Inovasi Kurikulum
https://ejournal.upi.edu/index.php/JIK

Improving assessment quality: Development of evaluation instruments for sixth-grade Mathematics learning

Eny Cahyaningsih[1], Wardani Rahayu[2], Riyan Arthur[3]
[1,2,3] Univesitas Negeri Jakarta, Jakarta, Indonesia
cahyaningsiheny@gmail.com[1]

ABSTRACT

Valid and reliable assessment instruments are crucial for accurately measuring student competencies, yet many fail. This study aims to develop a Mathematics learning outcome assessment instrument based on the Kurikulum Merdeka for sixth-grade elementary school students, focusing on validity and reliability using the Rasch model. This instrument is designed to encompass cognitive, affective, and psychomotor competencies following the needs of the Kurikulum Merdeka. Analysis was conducted on 213 students using a quantitative approach with Winsteps software to evaluate item quality, unidimensionality, reliability, difficulty level, and potential differential item functioning (DIF). The research results indicate that the instrument is highly reliable and meets the unidimensionality criteria. The distribution of question difficulty levels varies from very easy to very difficult, reflecting the instrument's ability to measure students with a range of abilities. All items meet the fit criteria based on fit statistics (Outfit MNSQ, ZSTD, and Point Measure Correlation). However, two items (Item 1 and Item 5) show significant DIF bias based on gender analysis. This study concludes that this assessment instrument based on the Kurikulum Merdeka is valid, reliable, and suitable for assessing students' mathematical abilities.

ARTICLE INFO

Article History:
Received: 10 Sep 2024
Revised: 23 Dec 2024
Accepted: 27 Dec 2024
Available online: 5 Jan 2025
Publish: 28 Feb 2025

Keyword:
educational assessment; instrument evaluation; kurikulum merdeka; Rasch model; reliability; validity

Open access
Inovasi Kurikulum is a peer-reviewed open-access journal.

ABSTRAK

Instrumen penilaian yang valid dan reliabel sangat penting untuk mengukur kemampuan peserta didik secara objektif, namun banyak instrumen yang ada belum merepresentasikan kompetensi peserta didik dengan akurat. Penelitian ini bertujuan untuk mengembangkan instrumen penilaian hasil belajar matematika berbasis Kurikulum Merdeka untuk peserta didik kelas VI Sekolah Dasar, dengan fokus pada validitas dan reliabilitas menggunakan model Rasch. Analisis dilakukan pada 213 peserta didik menggunakan pendekatan kuantitatif dengan software Winsteps untuk mengevaluasi kualitas item, unidimensionalitas, reliabilitas, tingkat kesulitan, dan potensi bias diferensial (Differential Item Functioning/ DIF). Hasil penelitian menunjukkan bahwa instrumen memiliki reliabilitas tinggi dan memenuhi kriteria unidimensionalitas. Sebaran tingkat kesulitan soal cukup bervariasi dari sangat mudah hingga sangat sulit, mencerminkan kemampuan instrumen dalam mengukur peserta didik dengan beragam tingkat kemampuan. Semua item memenuhi kriteria kesesuaian berdasarkan statistik kecocokan (Outfit MNSQ, ZSTD, dan Point Measure Correlation), namun terdapat dua item (Item 1 dan Item 5) yang menunjukkan bias DIF signifikan berdasarkan analisis gender. Kesimpulan dari penelitian ini adalah bahwa instrumen penilaian berbasis Kurikulum Merdeka ini valid, reliabel, dan cocok untuk digunakan dalam menilai kemampuan matematika peserta didik.

Kata Kunci: *evaluasi instrumen; Kurikulum Merdeka; Model Rasch; penilaian pendidikan; reliabilitas; validitas*

How to cite (APA 7)
Cahyaningsih, E., Rahayu, W., & Arthur, R. (2025). Improving assessment quality: Development of evaluation instruments for sixth-grade Mathematics learning. *Inovasi Kurikulum, 22*(1), 161-180.

Peer review
This article has been peer-reviewed through the journal's standard double-blind peer review, where both the reviewers and authors are anonymised during review.

https://doi.org/10.17509/jik.v22i1.77242

INTRODUCTION

autonomy in designing instruction that aligns with students' needs. Assessment plays a vital role in the learning process, serving as a means for teachers to evaluate the effectiveness of their instruction. This evaluation is crucial, as it provides a comprehensive view of students' mastery of specific topics, their challenges during learning, and their standing relative to their peers (Safitri & Widyanti, 2024; Saputri *et al*., 2024). In the context of mathematics, sixth-grade students are expected to develop logical and analytical thinking skills applicable to everyday situations. Effective assessment is essential for measuring and enhancing these skills.

The study found that although the critical thinking assessment instrument used in secondary schools demonstrated high reliability (r = 0.89), approximately 25% of its items were invalid and failed to fully capture the intended construct (Juliani & Erita, 2023). Another study reported that a student self-confidence assessment tool had adequate validity and reliability, though several items required revision to improve measurement accuracy (Maulana, 2022). These findings highlight the importance of carefully designing and analyzing assessment instruments to ensure they are valid and reliable in reflecting students' abilities. Consequently, developing assessment tools aligned with the Kurikulum Merdeka is an urgent priority to enhance the quality of learning and evaluation. A well-designed assessment instrument should incorporate critical thinking, creativity, and problem-solving skills that improve learning outcomes and increase student motivation.

The Rasch model approach in developing psychometrically based assessment instruments provides a comprehensive framework for evaluating the quality of test items. This model ensures that the instruments produced are valid, reliable, and aligned with students' needs (Tarigan *et al*., 2022). Assessment is crucial to learning, measuring student achievement, and providing feedback for teachers and learners. Evaluating the validity and reliability of an instrument is vital to ensure that it accurately assesses student performance. Validity refers to the extent to which the instrument aligns with its intended measurement objectives, while reliability reflects the consistency of results obtained under comparable conditions. The development of high-quality assessment instruments, therefore, plays a key role in improving learning effectiveness and the accuracy of student outcome evaluations.

Using the Rasch model also enables in-depth evaluation of test items, such as validity and unidimensionality (Abdullaev *et al*., 2024; Latifah *et al*., 2024; Yusuf *et al*., 2021). Through this approach, assessment instruments can be tailored to students' ability levels, resulting in a more equitable and objective measurement tool. Research analyzing the quality of General Biology Mid-Semester Exam (UTS) items using the Rasch model demonstrates the model's ability to identify misfitting items and offer recommendations to improve their effectiveness in measuring student achievement (Novriyanti & Arthur, 2024).

Psychometrics provides a solid theoretical foundation for developing assessment instruments based on Item Response Theory (IRT), which is more adaptive than Classical Test Theory (CTT). Recent studies have highlighted the advantages of IRT-based instruments in identifying invalid items and offering directions for refinement. IRT facilitates detailed analysis of item characteristics—including validity, reliability, difficulty level, and discrimination index—allowing assessment tools to be optimized for accurately measuring learners' abilities. Some studies have applied the 2PL IRT model to analyze item parameters and learner ability but have limited the focus to difficulty level and discrimination index, without addressing reliability, unidimensionality, or Differential Item Functioning (DIF) bias (Jumini & Retnawati, 2022). Moreover, these studies did not involve the development of a new instrument. In contrast, this study is more comprehensive as it develops a new instrument using the Rasch model. It also includes evaluations of DIF bias, unidimensionality, and reliability, producing a more thorough analysis.

Context-based assessment, which incorporates questions related to everyday life, has effectively enhanced students' motivation and critical thinking skills (Smith *et al*., 2022). This aligns with the principles of the Kurikulum Merdeka, which emphasizes practical applications in the learning process. Such an approach increases student engagement and helps them connect abstract concepts to real-world situations. At Phase C, sixth-grade students are expected to master number operations, fraction concepts, ratios, and measurement, and to apply these skills in real-life contexts. Recent studies have highlighted that innovative learning media, such as number cards and cooperative learning strategies, effectively improve students' understanding of fraction operations and other mathematical concepts (Nurdiana, 2023; Ruswan, 2020; Wibowo *et al.,* 2024). Moreover, the demonstration method has enhanced learning outcomes by promoting a more active and interactive learning process (Rustiati, 2023). The problem-based Learning (PBL) model also significantly strengthens problem-solving skills and deepens students' comprehension of fraction concepts in sixth grade (Adiyana, 2024).

These findings show that using appropriate learning strategies can significantly enhance the quality of instruction and student learning outcomes, particularly in understanding the concept of fractions, which often poses a challenge for elementary school students. Therefore, the development of assessment instruments based on the Kurikulum Merdeka must reflect this need to ensure the relevance of learning to holistic educational goals. Quantitative item analysis is an approach based on empirical data from tested questions. This data evaluates the items' quality and ensures that the measurement instrument demonstrates high validity and reliability. A quality test should accurately measure students' abilities, and its results must be trustworthy. A test is considered to have high validity if it can measure the intended objectives. This study aims to develop a valid and reliable mathematics learning outcome assessment instrument for sixth-grade students following Kurikulum Merdeka. Specifically, the objectives of this study are: (1) to analyze the quality of the assessment instruments currently in use; and (2) to produce assessment instruments that meet standards of validity and reliability to evaluate student learning outcomes objectively.

LITERATURE REVIEW

In educational settings, assessment plays an important role as a tool to evaluate the success of the learning process while providing constructive feedback for students and educators. Effective assessment not only measures students' academic achievement but also serves as an instrument to encourage the holistic development of their competencies and abilities. In this context, the validity and reliability of assessment instruments are crucial aspects that ensure evaluations can be conducted accurately and consistently. Various approaches, such as Classical Test Theory (CTT) and Item Response Theory (IRT), have been used to improve the quality of these instruments, including applications of the Rasch model. This model helps identify item validity and provides the ability to evaluate data fit with theoretical models, making it a very effective tool in developing competency-based assessments. By adding contextual elements into assessment instruments, such as ensuring questions are relevant to everyday life, the assessment process becomes more meaningful and applicable, aligning with modern educational demands like Kurikulum Merdeka*.*

Assessment Concept in Education

https://doi.org/10.17509/jik.v22i1.77242

Educational assessment is a crucial aspect of the teaching and learning process that cannot be overlooked. This process not only serves to measure student learning outcomes but also functions as a tool to provide constructive feedback for both students and teachers. Assessment should be developed as an integral part of the evaluation process to monitor student learning progress (Marwa *et al.*, 2024; Sholikhah & Hidayati, 2024). Evaluation itself is an essential subsystem within the education system. Through evaluation, the achievements of the learning process can be analyzed, enabling informed decision-making regarding necessary improvements to enhance future learning quality (Yektiana & Nursikin, 2023).

One of the main elements in educational assessment is the validity and reliability of the instruments used. Validity refers to how an assessment instrument measures what it intends to measure. For example, if a test is designed to assess students' mathematical abilities, it must specifically reflect those abilities without being influenced by other skills such as reading or writing. Meanwhile, reliability concerns the consistency of measurement results; a reliable instrument will yield consistent outcomes when administered repeatedly under similar conditions.

Rasch Analysis Model

The Rasch analysis model has proven effective in enhancing the quality of test items during instrument development. As a method grounded in Item Response Theory (IRT), Rasch analysis offers a precise mathematical framework for evaluating the alignment between empirical data and theoretical expectations. Using fit statistics such as Infit and Outfit, this model can identify items that do not align with the measured construct, thereby facilitating instrument refinement (Novriyanti & Arthur, 2024). One of the strengths of the Rasch model is its ability to address issues such as respondent bias and uneven item difficulty levels. Research has shown that Rasch analysis supports the creation of more effective items and ensures that each item accurately differentiates among participants' abilities (Nudin & Hidayatullah, 2023). Additionally, the model supports testing for local item independence—an essential instrument development assumption requiring items to function independently (Latifah *et al*., 2024; Abdullaev *et al*., 2024).

Rasch analysis also enables verification of an instrument’s unidimensionality, ensuring that all items measure the same underlying construct. Using Principal Component Analysis of Residuals (PCAR), this model can detect unintended additional dimensions within the instrument (Yusuf *et al*., 2021). As such, Rasch analysis confirms the instrument’s reliability and strengthens its validity in consistently measuring specific competencies. Moreover, item analysis using the Rasch model offers several distinct advantages. It can detect inconsistent responses, manage missing data, and demonstrate that a respondent’s ability is determined not solely by correct answers but also by the consistency of their response patterns (Azizah & Wahyuningsih, 2020; Eliza & Yusmaita, 2021).

Recent research also confirms that the Rasch model can be applied across diverse assessment contexts, including education, medical, and psychological settings. In education, Rasch-based instruments have demonstrated their effectiveness in enhancing the validity and reliability of competency-based assessments, such as those implemented in Indonesia's Kurikulum Merdeka (Widodo, 2020). This model facilitates the creation of instruments that are more responsive and adaptive to students’ needs, thereby promoting more inclusive and equitable evaluation practices. Psychometrics offers a robust framework for developing assessment tools based on Classical Test Theory (CTT) and Item Response Theory (IRT). Compared to CTT, using IRT in instrument development provides key advantages, particularly in the sample-independent estimation of item parameters and its ability to support the design of adaptive tests. IRT enables deeper analysis of assessment instruments, contributing to constructing efficient and relevant tests for students across various ability levels (Firdaus *et al*., 2022).

Recent research further underscores the critical role of psychometrics in developing assessment instruments. Instruments based on Item Response Theory (IRT) offer distinct advantages, particularly in identifying items with low validity and providing systematic guidance for item refinement. Unlike Classical Test Theory (CTT), the psychometric approach facilitates in-depth analysis of an instrument's reliability using parameters such as item difficulty, discrimination, and guessing elements not fully addressed by CTT. These capabilities highlight the importance of psychometrics in designing assessments that align with competency-based evaluation models, which are increasingly relevant in contemporary curricula like Indonesia's Kurikulum Merdeka (Jones *et al*., 2021). Moreover, psychometrics is crucial in ensuring that instruments adhere to the principle of unidimensionality, a key requirement for consistently measuring a single latent construct. As a specific application of IRT, Rasch analysis enables researchers to detect misfitting items and optimize the overall assessment scale. Rasch-based instruments are particularly effective in evaluating student achievement comprehensively, including critical competencies such as numeracy and problem-solving skills (Kim & Kim, 2022).

Modern psychometrics has also advanced into technology by developing computer-based assessments, particularly Computerized Adaptive Testing (CAT). CAT allows tests to dynamically adapt to each participant's ability level by selecting questions based on their previous responses, enabling more precise measurement with fewer items. When designed using IRT principles, CAT enhances the testing experience for students while generating data that is both more valid and reliable (Wang & Zheng, 2023). Supported by a robust theoretical foundation and growing empirical evidence, psychometrics remains a cornerstone in developing high-quality assessment instruments. Its applications ensure the validity and reliability of assessments and enhance their relevance and efficiency across diverse educational contexts.

Contextual Assessment

Context-based assessments—such as questions rooted in everyday life—have proven effective in enhancing student engagement and motivation. Examples include calculating change in a transaction or measuring the area of a geometric shape, offering more meaningful and relatable learning experiences. This approach aligns with the Kurikulum Merdeka principles, which emphasize the development of numeracy competencies and real-world problem-solving skills. In addition to increasing motivation, contextual assessments also support the cultivation of critical thinking and analytical abilities. Students exposed to real-life contexts are more likely to engage actively with problem-solving tasks, as they perceive the material as relevant to their daily lives (Widodo, 2020).

This approach enables learners to connect abstract concepts with real-world applications, deepening their comprehension of the learning material (Smith et al., 2022). Contextual assessments also play a vital role in fostering students' literacy and numeracy skills. Studies have shown that problems grounded in everyday scenarios, such as comparing prices or calculating travel times, enhance student engagement and facilitate a more meaningful understanding of mathematical concepts. These findings underscore the value of integrating contextual elements into the curriculum to support relevant and applicable learning experiences (Jones *et al*., 2021).

In competency-based education, contextual assessment is key in identifying gaps in student understanding. Using questions grounded in real-life situations, educators can better evaluate how well students apply their knowledge in practical contexts. Studies have shown that students exposed to contextualized learning demonstrate significantly stronger problem-solving abilities than those taught through traditional methods (Nguyen *et al*., 2023). Moreover, this assessment form fosters the development of essential 21st-century skills, including collaboration, creativity, and communication. Students engaging in discussions based on real-world problems are more likely to collaborate effectively

https://doi.org/10.17509/jik.v22i1.77242

and present their ideas innovatively. As a result, contextual assessment enhances academic performance and equips students with the skills needed to navigate real-world challenges.

The use of contextual assessment also aligns with adaptive and personalized learning approaches. Assessments tailored to learners' experiences and real-life contexts have proven more effective in increasing engagement and comprehension. When assessments are personally relevant, students tend to feel more motivated and involved, fostering an inclusive learning environment that supports diverse learner backgrounds. As the demand for relevant and applicable education grows, contextual assessments provide a strong foundation for implementing competency-based education. This approach ensures that students are preparing for tests and acquiring meaningful knowledge and skills in everyday life, an essential aim of the Kurikulum Merdeka. Ultimately, contextual assessments are vital in developing a generation equipped with critical thinking, collaboration, and adaptability to meet future challenges.

METHODS

This study employed a quantitative approach using Rasch model analysis as the psychometric method to evaluate the instrument's capacity. The Rasch model was applied to assess the instrument's validity and reliability, as well as to perform Differential Item Functioning (DIF) analysis to detect potential item bias (Dwilesanti & Yudiarso, 2022; Latifah, *et al*l. 2024). Two hundred thirteen sixth-grade students from Jakarta and South Tangerang participated in the study. Participants were selected through a combination of purposive and snowball sampling methods. Initially, purposive sampling was used to recruit sixth-grade students from Sekolah Prestasi Global who took the mid-semester summative assessment. The snowball sampling method was then employed to include additional sixth-grade students who voluntarily participated (Kennedy-Shaffer *et al*., 2021). The main inclusion criterion was active enrollment in the sixth grade of elementary school. Data analysis was conducted using WINSTEPS software to ensure the precision and accuracy of the measurement results.

RESULTS AND DISCUSSION

Unidimensionality

Two fundamental assumptions must be met in the analysis using the Rasch model: unidimensionality and local independence. Unidimensionality assumes that each item in the instrument measures only one main latent construct. To ensure unidimensionality, Principal Component Analysis of Residuals (PCAR) is conducted to evaluate residual patterns that may indicate additional dimensions.

The research explored the validity and reliability of the self-directed learning scale using the Rasch model. This analysis provides in-depth information regarding unidimensionality and local independence, which are critical to ensure that the instrument measures only one central construct (unidimensionality) and that the items are independent (Nizaruddin *et al*., 2024).

Unidimensionality testing is used to assess the validity of the Rasch model, which requires that the variables be unidimensional. The dimensionality map is analyzed using WINSTEPS software through the “raw variance explained by measure” value. Measurement unidimensionality is met if the raw variance explained by the measure is equal to or greater than 20% (Latifah *et al*., 2024). In the context of the Rasch model, the unidimensionality threshold should ideally reach at least 40%, and it is preferable if it exceeds this figure (Hidayat *et al*., 2020). The unidimensionality criteria are presented in **Table 1** as follows:

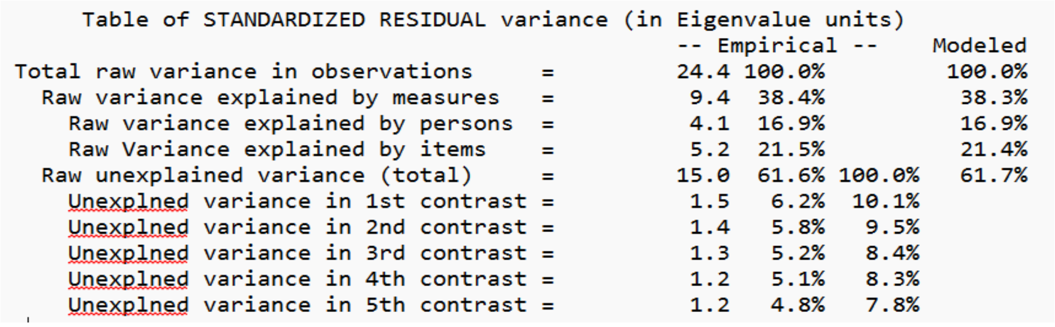
Table 1. Unidimensionality Criteria

https://doi.org/10.17509/jik.v22i1.77242

Percentage of Gross Variance Explained by Measurement	Criteria
> 60%	Very Good
40% - 60%	Good
20% - 40%	Acceptable

Source: *Latifah et al.., 2024.*

The results presented in **Figure 1** show that the total variance of the data is 24.4 (100%), with the variance explained by the model reaching 9.4 (38.4%). This analysis indicates that the model can sufficiently explain the variance, with contributions from inter-person variance at 4.1 (16.9%) and inter-item variance at 5.2 (21.5%).



```
      Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                                 -- Empirical --     Modeled
Total raw variance in observations     =         24.4 100.0%          100.0%
  Raw variance explained by measures   =          9.4  38.4%           38.3%
    Raw variance explained by persons  =          4.1  16.9%           16.9%
    Raw Variance explained by items    =          5.2  21.5%           21.4%
  Raw unexplained variance (total)     =         15.0  61.6% 100.0%    61.7%
    Unexplned variance in 1st contrast =          1.5   6.2%  10.1%
    Unexplned variance in 2nd contrast =          1.4   5.8%   9.5%
    Unexplned variance in 3rd contrast =          1.3   5.2%   8.4%
    Unexplned variance in 4th contrast =          1.2   5.1%   8.3%
    Unexplned variance in 5th contrast =          1.2   4.8%   7.8%
```

Figure 1. Standardized Residual Variance (in Eigenvalue units)
Source: *Author's Documentation 2024*

The dimensionality map also assesses local independence, which is analyzed through the unexplained variance in residual components (PCAR). This value indicates the degree to which items remain independent from any additional constructs not targeted by the main instrument. The criteria for unexplained variance are outlined in **Table 2.**

Table 2. Unexplained Variance Criteria

Unexplained Variance in Components 1-5 PCA Residuals	Criteria
< 3%	Excellent
3 - 5%	Very Good
5 - 10%	Good
10 - 15%	Acceptable
> 15%	Poor

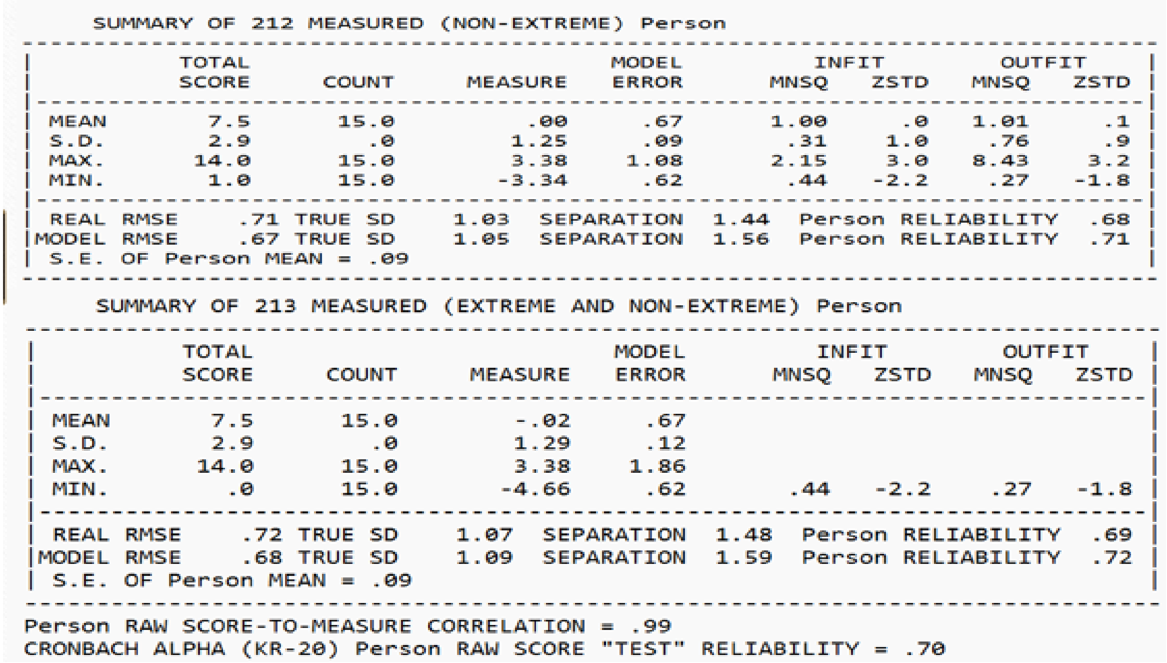
Source: *Ocy et al., 2023*

From **Figure 1,** the eigenvalue < 2.0 in the first contrast (1.5) is categorized as good, indicating that the instrument meets the unidimensionality criteria. Although the total unexplained variance of 61.6% is relatively high, the distribution still supports the measurement of a single main construct. Proportion of Variance Explained: The model accounts for 38.4% of the variance, which, although moderate, is sufficient to justify using the Rasch model. Given the low unexplained variance in the first contrast, it can be concluded that the items in this instrument measure one main construct unidimensionally. The model demonstrates adequate fit, making it a valid and reliable measurement tool for further psychometric analysis.

https://doi.org/10.17509/jik.v22i1.77242

Local Independence

Local independence refers to the assumption that learners' responses to an item are not influenced by their responses to other items after accounting for learners' latent abilities. A violation of this assumption can lead to bias in item parameter estimates, particularly when there is content similarity or logical linkage between items (Latifah *et al*., 2024; Abdullaev *et al*., 2024).



```
      SUMMARY OF 212 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE  ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       7.5       15.0         .00    .67      1.00     .0   1.01     .1 |
| S.D.       2.9         .0        1.25    .09       .31    1.0    .76     .9 |
| MAX.      14.0       15.0        3.38   1.08      2.15    3.0   8.43    3.2 |
| MIN.       1.0       15.0       -3.34    .62       .44   -2.2    .27   -1.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .71 TRUE SD    1.03  SEPARATION  1.44  Person RELIABILITY  .68 |
|MODEL RMSE    .67 TRUE SD    1.05  SEPARATION  1.56  Person RELIABILITY  .71 |
| S.E. OF Person MEAN = .09                                                   |
-------------------------------------------------------------------------------
      SUMMARY OF 213 MEASURED (EXTREME AND NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE  ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       7.5       15.0        -.02    .67                                |
| S.D.       2.9         .0        1.29    .12                                |
| MAX.      14.0       15.0        3.38   1.86                                |
| MIN.        .0       15.0       -4.66    .62       .44   -2.2    .27   -1.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .72 TRUE SD    1.07  SEPARATION  1.48  Person RELIABILITY  .69 |
|MODEL RMSE    .68 TRUE SD    1.09  SEPARATION  1.59  Person RELIABILITY  .72 |
| S.E. OF Person MEAN = .09                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .70
```

Figure 2. Person and Item Reliability
Source: *Author's Documentation 2024*

The results of the local independence test based on inter-residual correlation show that none of the inter-item correlation values exceed the critical threshold of ±0.2. Most correlation values are around zero or near zero, indicating that the relationship between items is relatively low or insignificant. This finding reflects that the items demonstrate good independence, aligning with one of the key assumptions of the Rasch model.

In addition, no high correlations (close to 1 or -1) were found between items, indicating that each item measures a different aspect of the assessed construct. This also eliminates potential multicollinearity that could compromise the validity of the measurement. Thus, all items in the instrument fulfill the assumption of local independence—that is, responses to an item are not influenced by responses to other items once latent ability is accounted for.

This result is reinforced by the visual data in **Figure 2**, which shows that none of the inter-item correlations exceed the critical threshold of ±0.2. The low residual correlation confirms the absence of systematic interrelationships between items, indicating that the instrument is free from local bias and meets the assumptions required for Rasch model analysis.

Reliability

Rasch reliability can be used to evaluate the stability of individuals and items within the instrument, serving as a key indicator provided by Rasch modeling. Reliability scores, ranging from 0 to 1, are often interpreted similarly to Cronbach's alpha and reflect the internal consistency of respondents and items. These values indicate whether the instrument demonstrates fair to excellent reliability. The criteria for individual and item reliability used in this analysis are summarized in **Table 3**.

Table 3. Person and Item Reliability Criteria

Person and Item Reliability	Criteria
> 0,94	Excellent
0,91 - 0,94	Very Good
0,81 - 0,90	Good
0,67 - 0,80	Acceptable
< 0,67	Poor

Source: Erfan, et al., 2020

The results of the analysis in **Table 3** provide a summary of the reliability of respondents from both the extreme and non-extreme groups. The Cronbach’s alpha coefficient obtained in this study was 0.70 for respondent reliability, which was classified as moderate. Meanwhile, item reliability reached 0.99, which is classified as excellent.

Thus, these results indicate that the tested instruments have high internal consistency and are reliable as measurement tools in research. These values also strengthen the instrument's validity and reliability, which aligns with psychometric standards referenced in the literature. (Erfan *et al*., 2020).

Item Suitability/Matching

In the context of Rasch model analysis, item fit evaluation aims to assess how items effectively measure the construct of interest. This process involves fit statistics, such as infit and outfit mean square (MNSQ), which provide an idea of the fit of participants' responses to the expected model. In addition, the Rasch model applies specific criteria to assess item quality (Latifah *et al*., 2024), which include:

1. Outfit MNSQ (Mean Square): 0,5 < MNSQ < 1,5
2. Outfit ZSTD (Z-Standard): -2,0 < ZSTD < +2,0
3. Point Measure Correlation: 0,4 < Pt Measure Corr < 0,85

Items in the instrument that do not meet these three criteria are considered a “misfit” and require revision or replacement. However, if an item meets at least two of the three criteria, then the item can still be declared fit with the model and is suitable for use in the instrument. (Azizah & Wahyuningsih, 2020).

Furthermore, the Outfit MNSQ value is a key indicator when assessing item fit. Items with values outside the preset range require further evaluation to ensure that the instrument maintains sufficient quality and is bias-free (Noben *et al*., 2021).

Table 4. Item Suitability/Matching

Item	Outfit			PT-MEASURE		Fit/misfit
	MNSQ	MNSQ Criteria	ZSTD	Correlation	PM Criteria	
Item10	1.34	Good	1.3	0.33	Good	Fit
Item5	1.3	Good	1.0	0.34	Good	Fit

https://doi.org/10.17509/jik.v22i1.77242

Item	Outfit			PT-MEASURE		Fit/misfit
	MNSQ	**MNSQ Criteria**	**ZSTD**	**Correlation**	**PM Criteria**	
Item9	1.17	Good	0.8	0.36	Good	Fit
Item8	1.11	Good	1.0	0.43	Excellent	Fit
Item6	1.07	Good	0.6	0.42	Excellent	Fit
Item12	1.07	Good	0.5	0.42	Excellent	Fit
Item15	0.85	Good	-0.5	0.36	Good	Fit
Item13	1.04	Good	0.5	0.48	Excellent	Fit
Item2	0.99	Good	0.0	0.36	Good	Fit
Item7	0.94	Good	-0.4	0.49	Excellent	Fit
Item3	0.83	Good	-1.0	0.47	Excellent	Fit
Item14	0.95	Good	-0.3	0.48	Sangat Baik	Fit
Item11	0.80	Good	0.8	0.33	Good	Fit
Item1	0.88	Good	-1.1	0.53	Excellent	Fit
Item4	0.77	Good	-1.7	0.55	Excellent	Fit

Source: Primary data 2024

Table 4 presents the results of the evaluation of item suitability based on three main parameters, namely Outfit MNSQ, ZSTD (Z-Standard), and Point Measure Correlation (PM).

1. **Outfit MNSQ Parameter**
   Outfit MNSQ measures the extent to which the data fits the Rasch model. The analysis results show that all items have values within the ideal range (0.5–1.5), which meets the model fit criteria. The item with the highest Outfit MNSQ value is Item10 (1.34), while the lowest is Item4 (0.77). No items show extreme misfit, thus all are considered to fit the model appropriately.
2. **ZSTD Parameter (Z-Standard)**
   ZSTD evaluates statistical deviations in the data. All items have ZSTD values within the ideal range of -2 to +2, indicating that no item significantly deviates from the model. The highest ZSTD value is found in Item10 (1.3), reflecting a slight deviation that still falls within the acceptable range. The lowest ZSTD value is in Item4 (-1.7), indicating a very good fit with the model.
3. **Point Measure Correlation Parameter (PM)**
   The PM parameter measures the relationship between item response patterns and participant ability, with an ideal value ranging from 0.4 to 0.85. Most items in the instrument have correlations above 0.4, indicating a good association with participant ability. Exceptions include Item10 (0.33), Item5 (0.38), and Item9 (0.39), which fall slightly below the ideal threshold. The item with the highest correlation is Item4 (0.55), reflecting a strong relationship with participant ability. Despite its slightly lower PM value, Item10 is still considered to fit the model.

All items were declared fit with the Rasch model based on the evaluation of the three parameters (Outfit MNSQ, ZSTD, and PM). No items needed to be deleted or revised due to model inconsistency. These results indicate that the instrument possesses excellent quality and can be relied upon to measure the intended construct. This evaluation also confirmed that participants' responses to the items were consistent with the assumptions of the Rasch model, thereby strengthening the validity and reliability of the instrument used.

Item Difficulty Level

The item difficulty level in Rasch modeling is divided into four categories based on the logit value and the logit item's standard deviation (SD) (Ocy *et al*., 2023). A higher logit value indicates a greater level of item difficulty. Prior to analyzing the item difficulty levels, the standard deviation (SD) was determined to be 1.38.

Table 5. Item Difficulty Level

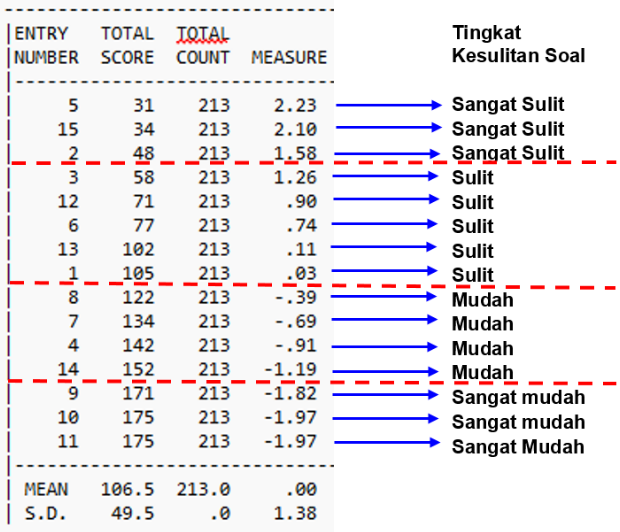
Logit Value	Category	Items
Greater than +1,38	Very Difficult	2,5,15
Greater than 0,0 logit end less than +01,38 SD	Difficult	1,3,6,12,13
Less than 0,0 logit - and greater than -1,38 SD	Easy	4,7,8,14
less than -1,38	Very Easy	9, 10,11

Source: Primary data 2024

Based on the data presented in **Table 5**, the item difficulty category is determined using a standard deviation (SD) of 1.38. Three items are classified as very difficult, namely items 2, 5, and 15. These questions have a high level of difficulty and tend to be answered correctly only by participants with very high abilities.

Questions with logit values between 0 and +1.38 are categorized as difficult. Difficult items include 1, 3, 6, 12, and 13. These questions are challenging but can still be answered by participants with above-average abilities. Items with logit values between 0 and -1.38 are categorized as easy. These questions include 4, 7, 8, and 14, which are relatively more straightforward and suitable for participants with medium to low abilities. Meanwhile, items with logit values less than -1.38 are categorized as very easy. These questions, namely 9, 10, and 11, are designed for participants with very low abilities, allowing them to achieve scores more easily. This can be seen more clearly in **Figure 3**.

https://doi.org/10.17509/jik.v22i1.77242



```
--------------------------------
|ENTRY   TOTAL  TOTAL                     Tingkat
|NUMBER  SCORE  COUNT  MEASURE            Kesulitan Soal
|--------------------------------
|     5     31    213     2.23  ------->  Sangat Sulit
|    15     34    213     2.10  ------->  Sangat Sulit
|     2     48    213     1.58  ------->  Sangat Sulit
|     3     58    213     1.26  ------->  Sulit
|    12     71    213      .90  ------->  Sulit
|     6     77    213      .74  ------->  Sulit
|    13    102    213      .11  ------->  Sulit
|     1    105    213      .03  ------->  Sulit
|     8    122    213     -.39  ------->  Mudah
|     7    134    213     -.69  ------->  Mudah
|     4    142    213     -.91  ------->  Mudah
|    14    152    213    -1.19  ------->  Mudah
|     9    171    213    -1.82  ------->  Sangat mudah
|    10    175    213    -1.97  ------->  Sangat mudah
|    11    175    213    -1.97  ------->  Sangat Mudah
|--------------------------------
| MEAN   106.5  213.0      .00
| S.D.    49.5     .0     1.38
```

Figure 3. Item Difficulty Level
Source: *Author's Documentation 2024*

Item Difficulty Distribution Analysis: The item difficulty distribution in this instrument shows various difficulty levels, ranging from very easy to very difficult. This reflects the instrument's ability to measure participants with various levels of ability effectively. However, the number of items in the very difficult and easy categories is relatively smaller than that in the difficult and easy categories. This distribution indicates that the instrument focuses more on measuring the participants' average ability. This structure can increase the instrument's sensitivity in assessing participants' abilities at the level most commonly found in the population.

Student Ability Levels in Item Responses

In line with the analysis of item difficulty levels, students' ability to answer items within the Rasch model is grouped into four categories based on the logit value and the standard deviation (SD) of the item logits (Ocy *et al*., 2023). These student ability levels are used to identify how well students answer the questions, which are sorted from the highest to the lowest logit value.

In this analysis, the standard deviation (SD) is set at 1.29, with an average person logit value of -0.02. This value is a reference point for categorizing student abilities into several groups.

Table 6. Student Ability Levels in Item Responses

Logit Value of Student Ability	Criteria	Number of Students
Logit > +1,29	Very High	33
-0.02 < logit ≤ 1,29	High	77
-1,29 ≤ logit ≤ -0,02	Low	67
Logit < -1,29	Very Low	36

Source: Primary data 2024

Analysis of Student Ability Distribution: Based on the data presented in Table 6, the distribution of student ability shows that the majority of students, 110 out of 213 students (51.6%), have abilities categorized as high to very high. However, 103 students (48.4%) fall into the low to very low category.

These results indicate that the measurement instrument effectively identifies variations in student abilities at various levels. However, these findings also indicate the need for more focused learning interventions for groups of students with low and very low abilities. This step is important to improve their competence and ensure an equal distribution of learning quality across ability groups.

Item Difficulty Level vs Student Ability Levels in Item Responses

Table 7 illustrates the relationship between the level of difficulty of the test items (logit item) and the students' (logit person) ability to answer certain items.

Table 7. Item Difficulty Level vs Student Ability Levels in Item Responses

Person Measure	Respondents/ Students	Item Number and Difficulty Level														
		-1.97	-1.97	-1.82	-1.19	-.91	-.69	-.39	.03	.11	.74	.90	1.26	1.58	2.10	2.23
		10	11	9	14	4	7	8	1	13	6	12	3	2	15	5
3.38	140P	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
3.38	290L	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
3.38	207P	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
2.50	001P	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
2.50	055L	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1
2.50	079P	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1
2.50	093L	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0
2.50	141L	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1
2.50	161L	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1
1.90	044L	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0
1.90	071L	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1
1.90	072L	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0
1.90	075P	1	1	1	1	1	1	0	1	1	1	1	0	1	0	1
1.90	109P	1	1	1	1	1	1	1	0	1	0	1	1	1	1	0
1.90	138L	1	1	1	0	1	0	1	1	1	1	1	0	1	1	1
1.41	026P	1	1	1	1	1	1	0	1	1	1	1	0	1	0	0
1.41	050L	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1
1.41	058P	1	1	0	1	1	1	1	1	1	1	0	1	0	1	0
1.41	060P	1	1	1	1	1	1	1	1	0	1	0	1	0	1	0
1.41	064P	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0

Source: Primary data 2024

Based on the data displayed in **Table 7**, there are several interesting findings related to the relationship between **student abilities** and **item difficulty levels**:

1. Students with High Logit Who Did Not Answer Difficult Questions. Several students with high logit values (ability above 2.5), such as 140P, 290L, 207P, 001P, and 055L, were above the difficulty level of question number 5 (logit 2.23). However, they were unable to answer this question correctly. This failure may have been caused by carelessness when answering the question, even though their abilities should have been sufficient to solve it correctly.
2. Students with Low Logit Who Successfully Answered Difficult Questions: On the other hand, students with lower logit values (1.90), such as 071L, successfully answered question number 5 (logit 2.23) correctly. This success may be attributed to a correct guess (lucky guess), considering that the question's difficulty level was above the student’s actual ability.
3. Polarization of Ability and Difficulty Level. In general, students with high ability (high logit values) tend to be able to answer questions with low to medium difficulty levels. However, several anomalies were observed. Some high-ability students failed to answer difficult questions, which may have been influenced by factors such as lack of accuracy, low motivation, or concentration issues. Conversely, some low-ability students succeeded in answering difficult questions, which may indicate the use of a correct guessing strategy.

https://doi.org/10.17509/jik.v22i1.77242

The relationship between learner ability and item difficulty in this instrument generally follows the expected pattern of the Rasch model. However, as previously discussed, the presence of anomalies highlights the need for further evaluation of the instrument. This evaluation aims to: 1) Identify non-ability factors that may influence the results, such as motivation and answering strategies; 2) Minimize the impact of guessing or lack of accuracy through retesting or additional training; and 3) Enhance the instrument's validity and reliability by revising items whose results are inconsistent with the Rasch model.

Item Characteristic Curve (ICC) Analysis in Item Response Theory (IRT)

The Item Characteristic Curve (ICC) in Item Response Theory (IRT) is a tool used to visualize the relationship between a test taker's ability and the probability of answering an item correctly. The ICC provides valuable insights into the difficulty and discrimination of an item, enabling a more in-depth analysis of the quality of a test instrument. This tool is effective for evaluating item quality, monitoring the consistency of item performance, and identifying items that require revision. Additionally, the ICC supports valid and reliable measurement by ensuring that items accurately and efficiently assess a test taker's ability. (Oktaviyanthi *et al*., 2024).

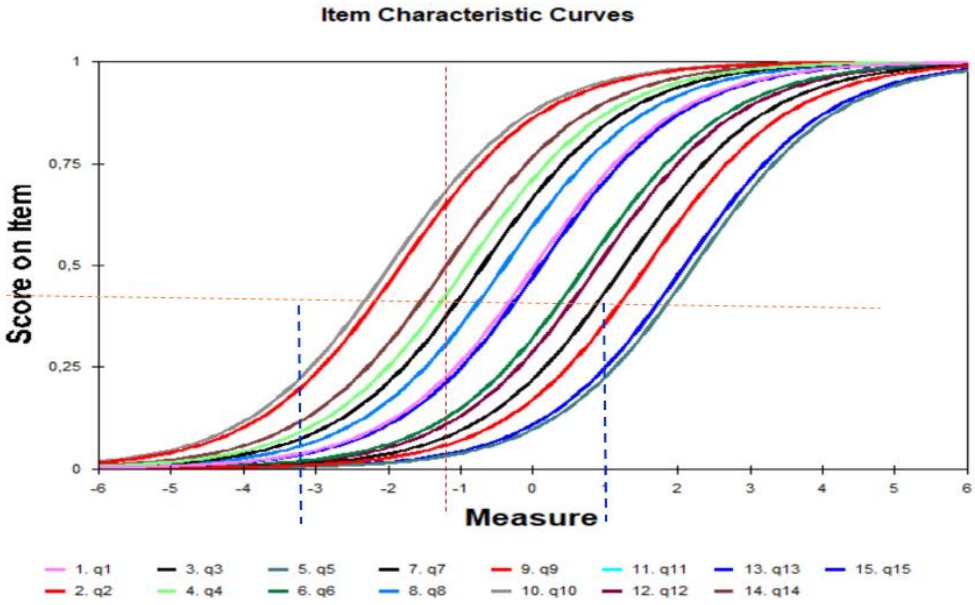


Figure 4. ICC Graph
Source: Author's Documentation 2024

Figure 4, which presents the Item Characteristic Curves (ICC), shows that items with curves shifted to the right, such as Item15 and Item5, have higher difficulty levels. Conversely, items with curves shifted to the left, such as Item11 and Item10, are associated with lower difficulty levels. The curve's slope reflects each item's discriminatory power, indicating how well the item can distinguish learners based on their abilities.

All items in the graph exhibit a uniform slope, indicating that each item has adequate discriminatory power. The ICC curve approaches a value of 1 (100% probability of answering correctly) for students with high ability and a value of 0 (0% probability of answering correctly) for students with low ability. This demonstrates that the Rasch model effectively reflects the relationship between student ability and item difficulty. The distribution of difficulty levels is well-balanced, covering items ranging from very easy to very difficult. This suggests that the test has a diverse set of items, effectively measuring students with varying ability levels.

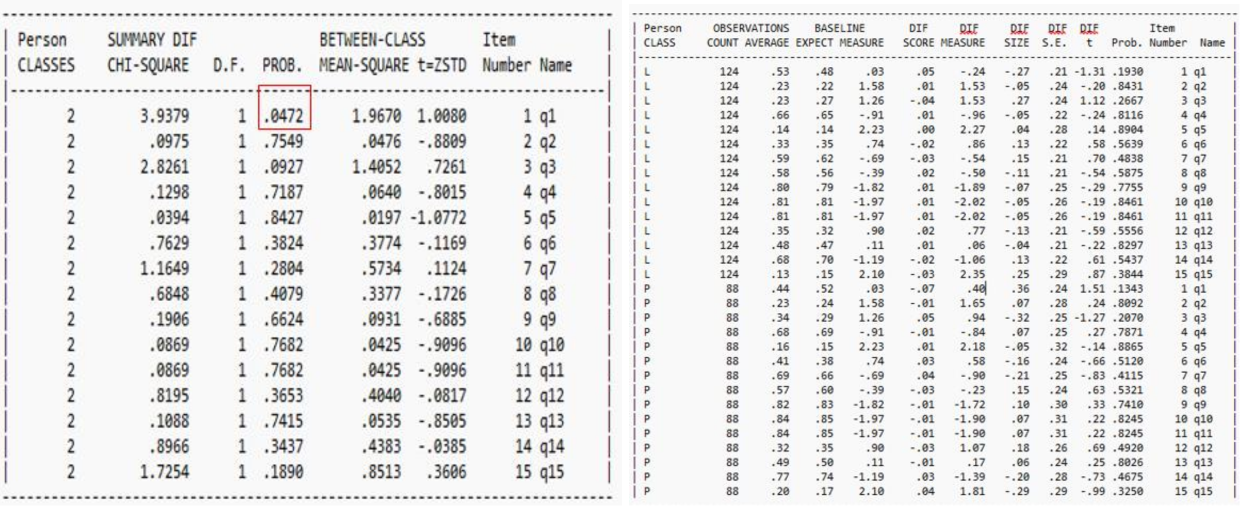
https://doi.org/10.17509/jik.v22i1.77242

The items in this test show good performance based on the characteristic curves. Items with high difficulty levels (such as **Item15** and **Item5**) provide useful information for measuring students with high abilities. On the other hand, **items** with low difficulty levels (such as **Item11** and **Item10**) are useful for measuring students with low abilities. Overall, this test has adequate quality in measuring students' abilities at various levels, in line with the principles of the Rasch model.

Differential Item Functioning **(DIF)**

Differential Item Functioning (DIF) is a phenomenon where items in a test exhibit different behaviors across groups of participants with similar abilities, which can lead to bias and reduce the validity of the measurement instrument. DIF is used to assess whether an item functions fairly across groups based on certain characteristics, such as gender, age, culture, or educational background. **(**El Fahmi *et al*, 2021; Peng *et al*., 2024).

Several studies have emphasized the importance of detecting and addressing Differential Item Functioning (DIF) to ensure the fairness and validity of measurement instruments. A Rasch analysis of the Autism Behavior Checklist (ABC) revealed the presence of DIF, highlighting the need for special consideration of population characteristics (Peng *et al*., 2024). Furthermore, semi-automatic methods for Rasch analysis that account for potential DIF simplify the complex analytical process and reduce subjectivity in decision-making (Wijayanto *et al*., 2023).



```
---------------------------------------------------------------------------
| Person    SUMMARY DIF                  BETWEEN-CLASS        Item        |
| CLASSES   CHI-SQUARE  D.F.  PROB.   MEAN-SQUARE  t=ZSTD  Number Name    |
|-------------------------------------------------------------------------|
|      2       3.9379     1  .0472      1.9670  1.0080      1 q1          |
|      2        .0975     1  .7549       .0476  -.8809      2 q2          |
|      2       2.8261     1  .0927      1.4052   .7261      3 q3          |
|      2        .1298     1  .7187       .0640  -.8015      4 q4          |
|      2        .0394     1  .8427       .0197 -1.0772      5 q5          |
|      2        .7629     1  .3824       .3774  -.1169      6 q6          |
|      2       1.1649     1  .2804       .5734   .1124      7 q7          |
|      2        .6848     1  .4079       .3377  -.1726      8 q8          |
|      2        .1906     1  .6624       .0931  -.6885      9 q9          |
|      2        .0869     1  .7682       .0425  -.9096     10 q10         |
|      2        .0869     1  .7682       .0425  -.9096     11 q11         |
|      2        .8195     1  .3653       .4040  -.0817     12 q12         |
|      2        .1088     1  .7415       .0535  -.8505     13 q13         |
|      2        .8966     1  .3437       .4383  -.0385     14 q14         |
|      2       1.7254     1  .1890       .8513   .3606     15 q15         |
---------------------------------------------------------------------------
```

```
-------------------------------------------------------------------------------------------------------
| Person     OBSERVATIONS     BASELINE        DIF       DIF     DIF   DIF   DIF            Item       |
| CLASS     COUNT AVERAGE  EXPECT MEASURE    SCORE MEASURE    SIZE  S.E.    t  Prob.  Number  Name   |
|------------------------------------------------------------------------------------------------------
| L          124     .53     .48     .03      .05    -.24    -.27   .21 -1.31 .1930       1 q1        |
| L          124     .23     .22    1.58      .01    1.53    -.05   .24  -.20 .8431       2 q2        |
| L          124     .23     .27    1.26     -.04    1.53     .27   .24  1.12 .2667       3 q3        |
| L          124     .66     .65    -.91      .01    -.96    -.05   .22  -.24 .8116       4 q4        |
| L          124     .14     .14    2.23      .00    2.27     .04   .28   .14 .8904       5 q5        |
| L          124     .33     .35     .74     -.02     .86     .13   .22   .58 .5639       6 q6        |
| L          124     .59     .62    -.69     -.03    -.54     .15   .21   .70 .4838       7 q7        |
| L          124     .58     .56    -.39      .02    -.50    -.11   .21  -.54 .5875       8 q8        |
| L          124     .80     .79   -1.82      .01   -1.89    -.07   .25  -.29 .7755       9 q9        |
| L          124     .81     .81   -1.97      .01   -2.02    -.05   .26  -.19 .8461      10 q10       |
| L          124     .81     .81   -1.97      .01   -2.02    -.05   .26  -.19 .8461      11 q11       |
| L          124     .35     .32     .90      .02     .77    -.13   .21  -.59 .5556      12 q12       |
| L          124     .48     .47     .11      .01     .06    -.04   .21  -.22 .8297      13 q13       |
| L          124     .68     .70   -1.19     -.02   -1.06     .13   .22   .61 .5437      14 q14       |
| L          124     .13     .15    2.10     -.03    2.35     .25   .29   .87 .3844      15 q15       |
| P           88     .44     .52     .03     -.07     .40     .36   .24  1.51 .1343       1 q1        |
| P           88     .23     .24    1.58     -.01    1.65     .07   .28   .24 .8092       2 q2        |
| P           88     .34     .29    1.26      .05     .94    -.32   .25 -1.27 .2070       3 q3        |
| P           88     .68     .69    -.91     -.01    -.84     .07   .25   .27 .7871       4 q4        |
| P           88     .16     .15    2.23      .01    2.18    -.05   .32  -.14 .8865       5 q5        |
| P           88     .41     .38     .74      .03     .58    -.16   .24  -.66 .5120       6 q6        |
| P           88     .69     .66    -.69      .04    -.90    -.21   .25  -.83 .4115       7 q7        |
| P           88     .57     .60    -.39     -.03    -.23     .15   .24   .63 .5321       8 q8        |
| P           88     .82     .83   -1.82     -.01   -1.72     .10   .30   .33 .7410       9 q9        |
| P           88     .84     .85   -1.97     -.01   -1.90     .07   .31   .22 .8245      10 q10       |
| P           88     .84     .85   -1.97     -.01   -1.90     .07   .31   .22 .8245      11 q11       |
| P           88     .32     .35     .90     -.03    1.07     .18   .26   .69 .4920      12 q12       |
| P           88     .49     .50     .11     -.01     .17     .06   .24   .25 .8026      13 q13       |
| P           88     .77     .74   -1.19      .03   -1.39    -.20   .28  -.73 .4675      14 q14       |
| P           88     .20     .17    2.10      .04    1.81    -.29   .29  -.99 .3250      15 q15       |
```

Figure 5. Person and Item Reliability
Source: *Author's Documentation 2024*

The data analysis using the Rasch model with Winsteps software is presented in **Figure 5**.

Most **items** exhibit relatively high probability values (> 0.05) in the Chi-Square column, suggesting no significant difference in item functioning across participant groups. However, **Item1** shows a probability value of 0.0472, which is below the 0.05 threshold, indicating potential DIF bias. (Wahyuni, 2022).

Further analysis of the Between-Class MEAN-SQUARE and t-ZSTD columns reveals that Item1 has a t-ZSTD value of 1.0800, which is approaching the significance threshold (±2). This further supports the indication of bias in the item. Additionally, Item5 exhibits a notable difference in the DIF Score Measure between the L group (-1.53) and the P group (2.18), suggesting that female participants are more likely to answer this item correctly compared to male participants.

DIF is employed to evaluate whether an item functions equitably across groups of participants based on specific characteristics. The majority of items in the table exhibit relatively high probability values (> 0.05)

https://doi.org/10.17509/jik.v22i1.77242

in the Chi-Square column, suggesting no significant differences in item functioning between participant groups. However, Item1 has a probability value of 0.0472, which falls below the 0.05 threshold, signaling potential DIF bias (Wahyuni, 2022). In the **Between-Class MEAN-SQUARE** and **t-ZSTD** columns, **Item1** shows a **t-ZSTD** value of 1.0800, approaching the significant threshold (±2), which further supports the indication of bias. Furthermore, the detailed table indicates that **Item5** displays a significant difference in the DIF Score Measure between class L (-1.53) and class P (2.18), suggesting that female participants are more likely to answer this item correctly than male participants.

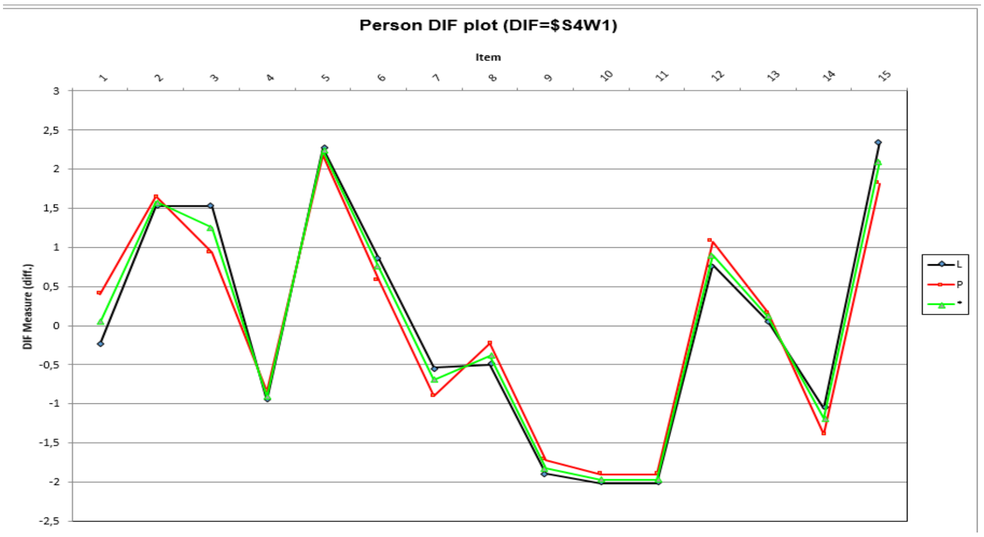


Figure 6. DIF Plot of Gender Bias on Significance of Answer Items
Source: *Author's Documentation 2024*

Figure 6 illustrates that certain items, such as **Item1** and **Item5**, exhibit non-overlapping lines, suggesting significant differences between groups. Conversely, items like **Item8** and **Item10** display overlapping lines, indicating no significant differences in item difficulty across groups.

Most items exhibited consistent functioning, indicating that the overall quality of the instrument is relatively high. However, items such as **Item1** and **Item5** require further scrutiny to ensure the fairness of the measurement. The DIF bias observed in these items may be influenced by external factors, including cultural context or the specific experiences of the participants. (Bialo & Li., 2024; Ray *et al*, 2024).

Differential Item Functioning (DIF) handling seeks to maintain the fairness and validity of the measurement instrument through a series of methodical steps. The process begins with the identification of problematic items, where items exhibiting DIF bias are detected using Rasch model-based statistical analysis. Following this, the next step involves item revision or removal. Items identified with bias can be revised to correct the imbalance, or, if revision proves ineffective, the item may be excluded from the instrument. The final stage is re-evaluation, where the revised instrument is re-administered to a more diverse population to ensure that no additional bias is present.

In addition, in the context of instrument validation, if the removal of items with DIF compromises content validity, those items should no longer be used in comparisons of individual measures, such as mean scores across sample groups. However, if items with DIF are retained, it is essential to take into account respondents' cultural, linguistic, and background factors when administering the instrument. These considerations are crucial to maintaining the fairness and accuracy of the instrument, ensuring that it objectively and consistently measures participants' abilities.

Discussion

This study highlights the development of a mathematics learning outcome assessment instrument based on the Kurikulum Merdeka, using the Rasch model approach. The main findings indicate that the instrument possesses high validity and reliability, supported by analysis using the Winsteps software. With a Cronbach's alpha of 0.70 and item reliability of 0.99, the instrument meets the criteria for unidimensionality and reflects a wide range of item difficulty levels, from very easy to very difficult. However, further analysis reveals the presence of DIF bias in Item 1 and Item 5 based on gender, indicating the need for revision or adjustment to ensure that the instrument measures student ability more fairly across different demographic groups.

This study reinforces the findings of previous research regarding the validity and reliability of instruments based on the Rasch model. The Rasch model demonstrates a strong capability to detect misfitting items and to provide insights into item discrimination and difficulty levels. Prior studies have highlighted the model's effectiveness in evaluating unidimensionality, ensuring that each item measures the same underlying construct (Latifah *et al*., 2024; Tarigan *et al*., 2022). Additionally, the detection of DIF bias has been emphasized as crucial for enhancing the fairness of assessments across demographic groups (Wahyuni, 2022). This study advances earlier research by comprehensively integrating both DIF analysis and unidimensionality evaluation, thereby supporting the Rasch model's role in identifying response anomalies, as previously stated (Eliza & Yusmaita, 2021).

This study also complements existing research that employs Item Response Theory (IRT) and psychometric approaches in evaluating the quality of assessment instruments. It highlights the development of competency-based instruments for contextual learning, emphasizing validity and reliability as essential components of effective evaluation (Komisia *et al*., 2021). The integration of reliability analysis, item discrimination, and DIF bias detection reinforces the methodological approach adopted in this study (Natanael *et al*., 2022). Furthermore, the use of machine learning to identify DIF bias offers promising insights for the development of more adaptive and advanced assessment instruments in the future (Peng *et al*., 2024).

The significant contribution of this study lies in the development of a new, valid, and reliable assessment instrument based on the Kurikulum Merdeka to evaluate the mathematics abilities of sixth-grade students. This study also introduces an integrated approach that combines DIF analysis, unidimensionality assessment, and item reliability evaluation, resulting in a comprehensive and high-quality measurement tool. Furthermore, it proposes a practical framework for addressing gender-biased items, thereby enhancing the fairness and effectiveness of the evaluation instrument.

CONCLUSION

This study successfully developed a mathematics learning outcome assessment instrument based on the Kurikulum Merdeka for sixth-grade Elementary School students with a focus on validity and reliability using the Rasch model. This instrument covers cognitive, affective, and psychomotor aspects, and has been tested on 213 students with a quantitative approach using Winsteps software. The results of the analysis showed that the instrument has high reliability (Alpha Cronbach = 0.70; item reliability = 0.99) and meets the unidimensionality criteria. The distribution of item difficulty levels that vary from very easy to very difficult reflects the instrument's ability to measure students with varying levels of ability. All items meet the suitability criteria based on fit statistics (Outfit MNSQ, ZSTD, and Point Measure Correlation). However,

https://doi.org/10.17509/jik.v22i1.77242

two items (Item 1 and Item 5) showed significant Differential Item Functioning (DIF) bias based on gender analysis, which requires further evaluation and adjustment. Thus, the instrument developed in this study is declared valid, reliable, and suitable for use as a tool for evaluating the mathematics abilities of sixth-grade students in accordance with the principles of the Kurikulum Merdeka. The use of the Rasch model has proven effective in analyzing item quality in depth, ensuring that the resulting instrument is able to provide fair, accurate, and comprehensive measurements. The results of this study are expected to be a reference for the development of other assessment instruments that are oriented towards improving the quality of learning and evaluating student learning outcomes. Improving the quality of assessment instruments requires re-examination of items with significant DIF bias, especially in the context of language, culture, and participant experience. Modification or replacement of items can be done to ensure measurement fairness across groups. In addition, it is important to increase the number of items in the very easy and very difficult categories in order to expand the range of abilities that can be measured, so that the instrument is more adaptive to groups of students with extreme abilities. Training for teachers is also a priority so that they can understand the use of Rasch model-based instruments, read the analysis results correctly, and provide more effective feedback to students. Further validation tests involving larger and more diverse samples from various regions are also needed to validate the instrument and increase the generalizability of the research results. Finally, this instrument can be practically implemented as an assessment tool in Kurikulum Merdeka-based learning, supporting more holistic and fair measurements for students.

AUTHOR'S NOTE

The author declares that there is no conflict of interest regarding the publication of this article. The author confirms that the data and content of the article are free from plagiarism.

REFERENCES

Abdullaev, D., Shukhratovna, D. L., Rasulovna, J. O., Umirzakovich, J. U., & Staroverova, O. V. (2024). Examining local item dependence in a cloze test with the Rasch Model. *International Journal of Language Testing*, *14*(1), 75-81.

Adiyana, S. (2024). Peningkatan kemampuan menghitung pecahan melalui Model Problem Based Learning pada siswa kelas VI SDN 01 Ngunut. *Jurnal Edukasi Indonesia, 12*(3), 120-136.

Azizah, & Wahyuningsih, S. (2020). Penggunaan model Rasch untuk analisis instrumen tes pada Mata kuliah Matematika Aktuaria**.** *Jurnal Pendidikan Matematika (Jupitek)*, *3*(1), 45-50.

Bialo, J. A., & Li, H. (2024). An analysis of DIF and sources of DIF in achievement motivation items using anchoring vignettes. *Educational Assessment*, *29*(4), 293-318.

Dwilesanti, W. G., & Yudiarso, A. (2022). Rasch analysis of the Indonesian version of INDIVIDUAL Work Performance Questionnaire (IWPQ). *JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)*, *11*(2), 153-167.

El Fahmi, E., Khoirot, U., & Astutik, F. (2021). Analisis psikometri aitem need of agression tes EPPS pada remaja akhir. *Psikoislamika: Jurnal Psikologi dan Psikologi Islam, 18*(2), 295-306.

Eliza, W., & Yusmaita, E. (2021). Pengembangan butir soal literasi Kimia pada materi sistem koloid kelas XI IPA SMA/MA. *Jurnal Eksakta Pendidikan (JEP)*, *5*(2), 197-204.

Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). Analisis kualitas soal kemampuan membedakan rangkaian seri dan paralel melalui teori tes klasik dan model rasch. *Indonesian Journal of Educational Research and Review*, *3*(1), 11-19.

https://doi.org/10.17509/jik.v22i1.77242

Firdaus, F., Huda, A., Irfan, D., & Hebdriyani, Y. (2022). Pengembangan sistem Computer Adaptive Test (CAT) dengan pendekatan Item Response Theory (IRT). *EduTech: Jurnal Teknologi Pendidikan*, *21*(3), 272-286.

Hidayat, R., Patras, Y. E., Harijanto, S., & Hasanah, L. (2020). Analisis instrumen dan prioritas tindakan untuk kepuasan kerja guru di Indonesia berdasarkan pemodelan Rasch. *Kelola: Jurnal Manajemen Pendidikan*, *7*(2), 110-130.

Jones, R. J., Brown, D. E., & Smith, T. L. (2021). Competency-based assessment in modern curriculum: A contextual approach. *Educational Measurement Quarterly, 45*(3), 150-168.

Juliani, R. P., & Erita, S. (2023). Analisis validitas dan reliabilitas instrumen penilaian kemampuan berpikir kritis dalam konteks sekolah menengah. *JEID: Journal of Educational Integration and Development*, *3*(3), 169-179.

Jumini, J., & Retnawati, H. (2022). Estimating item parameters and student abilities: An IRT 2PL analysis of mathematics examination. *Al-Ishlah: Jurnal Pendidikan*, *14*(1), 385-398.

Kennedy-Shaffer, L., Qiu, X., & Hanage, W. P. (2021). Snowball sampling study design for serosurveys early in disease outbreaks. *American Journal of Epidemiology*, *190*(9), 1918-1927.

Kim, S., & Kim, J. (2022). Advancing Rasch analysis for holistic student assessment. *Journal of Educational Measurement, 59*(1), 78-95.

Komisia, F., Tukan, M. I. B., & Leba, M. A. U. (2021). Pengembangan perangkat pembelajaran berbasis pendekatan kontekstual untuk siswa SMA. *Indonesian Journal of Educational Science (IJES)*, *3*(2), 98-104.

Latifah, M., Saripah, I., Suryana, D., & Sunarya, Y. (2024). Validity and reliability of self-concept instrument using Rasch Model. *Jurnal Kajian Bimbingan dan Konseling*, *9*(1), 26-35.

Marwa, N. W. S., Pitria, P. R., & Madani, F. (2024). Development of authentic assessment of 21st-century skills in kurikulum merdeka. *Inovasi Kurikulum, 21*(2), 635-646.

Maulana, A. (2022). Analisis validitas, reliabilitas, dan kelayakan instrumen penilaian rasa percaya diri siswa. *Jurnal Kualita Pendidikan*, *3*(3), 133-139.

Natanael, Y., Salsabilla, R., Aulia, D., Khoirunnisa, D., Munawar, H. N., Hidayat, N. S., & Firdaus, R. F. (2022). Rasch rating scale model: Bias detection and validation test of Indonesian-adolescent life satisfaction scale. *Psympathic: Jurnal Ilmiah Psikologi, 9*(1), 31-44.

Nguyen, T., Pham, L., & Tran, H. (2023). Context-based learning and its impact on problem-solving skills*. Educational Research Review, 58*(1), 45-67.

Nizaruddin, N., Muhtarom, M., Murtianto, Y. H., & Sutrisno, S. (2024). Examining the self-regulated learning scale using the Rasch model approach. *Indonesian Journal of Science and Mathematics Education*, *7*(3), 518-530.

Noben, I., Maulana, R., Deinum, J. F., & Hofman, W. A. (2021). Measuring university teachers' teaching quality: A Rasch modelling approach. *Learning Environments Research*, *24*(1), 87-107.

Novriyanti, E., & Arthur, R. (2024). Analisis kualitas butir soal ujian tengah semester Biologi umum menggunakan Model Rasch. *JagoMIPA: Jurnal Pendidikan Matematika dan IPA*, *4*(4), 718-733.

Nudin, I., & Hidayatullah, R. S. (2023). Analisis butir soal penilaian tengah semester menggunakan model Rasch di SMK Negeri 5 Surabaya. *JPTM*, *12*(2), 132-139.

Nurdiana, N. (2023). Meningkatkan hasil belajar operasi hitung bilangan pecahan dengan kartu bilangan siswa kelas VI SD Negeri Krueng Baung. *Jurnal Bima: Pusat Publikasi Ilmu Pendidikan Bahasa dan Sastra*, *1*(3), 338-348.

Ocy, D. R., Rahayu, W., & Makmuri, M. (2023). Rasch model analysis: Development of hots-based mathematical abstraction ability instrument according to Riau Islands Culture. *Aksioma: Jurnal Program Studi Pendidikan Matematika*, *12*(4), 3542-3560.

https://doi.org/10.17509/jik.v22i1.77242

Oktaviyanthi, R., Agus, R. N., Garcia, M. L. B., & Lertdechapat, K. (2024). Cognitive load scale in learning formal definition of limit: A Rasch model approach. *Infinity Journal of Mathematics Education, 13*(1), 99-118.

Peng, K., Chen, M., Zhou, L., & Weng, X. (2024). Differential item functioning in the autism behavior checklist in children with autism spectrum disorder based on a machine learning approach. *Frontiers in Psychiatry, 15*(1), 1-14.

Ray, J. V., Baker, T., & Peck, J. H. (2024). An examination of differential item functioning in a measure of self-reported offending across race and ethnicity among a sample of justice-involved youth. *Justice Quarterly, 1*(1), 1-25.

Rustiati, T. (2023). Upaya meningkatkan hasil belajar siswa kelas VI SD pada konsep operasi hitung pecahan pada mata pelajaran Matematika melalui metode demostrasi. *Jurnal Pendidikan Abad Ke-21*, *1*(1), 17-29.

Ruswan, R. (2020). Penggunaan pendekatan kooperatif dalam pembelajaran Matematika tentang operasi hitung pecahan untuk meningkatkan hasil belajar siswa sekolah dasar. *Pedadidaktika: Jurnal Ilmiah Pendidikan Guru Sekolah Dasar*, *7*(3), 58-67.

Safitri, E., & Widyanti, E. (2024). Analisis penilaian guru yang efektif pada pencapaian kompetensi pengetahuan siswa. *Ihsan: Jurnal Pendidikan Islam, 2*(2), 227-235.

Saputri, R. E., Firmansyah, R., & Silfiya, S. (2024). Pentingnya evaluasi pembelajaran untuk meningkatkan kompetensi peserta didik di sekolah dasar. *Sindoro: Cendikia Pendidikan*, *3*(8), 21-30.

Smith, J. K., Lee, M., & Davis, K. (2022). Integrating real-life scenarios into classroom assessments. *Journal of Educational Innovation, 20*(4), 101-120.

Sholikhah, M., & Hidayati, Y. M. (2024). Summative assessment planning in the kurikulum merdeka on two-dimensional figure materials. *Inovasi Kurikulum, 21*(1), 467-480.

Tarigan, E. F., Nilmarito, S., Islamiyah, K., Darmana, A., & Suyanti, R. D. (2022). Analisis instrumen tes menggunakan Rasch model dan Software SPSS 22.0. *Jurnal Inovasi Pendidikan Kimia*, *16*(2), 92-96.

Wahyuni, A. (2022). Detection of gender biased using DIF (Differential Item Functioning) analysis on item test of school examination Yogyakarta. *Jurnal Evaluasi Pendidikan*, *13*(1), 46-49.

Wang, X., & Zheng, Y. (2023). Improving adaptive testing through psychometric modeling. *International Journal of Educational Technology, 14*(2), 112-126.

Wibowo, S. A., Degeng, M. D. K., & Praherdhiono, H. (2024). Interactive video for learning Mathematics element of measurement in elementary school. *Inovasi Kurikulum, 21*(2), 723-736.

Widodo, H. (2020). Penilaian kontekstual untuk meningkatkan kompetensi numerasi. *Jurnal Pendidikan dan Kebudayaan, 26*(4), 127-140.

Wijayanto, F., Bucur, I. G., Mul, K., Groot, P., van Engelen, B. G., & Heskes, T. (2023). Semi-automated Rasch analysis with differential item functioning. *Behavior Research Methods*, *55*(6), 3129-3148.

Yektiana, N., & Nursikin, M. (2023). Konsep dasar pengukuran, penilaian, dan evaluasi hasil belajar pendidikan agama Islam. *J-Ceki: Jurnal Cendekia Ilmiah*, *2*(2), 263-266.

Yusuf, S., Budiman, N., Yudha, E. S., Suryana, D., & Yusof, S. M. J. B. (2021). Rasch analysis of the Indonesian mental health screening tooals. *The Open Psychology Journal*, *14*(1), 198-203.

https://doi.org/10.17509/jik.v22i1.77242