, items Z29 ("I motivate myself to do better than before") and Z1 ("I study in a suitable place where I can concentrate") represent the easiest statements for respondents to endorse, in stark contrast to Z16 at the apex, suggesting that while students broadly affirm foundational behaviors related to self-motivation and study environment, they find it considerably more challenging to consistently practice advanced self-regulatory strategies such as systematic note-taking, goal-setting, and strategic time allocation. In conclusion, the majority of participants exhibit fundamental self-regulation behaviors, while higher-order planning and performance control behaviors remain more demanding for students to consistently endorse.

Item Bias Distributions

Analysis using Person Differential Item Functioning (DIF) was carried out to systematically discover whether any items on the essay instrument contain bias or unfairly benefit a specific demographic subgroup of respondents. The rigorous psychometric threshold for identifying item bias dictates that a flagged item must exhibit a DIF contrast value greater than 0.50 logits accompanied by a statistically significant probability value ($p < 0.05$).

Table 3. Item Bias for the Mathematical Metacognitive and Self-Regulated Learning

Mathematical Metacognitive

No	Item	Status	DIF Measure	DIF Contrast	t	Probability
1	u2	H	.00	.53	2.40	.0350
		I	.04	.56	2.54	.0259
		A	-.16	.36	1.65	.1265
		J	-.52	-.53	-2.40	.0350
2	u7	H	.23	-.23	-2.22	.0275
		I	.46	.08	.67	.5049
		A	.39	.20	1.50	.1349
		J	.19	.04	.14	.8908
3	u10	H	.23	-.29	-2.80	.0056
		I	.52	-.34	-3.12	.0021
		A	.57	.34	3.12	.0021
		J	.76	-.53	-1.18	.2695
4	u11	H	-.31	.25	3.01	.0030

Mathematical Metacognitive

No	Item	Status	DIF Measure	DIF Contrast	t	Probability
		I	-.56	-.37	-4.25	.0000
		A	-.19	.12	1.47	.1423
		J	-.14	-.18	-.68	.5115

Self-Regulated Learning

No	Item	Status	DIF Measure	DIF Contrast	t	Probability
1	Z4	H	1.96	.39	2.16	.0324
		I	1.95	.38	2.08	.0387
		A	1.57	.20	1.12	.2630
2	Z9	H	-.44	-.42	-2.10	.0375
		I	-.23	-.20	-1.09	.2782
		A	-.02	-.42	-2.10	.0375
3	Z11	H	.23	.45	2.32	.0214
		I	.22	.44	2.25	.0257
		A	-.22	.22	1.18	.2400
4	Z12	H	1.18	.53	2.82	.0053
		I	.99	.33	1.76	.0793
		A	.65	.53	2.82	.0053
5	Z14	H	-.09	.42	2.10	.0373
		I	-.51	.23	1.17	.2421
		A	-.33	.18	.89	.3722
6	Z15	H	-.02	.43	2.17	.0312
		I	-.45	-.45	-2.23	.0272
		A	.00	-.02	-.11	.9165
7	Z16	H	2.24	.66	3.70	.0003
		I	2.53	.96	5.34	.0000
		A	1.57	.66	3.70	.0003
8	Z20	H	-.24	.25	1.23	.2202
		I	-.49	-.41	-2.02	.0447
		A	-.08	-.16	-.83	.4091
9	Z21	H	-.42	-.44	-2.23	.0269
		I	.02	.10	.50	.6183

Self-Regulated Learning

No	Item	Status	DIF Measure	DIF Contrast	t	Probability
		A	-.08	-.34	-1.69	.0922
10	Z23	H	.45	.61	3.20	.0016
		I	-.17	.48	2.49	.0136
		A	-.04	.30	1.58	.1150
11	Z24	H	1.00	.64	3.39	.0009
		I	.78	.43	2.25	.0257
		A	.35	.64	3.39	.0009
12	Z25	H	.48	.46	2.37	.0188
		I	.26	.24	1.24	.2167
		A	.02	.18	.92	.3580
13	Z28	H	-.80	-.55	-2.66	.0084
		I	-.82	-.58	-2.74	.0067
		A	-.24	-.55	-2.66	.0084
14	Z29	H	-1.51	-1.08	-4.65	.0000
		I	-1.32	-.89	-3.91	.0001
		A	-.43	-1.08	-4.65	.0000
15	Z30	H	-.93	-.61	-2.87	.0045
		I	-1.26	-.93	-4.18	.0000
		A	-.33	-.61	-2.87	.0045
16	Z31	H	.05	-.42	-2.20	.0291
		I	.47	.35	1.81	.0725
		A	.12	-.07	-.34	.7322
17	Z32	H	.81	.40	2.10	.0375
		I	1.09	.68	3.58	.0004
		A	.41	.40	2.10	.0375
18	Z36	H	-1.27	-.88	-3.96	.0001
		I	-.80	-.47	-2.08	.0389
		A	-.39	-.88	-3.96	.0001

As shown in Table 3, the DIF analysis of the essay instrument identified three items exhibiting statistically significant bias across respondent groups. Item u7 demonstrated a notable difference between groups H and I ($t = -2.22$, $p = .0275$), suggesting that group I found this item comparatively more difficult than group H. Item u10 revealed bias in two separate comparisons, between H and I ($t = -2.80$, $p = .0056$) and between H and J ($t = -3.12$, $p = .0021$), whereby the progressively larger absolute t-values indicate increasingly stronger differential functioning across groups. Item u11 similarly exhibited significant bias in two comparisons, between H and I (t = 3.01, p = .0030) and between I and A (t = -4.25, p = .0000), with the latter representing the most substantial group difference detected within the essay instrument. Across all flagged items, a consistent inverse relationship between the absolute t-value and the probability value was observed, whereby larger t-values corresponded to smaller probability values, thereby providing stronger statistical evidence that the items functioned differently across the compared respondent groups (Linacre, 2020).

Turning to the questionnaire instrument, nine unique items demonstrated statistically significant DIF, revealing a more complex pattern of differential item functioning. Item Z16 recorded the highest t-value across both instruments (t = 5.34, p = .0000) in the comparison between groups I and J, establishing it as the most severely biased item in the entire study. Item Z29 similarly exhibited highly significant bias in comparisons between H and J (t = -4.65, p = .0000) and I and J (t = -3.91, p = .0001), indicating that group J consistently interpreted this item differently from the other respondent groups. In contrast, items such as Z12 (t = 2.82, p = .0053) and Z28 (t = -2.66, p = .0084) reflected more moderate levels of differential functioning, with t-values that, while statistically significant, were comparatively smaller in magnitude. These findings collectively suggest that respondents' occupational background and teaching experience meaningfully shape how certain self-regulated learning items are perceived and interpreted, a pattern consistent with prior evidence that professional context and demographic background influence how individuals respond to psychometric instruments (Chan & Subramaniam, 2020; Mulyono et al., 2020).

CONCLUSION

` The present study aimed to validate an integrated assessment tool for assessing pupils' mathematics metacognition and self-regulated learning (SRL) using the Rasch measurement model, with specific focus on unidimensionality, item and person reliability, rating scale effectiveness, Wright Map analysis, and Differential Item Functioning (DIF). The findings demonstrate that both the Mathematics Metacognitive Essay Assessment and the Self-Regulated Learning Questionnaire possess satisfactory psychometric properties. Both instruments confirmed unidimensionality, with raw variance explained by Rasch measures exceeding the minimum recommended threshold of 20% (Mathematics Metacognitive: 46.8%; SRL: 40.1%), and eigenvalues of the first contrast remaining below or near 2.0 logits across all subscales. Item separation reliability was classified as excellent at the global level for both instruments (.99 and .98, respectively), and internal consistency was further supported by very good to excellent Cronbach Alpha values (.89 and .93). Person reliability ranged from fair to good, indicating sufficient capacity to differentiate respondents along the ability continuum, although the Self-Reflection subscale of the SRL instrument demonstrated lower indices, suggesting the need for targeted item revision. Wright Map analysis revealed that items covered a broad difficulty range, with the majority clustered around the mean logit, indicating strong sensitivity for assessing respondents of average metacognitive and self-regulatory ability. Rating scale analysis showed that the five-point Likert scale functioned adequately, whereas the essay test's nine-point rubric exhibited disordered thresholds at mid-range categories, suggesting the scoring criteria require recalibration. DIF analysis identified three items in the essay instrument (u7, u10, u11) and nine items in the questionnaire as exhibiting statistically significant differential functioning across groups, with items Z16 and Z29 showing the most substantial bias, indicating that certain items may be perceived differently depending on respondents' school type or grade level. Overall, these findings confirm that the integrated instrument demonstrates sound validity, reliability, and measurement fit as a practically applicable assessment tool in mathematics learning contexts; however, targeted revision of flagged items and expansion of the participant sample beyond East Jakarta schools are recommended to further strengthen measurement quality and generalizability in future research..

ACKNOWLEDGEMENT

The authors would like to sincerely thank all of the East Jakarta schools that allowed and assisted with the data collection process during this study. Deepest appreciation is also extended to the validators who generously dedicated their time and expertise in providing thorough evaluation and constructive feedback on the research instruments. The authors further wish to thank Universitas Muhammadiyah Prof. DR. HAMKA (UHAMKA) for the academic support and facilities provided during the completion of this research. This article was produced as a course output for the Evaluation of Learning Process and Outcomes (EPHB) subject, which has provided the theoretical and methodological foundation for the research team in designing, conducting, and reporting this instrument validation study.

REFERENCE

Ayu Faradillah, C. S. (2022). *MATHEMATICAL RESILIENCE: VALIDITY AND RELIABILITY WITH RASCH MODEL AND SPSS IN SENIOR HIGH SCHOOL*. *11*(4), 3545–3555.

Daher, W. (2022). *Mathematics Teachers ' Encouragement of Their Students ' Metacognitive Processes*. 1272–1284.

Deepika Nambiar. (2022). *Development and Validation of Academic Self-regulated Learning Questionnaire (ASLQ)*.

Eda Yılmazer, Z. H. 1 and F. T. (2024). *Psychometric properties and validation of the metacognitive self-assessment scale (MSAS) in a Turkish sample*. *December*. https://doi.org/10.3389/fpsyg.2024.1411733

Febryana, U. I., Rahmania, F., Faradillah, A., & Jusra, H. (2024). *Evaluation of Validity and Reliability Using Rasch Model and SPSS on Students ' Mathematical Creative Thinking Ability*. *5*(1), 1–10.

Gaeta, M. L., Mella-norambuena, J., & Awidi, I. T. (2024). *Validation of the self-regulation of learning instrument for undergraduates*. *November*, 1–10. https://doi.org/10.3389/feduc.2024.1464424

Khotimah, H., Aini, L. N., Hariyati, T., Indah, N., Balikpapan, U., Pupuk, J., Balikpapan, R., Balikpapan, U., Pupuk, J., Balikpapan, R., Balikpapan, U., Pupuk, J., & Balikpapan, R. (2022). *PROFILE OF SELF-REGULATED LEARNING OF STUDENTS ON ONLINE MATHEMATICS LEARNING PROFIL SELF-REGULATED LEARNING SISWA PADA*. *7*, 229–235.

Lim, L., Lim, S. H., Ying, W., & Lim, R. (2022). *behavioral sciences A Rasch Analysis of Students ' Academic Motivation toward Mathematics in an Adaptive Learning System*.

Linacre, J. M. (2020). *Winsteps Help for Rasch Analysis 4.5.2*.

Mahamde, R. D., Darmawan, F. A., & Malago, J. D. (2021). *Metacognitive skill assessment model through the blended learning management system in vocational education*. *11*(1), 1–13.

Mulyono, H., Ningsih, S. R. I. K., & Sholeha, B. (2023). *Validating the Academic Writing Creativity and Self-Efficacy Scale : A Rasch Model Analysis*. *7*(1), 121–135.

Mumtaz Ali Memona, Ramayah Thurasamyd, Hiram Tingf, and J.-H. C. (2025). *PURPOSIVE SAMPLING : A REVIEW AND GUIDELINES FOR*. *9*(1), 1–23. https://doi.org/10.47263/JASEM.9(1)01

Muncer, G., Higham, P. A., & Gosling, C. J. (2022). *A Meta-Analysis Investigating the Association Between Metacognition and Math Performance in Adolescence* (Vol. 2012). Educational Psychology Review.

Ningsih, S. K., Mulyono, H., Rahmah, R. A., & Aprilia, N. (2021). A Rasch-based validation of EFL teachers ' received online social support scale A Rasch-based validation of EFL teachers ' received online social support scale. *Cogent Education*, *8*(1). https://doi.org/10.1080/2331186X.2021.1957529

Panadero, E. (2017). *A Review of Self-regulated Learning : Six Models and Four Directions for Research*. *8*(April), 1–28. https://doi.org/10.3389/fpsyg.2017.00422

Ramadhani, R., Syahputra, E., & Simamora, E. (2024). *The construct validity of self-regulated learning questionnaire for senior high school students : a Rasch model analysis*. *13*(3), 2032–2038. https://doi.org/10.11591/ijere.v13i3.26816

Setyaningsih, N., & Rahmawati, E. Z. (2022). *Metacognitive skills of student in solving problems of two-variable linear equation systems in terms of self-regulated learning*. *9*(1), 82–92.

Sofroniou, A., & Judijanto, L. (2024). *Reliability and separation index analysis of mathematics questions integrated with the cultural architecture framework using the Rasch model*. *11*(3), 499–509. https://doi.org/10.20448/jeelr.v11i3.5861

Thi-nga, H. (2024). *Metacognition in mathematics education : From academic chronicle to future research scenario – A bibliometric analysis with the Scopus database*. *20*(4).

Widhiarso, W. (2013). *Jurnal Penelitian dan Evaluasi Pendidikan APLIKASI MODEL RASCH CAMPURAN DALAM MENGEVALUASI*. *17*(1). https://doi.org/10.21831/pep.v17i1.1367

Widyaningrum, A., Lutfiyana, R. D., Faradillah, A., & Miatun, A. (2024). *Survey of Rasch Model Analysis of Mathematical Literacy Abilities of Senior High School Students*. *5*(1), 57–66.

Yunika, V. R., Rohmah, A. N., Istiqomah, S. N., & Faradillah, A. (2021). *Validitas dan Reliabilitas Anxiety Questioner dalam Pembelajaran Matematika dengan menggunakan Rasch Models*. *2*, 161–169.

Zimba, O. (2023). *Designing , Conducting , and Reporting Survey Studies : A Primer for Researchers*. *38*(48), 1–11.