Jurnal Penelitian dan Evaluasi Pendidikan
Volume 28, No. 2, December 2024 (195-209)
Online: http://journal.uny.ac.id/index.php/jpep

# Item analysis test of science, Indonesian language, and mathematics using the rasch model in elementary schools

**Ernawati Ernawati[1]*; Rini Yaumi Habibah[1]; Nur Syarifah[1]; Firmansyah Firmansyah[1]; Has'ad Rahman Attamimi[2]**
[1]Universitas Muhammadiyah Prof. Dr. HAMKA, Indonesia
[2]STIKES Griya Husada, Sumbawa, Indonesia
*Corresponding Author. E-mail: ernawati.pep@uhamka.ac.id

**ARTICLE INFO**

**Article History**
**Submitted:**
26 June 2024
**Revised:**
08 July 2024
**Accepted:**
01 October 2024

**Keywords**
rasch model; item analysis; quest application

**Scan Me:**

**ABSTRACT**

Assessment of learning outcomes is an important component in the learning process to measure student learning achievements and assist teachers in determining appropriate learning strategies. This study aims to evaluate the quality of test items and student ability levels in Science, Indonesian Language, and Mathematics subjects using the Rasch model. The evaluation includes item reliability, person reliability, and the overall fit of the test items to the model. The research was conducted with 187 students from four public elementary schools in West Jakarta, using a quantitative method with a descriptive design. Data collection involved administering tests to the students, which consisted of 12 items in Science, eight items in Indonesian Language, and 10 items in Mathematics. The data were analyzed using the Quest software to provide comprehensive Rasch analysis results. The findings revealed that the consistency of student responses was weak, but the quality of the test items was good, as evidenced by high item reliability and low person reliability. In terms of model fit, all Science items met the Rasch model criteria, while in the Indonesian Language, one item (item 19) did not fit, and in Mathematics, three items (items 21, 27, and 28) failed to meet the criteria. The analysis of item difficulty levels showed a predominance of medium difficulty. The validity results indicated that 26 items were valid, and four items (1 in Indonesian Language and 3 in Mathematics) were invalid. Most students fell into the medium ability category across all subjects, indicating the need for further tutoring and personalized learning strategies to improve student performance.



**To cite this article (in APA style):**
Ernawati, E., Habibah, R. Y., Syarifah, N., Firmansyah, F., & Attamimi, H. R. (2024). Item analysis test of science, Indonesian language, and mathematics using the rasch model in elementary schools. *Jurnal Penelitian dan Evaluasi Pendidikan, 28(2)*, 195-209 doi: https://doi.org/10.21831/pep.v28i2.75448

## INTRODUCTION

Assessment of learning outcomes is an important component in the learning process to measure student learning achievements and assist teachers in determining appropriate learning strategies. According to the Regulation of the Minister of Education, Culture, Research and Technology of the Republic of Indonesia, Number 21 of 2022, concerning Educational Assessment Standards in Early Childhood Education, Basic Education Levels and Secondary Education Levels, "Assessment is the process of collecting and processing information to determine learning needs and developmental achievements or learning outcomes of students." In the same regulation, specifically article 3, the procedure for assessing student learning outcomes consists of formulating objectives, selecting and/or developing assessment instruments, implementing assessments, processing assessment results, and reporting assessment results (Menteri Pendidikan Kebudayaan Riset dan Teknologi, 2022).

The learning outcomes assessment procedures mentioned above have various challenges. In a study aimed at producing standardized Indonesian and English instrument models for Item Response Theory, preliminary information was obtained that there was a lack of teacher knowledge in developing standardized instruments (Suyata et al., 2014). Furthermore, Istiyono


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v28i2.75448

et al. (2020) revealed that the majority of junior high school teachers in Sleman Regency are still confused about preparing instruments and analyzing items using Item Response Theory. Then, Ernawati et al. (2022) explained that most of the teachers who were respondents had not been able to create items based on the criteria for good items. Therefore, questions cannot be included in the question bank.

Based on several references that have been mentioned, the teacher's lack of knowledge in creating questions can have an impact on the teacher's attitude in presenting subject assessment instruments. Not infrequently, teachers will take shortcuts in creating exam questions. Teachers will take questions from textbooks available on the market. This is in accordance with the findings obtained by Hartini et al. (2021). According to Hartini et al. (2021), the teachers who were respondents to the research admitted that the exam questions given to students were obtained from the students' textbooks. This is considered easy and practical. The use of questions without validation can reduce the quality of the assessment because it does not meet the required validity and reliability criteria (Malonisio & Malonisio, 2023).

The challenges of implementing learning outcome assessment procedures were also illustrated in a study conducted by Setiadi (2016). The study used 330 teachers as respondents (45 SD/MI teachers, 140 SMP/MTs teachers, and 145 SMA/MA and SMK teachers) who are in 15 provinces. In this research, the percentage of teachers who revised instruments at the elementary, middle and high school levels was 53%, 41% and 53%, respectively. Then, the percentage of teachers who chose questions at the elementary, middle and high school levels was 29%, 42% and 29% respectively. Revising the instrument and selecting question items is part of the planning stage in obtaining a valid and reliable assessment instrument.

According to the Pusat Penilaian Pendidikan (2019), in developing valid and reliable instruments, you can follow the rules for preparing standard instruments, namely determining objectives, compiling grids, compiling questions, qualitative analysis, trials and quantitative analysis. In the context of obtaining valid and reliable instruments, teachers can use questions in formative or summative exams in class so that they do not need to spend special time testing questions (Setiadi, 2016). Questions that have been used in the results can be used to fill in report cards and can be analyzed quantitatively so those questions can be saved into a question bank (Setiadi, 2016).

Allen & Yen (1979) explained quantitative analysis of question items includes classical test theory and response theory. Classical test theory is the sum of the actual score and the measurement error score (Retnawati, 2016). The measurement error in question is a random error and does not include systematic measurement error (Retnawati, 2016). There are several weaknesses in this test. First, statistics on the level of difficulty and differentiating power of questions really depend on the sample used in the analysis (Retnawati, 2016). Second, the student scores obtained on a test are very limited to the test used, so the test cannot be generalized outside the test used (Retnawati, 2016). Third, the concept of reliability, which is based on the alignment of test equipment, is very difficult to fulfil (Retnawati, 2016). Fourth, classical test theory does not provide a basis for determining how a test taker will respond when given certain items (Retnawati, 2016). Fifth, the standard error index of measurement is assumed to be the same for each test participant (Retnawati, 2016). Finally, procedures such as item bias testing and test equalization are not practical and difficult to carry out (Retnawati, 2016).

Item response theory has a mathematical model, which means that the probability of a test taker answering an item correctly depends on the participant's ability and the characteristics of the item (Retnawati, 2014). This means that test participants with high ability will have a greater probability of answering correctly than participants with low ability (Retnawati, 2014). In item response theory, there are three basic assumptions, namely unidimensionality, local independence, and parameter invariance. Unidimensional means that each test item only measures one ability (Retnawati, 2014). Local independence is a condition where if the


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v28i2.75448

influencing factors are constant, then the subjects' responses to any pair of items will be statistically independent of each other. Finally, parameter invariance provides information that the characteristics of an item do not depend on the distribution of the test taker's abilities, and conversely, the parameters that characterize the test taker do not depend on the characteristics of an item (Retnawati, 2014).

One part of item response theory is the Rasch model, which is usually called one logistic parameter. The Rasch model assumes that the participant's ability and question difficulty are on the same basic scale. The Rasch model estimates the relationship between items and respondents' true scores, making it possible to evaluate the validity of items and measurement instruments (Awaluddin et al., 2023). Rasch models are capable of analyzing the quality of scoring instruments in various fields, such as music (Wolfs et al., 2023), Natural Science in Elementary School (Maryani et al., 2021), and Calculus in College Students (Taylor et al., 2020). The use of the Rasch model is not only limited to student assessment but has also been applied to assessing teachers' teaching skills (van de Grift et al., 2019).

According to Suseno & Sasongko (2021), the Rasch Model fulfils five criteria for educational assessment measurement, namely 1. providing a linear measure with the same interval, 2. carrying out a precise estimation process, 3. finding items that are inappropriate (misfits) or are not common (outliers), 4. overcome missing data, 5. produce replicable measurements (independent of the parameters studied). Mardapi (2012) explains that evaluation of test instruments using the Rasch model can be done through the following series of steps: First, assess the item fit statistics. This step is important to identify items that fit the Rasch model, and if items that do not fit are found, they can be removed. Second, assess person fit statistics. This stage aims to determine which test takers fit the Rasch model. Finally, the item fit and person fit are determined using the Rasch model through goodness of fit analysis.

Rasch analysis enables a comprehensive approach to several measurement issues, all necessary for the validity of transforming data to an interval scale: testing the internal construct validity of the scale for unidimensionality, which is required for a valid raw (ordinal) score; testing item invariance (i.e., the ratio of difficulties between item pairs remains constant across the respondents' ability levels), which is necessary for interval scaling; appropriate category ordering (whether the order of categories in polytomous items functions as expected); and differential item functioning (DIF; whether bias exists for an item among subgroups in the sample) (Tennant & Conaghan, 2007).

Rasch analysis is used when a set of questionnaire items or items from a given scale are intended to be summed together to provide a total score, including several subscale totals and an overall score. Its applications include: first, in the development of new scales, allowing for the design of items that meet the model's criteria. Second, it reviews the psychometric properties of existing ordinal scales. Third, Rasch's analysis tests hypotheses about the dimensional structure of ordinal scales, including higher-order constructs from different subscales. Fourth, it is used in constructing item banks for computer adaptive testing. Lastly, Rasch analysis is applied whenever change scores need to be calculated from ordinal scales, ensuring data meet model expectations for interval-based estimates. Common software for Rasch analysis includes WINSTEPS, RUMM2020, and ConQuest, each ensuring that observed response patterns match theoretical expectations and verifying data fit (Tennant & Conaghan, 2007).

Item analysis is an important part of educational evaluation. Item analysis aims to determine the quality of items and assessment instruments as a whole. The Rasch model is one of the most widely used item analysis methods and has several advantages over other methods. Previous studies have shown that the Rasch Model can help improve the quality of assessment instruments and improve student learning outcomes. This research is important to provide an overview of the quality of learning outcomes assessment instruments in public primary schools in West Jakarta and help in improving the quality of assessment and student learning outcomes.


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v28i2.75448

## RESEARCH METHOD

### Instrument Spesification

This study uses a quantitative method with a descriptive design to analyze the quality of test items created by teachers. The research was conducted with 187 students from four public elementary schools in West Jakarta. The students participating in the study are from grades V and VI of elementary school. The test trial was conducted from April to Mei 2023.

The instrument used in this study is a learning outcome test covering subjects in science, Indonesian language, and mathematics. The instrument tested consists of 30 multiple-choice items, including 12 items in Science, eight items in Indonesian Language, and 10 items in Mathematics. The following indicators were used in the item specifications: for Science, the indicators include determining the appropriate material to prevent echo, plant reproduction methods, reasons fishermen go to sea, the process of dry ice sublimation, reasons plants survive in dry places, magnetic properties, the apparent daily motion of the sun, examples of mutualistic symbiosis, examples of evaporation events, the effect of force on motion, determining types of animal food based on tables, and the impact of changes in the paddy field ecosystem. For the Indonesian Language, the indicators include determining question sentences according to the paragraph content, determining invitation sentences in speeches, main ideas in texts, the moral of poems, differences between poems and prose, suitable advertisement sentences, and main ideas in texts. For Mathematics, the indicators include results of mixed arithmetic operations, calculating the area of a circle, properties of solid figures, results of fraction operations, elements of a circle, determining the number of vases, calculating the volume of water in an aquarium, the number of longan trees based on ratios, the number of items sold from diagram data, and determining departure times.

### Data Analysis

The research procedure began with data collection through the administration of tests to all students in the sample. After the data was collected, analysis was conducted using the Quest software. Adams & Kho (1996) explain that Quest is an application that offers comprehensive Rasch analysis, including item estimation, case estimation, and fit statistics. The results of the analysis in the Quest application can be accessed through various informative tables and maps. Additional analyses are available, including item and person reliability.

Data analysis using the Rasch model is divided into six parts, namely 1. Validity Instrument, 2. Estimating reliability, 3. Estimating the suitability of items and persons, 4. Estimating passing items, 5.Estimating the level of difficulty, 6.Estimating the ability level of test participants (Dewi et al., 2023; Hanna & Retnawati, 2022; Pratama, 2020). Through analysis using the Rasch model with the Quest program, this study is expected to provide practical recommendations for teachers in developing valid and reliable assessment instruments before they are used in actual exams.

## FINDINGS AND DISCUSSION

### Findings

#### *Realibility Estimation*

In estimating reliability, there are reliability criteria presented in Table 1 below (Sumintono & Widhiarso, 2014):


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v28i2.75448

Table 1. Person Reliability Criteria and Item Reliability

| Realibility Score | Category |
|---|---|
| <0.67 | Weak |
| 0.67-0.80 | Avarage |
| 0.81-0.90 | Good |
| 0.91-0.94 | Very good |
| >0.94 | Excellent |

A reliability analysis was conducted to measure the reliability level of the research instrument used to measure variables. The results of the reliability analysis are presented in Table 2. The item reliability scores for each subject are categorized as excellent, with values of 0.96 for Science, 0.95 for Indonesian Language, and 0.93 for Mathematics. In contrast, the person reliability scores fall into the weak category, with values of 0.47 for Science, 0.26 for Indonesian Language, and 0.59 for Mathematics. In general, Cronbach's Alpha reliability coefficients for the three subjects, namely Science, Indonesian Language, and Mathematics, show fairly good values. This indicates that the items in these research instruments consistently measure the same construct.

Table 2. The Result of Person Reliability and Item Reliability in Each Subject

| Subject | Estimates | Mean | SD | SD (adj) | Reliability |
|---|---|---|---|---|---|
| Science | Item | 0.00 | 0.94 | 0.93 | 0.96 |
| | Case | 0.42 | 0.95 | 0.65 | 0.47 |
| Indonesia Languange | Item | 0.00 | 0.77 | 0.75 | 0.95 |
| | Case | 0.71 | 1.02 | 0.52 | 0.26 |
| Mathematics | Item | 0.00 | 0.69 | 0.67 | 0.93 |
| | Case | 0.11 | 1.22 | 0.94 | 0.59 |

### *Item Fit Estimation*

a. Infit Mean Square Criteria

In determining the suitability of the model for each item, the infit mean square value for each item can be matched with the criteria in Table 3 below (Setyawarno, 2017).

Table 3. Infit Mean Square Criteria

| Infit MNSQ | Interpretation |
|---|---|
| >1.33 | Misfit |
| 0.77-1.33 | Fit |
| <0.77 | Misfit |

The results of the item fit analysis for science, Indonesian and mathematics subjects are summarized in Table 4. Based on the analysis using Infit Mean Square (MNSQ) values, it is evident that most test items in the Science, Indonesian Language, and Mathematics subjects fall within the acceptable fit range of 0.77 to 1.33. All test items in Science show a good fit within this range. In the Indonesian Language subject, although all items fall within the Infit Mean Square range, one item (item 19) requires attention. Despite its Infit Mean Square value being within the range, it is very close to the boundary and was deemed invalid in the validity results. Therefore, item 19 needs improvement. In Mathematics, items 21 and 28 are categorized as misfits, with MNSQ values of 1.34 and 0.73, respectively. Additionally, item 27 slightly exceeds the lower criterion of 0.77. Thus, items 21, 27, and 28 in Mathematics require revision.


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v28i2.75448

Table 4. Infit Mean Square Estimation Results for Each Subject

| Science | | | Indonesian Language | | | Mathematic | | |
|---|---|---|---|---|---|---|---|---|
| Items | Infit MNSQ | Criterion Infit MNSQ | Items | Infit MNSQ | Criterion Infit MNSQ | Items | Infit MNSQ | Criterion Infit MNSQ |
| 1 | 0.95 | Fit | 11 | 0.94 | Fit | 21 | 1.34 | Misfit |
| 2 | 0.96 | Fit | 12 | 1.01 | Fit | 22 | 1.2 | Fit |
| 3 | 1.06 | Fit | 13 | 1.01 | Fit | 23 | 1.08 | Fit |
| 4 | 1.00 | Fit | 14 | 0.92 | Fit | 24 | 0.84 | Fit |
| 5 | 0.94 | Fit | 15 | 0.90 | Fit | 25 | 1.09 | Fit |
| 6 | 0.96 | Fit | 16 | 0.96 | Fit | 26 | 1.20 | Fit |
| 7 | 0.99 | Fit | 17 | 1.30 | Fit | 27 | 0.79 | Fit |
| 8 | 0.99 | Fit | 18 | 0.94 | Fit | 28 | 0.73 | Misfit |
| 9 | 0.93 | Fit | 19 | 0.94 | Misfit | 29 | 0.91 | Fit |
| 10 | 1.16 | Fit | 20 | 1.01 | Fit | 30 | 0.84 | Fit |
| 11 | 1.00 | Fit | | | | | | |
| 12 | 1.03 | Fit | | | | | | |

b. Outfit t criteria

To find out whether an item has passed or failed, the outfit t value of each item can be matched with the criteria listed in Table 5 below (Setyawarno, 2017)

Table 5. Outfit t Criteria

| Oufit t | Interpretation |
|---|---|
| -2.0 ≤ Outfit t ≤ 2.0 | Passed |
| Outfit t < -2.00 atau Outfit t > 2.00. | Failed |

The results of the analysis of passing and failing item estimates in the subjects of Science , Indonesian Language, and Mathematics, are summarized in Table 6 below.

Table 6. Passing and Failing Items in Each Subject

| Science | | | Indonesia Languange | | | Mathematics | | |
|---|---|---|---|---|---|---|---|---|
| Items | Outfit t | Criterion Outfit t | Items | Outfit t | Criterion Outfit t | Items | Outfit t | Criterion Outfit t |
| 1 | -0.50 | Passed | 13 | -0.60 | Passed | 21 | 2.70 | Failed |
| 2 | -0.60 | Passed | 14 | -0.40 | Passed | 22 | 1.90 | Passed |
| 3 | 0.50 | Passed | 15 | 0.10 | Passed | 23 | 0.20 | Passed |
| 4 | 0.20 | Passed | 16 | -0.30 | Passed | 24 | -1.50 | Passed |
| 5 | -0.50 | Passed | 17 | -1.10 | Passed | 25 | 1.10 | Passed |
| 6 | -0.40 | Passed | 18 | -0.70 | Passed | 26 | 1.30 | Passed |
| 7 | 0.00 | Passed | 19 | 3.80 | Failed | 27 | -2.30 | Failed |
| 8 | 0.70 | Passed | 20 | -1.10 | Passed | 28 | -2.70 | Failed |
| 9 | -0.80 | Passed | | | | 29 | -0.10 | Passed |
| 10 | 1.80 | Passed | | | | 30 | -1.80 | Passed |
| 11 | -0.30 | Passed | | | | | | |
| 12 | 0.90 | Passed | | | | | | |

Based on the analysis using Outfit t values, all items in the Science subject met the Rasch model criteria, with Outfit t values ranging from -0.80 to 1.80. In the Indonesian Language subject, 1 out of 8 items did not meet the Outfit t criteria. The Outfit t values for the items in the Indonesian Language that met the criteria ranged from -0.70 to 0.10. Additionally, 3 out of


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v28i2.75448

10 items in the Mathematics subject did not meet the Outfit t criteria. The Outfit t values for Mathematics ranged from -1.80 to 1.90.

### *Item Validity*

Item validity can be determined by identifying those items that fall into the category of misfit. The Indonesian language item 19 showed a significant misfit, as indicated by its elevated infit mean square and outfit t values. This suggests that the item may have compromised the overall reliability and validity of the assessment. Similarly, the Mathematics items 21, 27, and 28 also exhibited misfit, warranting further investigation and potential item revision or removal. The summary of the invalid items is as follows (Table 7).

Table 7. Summary of Invalid Items

| Subject | Infit Mean Square | Outfit t |
|---|---|---|
| Science | 0 | 0 |
| Indonesian Languange | 1.30 (Item 19) | 3.80(Item 19) |
| Mathematics | 1.34 (Item 21), 0.79 (item 27), 0.73 (Item 28) | 2.70 (Item 21), -2.30 (Item 27), -2.70 (Item 28) |

### *Item Difficulty Level*

To determine the difficulty level of items, the threshold value for each item can be matched with the criteria listed in Table 8 (Setyawarno, 2017).

Table 8. Criteria for Item Difficulty Level

| Treshold Value | Interpretation |
|---|---|
| b>2 | Very difficult |
| 1<b≤2 | Difficult |
| -1≤b≤1 | Moderate |
| -1>b≥-2 | Easy |
| b<-2 | Very Easy |

Table 9. Results of Item Difficulty Levels in Each Subject

| Science | | | Indonesia Language | | | Mathematics | | |
|---|---|---|---|---|---|---|---|---|
| Item | Threshold Value | Interpretation | Item | Threshold Value | Interpretation | Item | Threshold Value | Interpretation |
| 1 | 0.53 | Moderate | 13 | -0.81 | Easy | 21 | -0.12 | Moderate |
| 2 | 0.44 | Moderate | 14 | 0.05 | Moderate | 22 | 1.73 | Difficult |
| 3 | 0.48 | Moderate | 15 | -0.26 | Moderate | 23 | 0.15 | Moderate |
| 4 | 1.90 | Difficult | 16 | -0.43 | Moderate | 24 | -0.82 | Moderate |
| 5 | 0.99 | Moderate | 17 | -0.61 | Moderate | 25 | -0.01 | Moderate |
| 6 | -0.67 | Moderate | 18 | 0.17 | Moderate | 26 | -0.25 | Moderate |
| 7 | -0.29 | Moderate | 19 | 1.68 | Difficult | 27 | 0.26 | Moderate |
| 8 | 0.08 | Moderate | 20 | 0.20 | Moderate | 28 | -0.22 | Moderate |
| 9 | -0.22 | Moderate | | | | 29 | -0.09 | Moderate |
| 10 | -0.79 | Moderate | | | | 30 | -0.64 | Moderate |
| 11 | -1.60 | Easy | | | | | | |
| 12 | -0.85 | Moderate | | | | | | |


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v28i2.75448

The analysis of item difficulty levels revealed variations across subjects. Science items consisted of 1 easy item, 10 items of moderate difficulty, and 1 difficult item. Indonesian language items comprised 1 easy item, 6 items of moderate difficulty, and 1 difficult item. Mathematics items, on the other hand, included 9 item of moderate difficulty and 1 difficult item. The following is the item difficulty level table (Table 9).

### *Student Ability*

The analysis of ability levels has estimate value criteria as shown in Table 10 (Setyawarno, 2017).

Table 10. Criteria for Student Ability Using the Rasch Model

| Nilai Estimate | Description |
|---|---|
| >1.00 | High Ability |
| -1.00 s.d +1.00 | Medium Ability |
| <-1.00 | Low Ability |

Based on the Quest output, student abilities can be assessed using the estimated ability values, which are divided into three categories: high (>1.00), medium (-1.00 to 1.00), and low (<-1.00). In the Science subject, high ability estimates range from 1.28 to 2.73, medium ability from -0.82 to 0.8, and low ability from unreadable to -1.28. In the Indonesian Language subject, high ability estimates range from 1.2 to perfect, medium ability from -0.59 to 0.54, and low ability from -2.11 to -1.22. In the Mathematics subject, high ability estimates range from 1.5 to perfect, medium ability from -0.94 to 0.9, and low ability from unreadable to -1.5. The summary of the abilities of 187 students in this study is available in Table 11.

Table 11. Summary of Respondent's Abilities

| Ability Level | Science | | | | Indonesian Language | | | | Mathematics | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Item Estimate** | **Total Correct Answers** | **Total Students** | **%** | **Item Estimate** | **Total Correct Answers** | **Total Students** | **%** | **Item Estimate** | **Total Correct Answers** | **Total Students** | **%** |
| Low | No score | 0 | 1 | | No score | 0 | 0 | | No score | 0 | 4 | |
| | -2.7 | 1 | 1 | | -2.11 | 1 | 2 | | -2.34 | 1 | 8 | |
| | -1.86 | 2 | 4 | | -1.22 | 2 | 12 | | -1.5 | 2 | 20 | |
| | -1.28 | 3 | 3 | | | | | | | | | |
| | Total Students | | 9 | 5 | Total Students | | 14 | 7 | Total Students | | 32 | 17 |
| Moderate | -0.82 | 4 | 14 | | -0.59 | 3 | 16 | | -0.94 | 3 | 23 | |
| | -0.41 | 5 | 26 | | -0.03 | 4 | 27 | | -0.47 | 4 | 19 | |
| | -0.01 | 6 | 28 | | 0.54 | 5 | 43 | | -0.03 | 5 | 19 | |
| | 0.38 | 7 | 38 | | | | | | 0.41 | 6 | 29 | |
| | 0.8 | 8 | 19 | | | | | | 0.9 | 7 | 24 | |
| | Total Students | | 125 | 67 | Total Students | | 86 | 46 | Total Students | | 114 | 61 |
| High | 1.28 | 9 | 31 | | 1.2 | 6 | 49 | | 1.5 | 8 | 22 | |
| | 1.86 | 10 | 18 | | 2.15 | 7 | 37 | | 2.39 | 9 | 14 | |
| | 2.73 | 11 | 4 | | Perfect | 8 | 1 | | Perfect | 10 | 5 | |
| | Perfect | 12 | 0 | | | | | | | | | |
| | Total Person | | 53 | 28 | Total Student | | 87 | 47 | Total Student | | 41 | 22 |


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v28i2.75448

## Discussion

### *Reliability Estimation*

High item reliability indicates that the test items fit the model well, ensuring that most items are appropriate and provide the expected information. On the other hand, low person reliability reflects inconsistency in student responses, where high-ability students sometimes answer incorrectly, and low-ability students sometimes answer correctly. This highlights the need to improve student performance consistency to achieve more reliable assessment results (Setyawarno, 2017).

The difference between high item reliability and low person reliability suggests that while the quality of the test items is good, the consistency of student responses is weak (Sumintono & Widhiarso, 2014). Several factors can influence reliability levels, including the number of items, score distribution, item difficulty level, objectivity, the psychological and health conditions of students, teacher/grader subjectivity in scoring, and having too many items, leading to fatigue and haste in answering (Sumardi, 2020). These factors should be considered for improving or revising the test items.

### *Item Fit Estimation*

In the Science subject, the outfit t values for 12 items show good results, with all items meeting the "Passed" criteria. The outfit t values range from -0.80 to 1.80, indicating that these items function according to the Rasch model without significant anomalies. This is consistent with Azizah and Supahar's (2023) study, which analyzed the quality of Physics questions for 10th grade. In her model fit analysis results, 18 questions were found to fit the Rasch model, and two did not. The Rasch model fit criteria in Azizah & Supahar (2023) study were an infit mean square value between 0.77 and 1.33 and an outfit t value ranging from -2 to 2.

Additionally, Maulana et al.'s (2023) research showed that 20 diagnostic Science items met the Rasch model criteria. Ngadi's (2023) study on the analysis of questions on the subject of motorcycle electrical system maintenance showed that 30 items met the Rasch model criteria. The Rasch model criteria used by Maulana et al. (2023) and Ngadi (2023) were: 1) Outfit Mean Square (MNSQ): $0.5 < MNSQ < 1.5$, 2) Outfit Z.Standard (ZSTD): $-2.0 < ZSTD < +2.0$, and 3) Point Measure Correlation (Pt Mean Corr): $0.4 < Pt\ Mean\ Corr < 0.85$.

In the Indonesian Language subject, although all items fell within the Infit Mean Square range, one item (item 19) requires special attention. The infit mean square value for item 19 is within the acceptable range but very close to the boundary. Therefore, item 19 needs improvement. Items meeting the outfit t criteria < 2.00 align with Anggraini and Suyata's (2014) study on the characteristics of the UASBN for the Indonesian Language subject. Their Rasch model analysis with Bigstep found that 49 items (98%) fit the Rasch model (outfit < 2.00), and 1 item (2%) did not fit the model (outfit t > 2.00).

In the Mathematics subject, items 21 and 28 fall into the misfit category with MNSQ values of 1.34 and 0.73, respectively. Additionally, item 27 requires attention as its MNSQ value slightly exceeds the lower limit of 0.77. Therefore, items 21, 28, and 27 require improvement. Items meeting the outfit t and infit mean square criteria align with Rahmayani et al.'s (2022) study on the analysis of 4th-grade Mathematics questions. In their study, 10 items met the infit mean square and outfit t criteria.

Furthermore, Hanna & Retnawati's (2022) study on the quality of Mathematics questions in 10th-grade science classes showed that out of 12 items, two items did not fall within the infit mean square and outfit t range (items 1 and 3), while 10 items met the criteria. The model fit criteria used by Rahmayani et al. (2022) and Hanna & Retnawati (2022) were an infit mean square of 0.77 to 1.33 and an outfit t of -2 to 2.


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v28i2.75448

Finally, Mutakin et al.'s (2022) study on the analysis of 4th-grade Mathematics questions showed that out of 47 items, 26 items fit the model, and 21 items did not. The model fit criteria used in their study were: 1) Outfit Mean Square (MNSQ): 0.5 < MNSQ < 1.5, 2) Outfit Z.Standard (ZSTD): -2.0 < ZSTD < +2.0, and 3) Point Measure Correlation (Pt Mean Corr): 0.4 < Pt Mean Corr < 0.85.

Based on the output results of Infit mean square and outfit t, the following is a summary of items (Table 12) that need improvement and do not meet the Infit mean square and outfit t criteria.

Table 12. Summary of Item Revision

| Subject | Jumlah Item | Nomor Item |
|---|---|---|
| Science | 0 | 0 |
| Indonesian Languange | 1 | 19 |
| Mathematics | 3 | 21, 27, 28 |

The item fit analysis yields important implications for the research. Firstly, it provides critical insights into the quality of the research instrument. A high number of misfit items can significantly compromise the instrument's reliability and validity. Secondly, this analysis necessitates a thorough review and potential revision of the instrument. Misfit items should be identified and subjected to further scrutiny, which may lead to reformulating questions, removing problematic items, or introducing new ones to enhance the instrument's effectiveness. Lastly, the results of the item fit analysis play a crucial role in the interpretation of overall research findings. In cases where numerous items are found to be misfit, researchers must exercise caution when drawing conclusions from the data analysis. This careful approach ensures that the research outcomes are interpreted within the context of the instrument's limitations, maintaining the integrity and accuracy of the study's conclusions.

### *Item Validity*

The validity of tests in the Rasch model can be seen from model fit (Azizah & Supahar, 2023). The criteria for the Rasch model can be observed from the infit mean square and outfit t values. In this study, the infit mean square criteria use a range of 0.77 to 1.33, and the outfit t criteria use a range of -2 to 2 (Setyawarno, 2017). Items that fall within these ranges are considered valid, while items that do not fit the model are considered invalid. According to the summary of the infit mean square and outfit t output for Science, Indonesian Language, and Mathematics items, there are 4 items that do not meet these criteria. Therefore, these four items are considered invalid and need improvement.

### *Item Difficulty Level*

Based on the threshold values for the Science subject, there are eight items (80%) with a medium difficulty level, one item (10%) with a difficult level, and one item (10%) with an easy level. This is consistent with the study by Ngadi (2023), which shows the distribution of questions based on difficulty levels with 70% in the medium category, 20% difficult, and 10% easy out of 30 questions. The item difficulty criteria in Ngadi's (2023) study are: a) Measure logit > 1: Very difficult, b) 0.5 < Measure logit < 1: Difficult, c) -0.5 < Measure logit < 0.5: Medium, d) -0.5 < Measure logit < -1: Easy, e) Measure logit < -1: Very easy.

The study by Maulana et al. (2023) found that out of 20 diagnostic Science test questions, there are 2 difficult items (items 20 and 18), 3 easy items (items 6, 7, and 8), and 15 medium items. The difficulty levels in Maulana et al.'s (2023) study use the criteria: a) Measure greater than mean + S.D.: Difficult, b) Measure between mean - S.D. and mean + S.D.: Medium, c) Measure less than mean - S.D.: Easy. Azizah & Supahar (2023) also shows an ideal distribution


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v28i2.75448

of difficulty levels with 5% very easy, 25% easy, 40% medium, and 30% difficult. The difficulty criteria in Azizah & Supahar (2023) study are threshold > 2 very difficult, 1 < threshold < 2 difficult, -1 < threshold < 1 medium, -1 > threshold > -2 easy, and threshold < -2 very easy.

In the Indonesian Language subject, there are eight items (80%) with a medium difficulty level, one item (10%) with a difficult level, and one item (10%) with an easy level. This aligns with the study by Anggraini & Suyata (2014), which shows the difficulty levels of Indonesian Language questions with 2 easy questions (4.08%), 44 medium questions (89.9%), and 3 difficult questions (6.12%). The difficulty criteria in this study are Easy < -2.00, Medium -2.00 – 2.00, Difficult > 2.00. The study by Azizah et al. (2021) on Indonesian Language exam questions at PKN STAN shows there are 1 very difficult item, 12 difficult items, 11 easy items, and 1 very easy item. The difficulty levels in this study are measured with measure value < -1 very easy, -1 to 0 easy, 0 to 1 difficult, and > 1 very difficult.

In the Mathematics subject, there are nine items (90%) with a medium difficulty level and one item (10%) with a difficult level. The study by Mutakin et al. (2022) shows the distribution of mathematics difficulty levels as follows: X > M + 2SD (very difficult), M + 2SD ≤ X < M + 1SD (difficult), M + 1SD ≤ X < 0 (medium), 0 ≤ X < M - 1SD (easy), X < M - 1SD (very easy). Out of 26 questions that fit the Rasch model, there are 2 very difficult items, 0 difficult items, 13 medium items, 10 easy items, and 1 very easy item.

The studies by Rahmayani et al. (2022) and Hanna & Retnawati (2022) use the criteria b > 2 very difficult, 1 < b < 2 difficult, -1 < b < 1 medium, -1 > b > -2 easy, and b < -2 very easy. Rahmayani et al. (2022) found 2 difficult items, 1 very easy item, and 7 medium items. Meanwhile, Hanna & Retnawati (2022) found 4 difficult items (33.3%), 4 medium items (33.33%), 3 easy items (25%), and 1 very easy item (8.3%). The difficulty levels in Hanna & Retnawati's study show an ideal composition.

Among the three subjects, the composition of easy, medium, and difficult questions does not yet meet the recommended composition. According to Kunandar (2015), the ideal composition of questions consists of 25% easy items, 50% medium items, and 25% difficult items. Several studies that have an ideal composition include Azizah & Supahar (2023) for the Science subject, Azizah et al. (2021) for the Indonesian Language subject at PKN STAN, and the study by Hanna & Retnawati (2022) for the Mathematics subject. The summary of the difficulty levels for each subject is available in the Table 13.

Table 13. Summary of Item Difficulty Levels

| Levels | Science | | | Indonesia Languange | | | Mathematics | | |
|---|---|---|---|---|---|---|---|---|---|
| | Item Count | Item Number | % | Item Count | Item Number | % | Item Count | Item Number | % |
| Very difficult | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Difficult | 1 | 4 | 8 | 1 | 19 | 13 | 1 | 22 | 10 |
| Moderate | 10 | 1,2,3,5,6, 7,8,9,10,1 2 | 83 | 7 | 13,14,15,16,1 7,18,20 | 88 | 9 | 21,23,24,25 ,26,27,28,2 9,30 | 90 |
| Easy | 1 | 11 | 8 | 0 | 0 | 0 | 0 | 0 | 0 |

### *Student Ability*

Based on the Quest output, student abilities can be assessed using the estimated ability values, which are divided into three categories: high (>1.00), medium (-1.00 to 1.00), and low (<-1.00). In the Science subject, high ability estimates range from 1.28 to 2.73, medium ability


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v28i2.75448

from -0.82 to 0.8, and low ability from unreadable to -1.28. In the Indonesian Language subject, high ability estimates range from 1.2 to perfect, medium ability from -0.59 to 0.54, and low ability from -2.11 to -1.22. In the Mathematics subject, high ability estimates range from 1.5 to perfect, medium ability from -0.94 to 0.9, and low ability from unreadable to -1.5.

In the Science subject, nine students (5%) are in the high ability category, 125 students (67%) are in the medium ability category, and 53 students (28%) are in the low ability category. This aligns with Maulana et al.'s (2023) study, which showed that in the diagnostic test for Science, there were 17 participants with high ability, 14 participants with low ability, and 52 participants with medium ability. Ngadi's (2023) study shows that in the Electrical System Maintenance subject test, out of 49 students, 44.9% had a high ability with logits from 2.85 to 4.83, 46.94% had a medium ability (-1.78 to 1.33 logits), 4.08% had the low ability (-2.42 logits), and 4.08% had the very low ability (-3.65 logits).

In the Indonesian Language subject, 14 students (7%) are in the high ability category, 86 students (46%) are in the medium ability category, and 87 students (47%) are in the low ability category. This is consistent with Anggraini & Suyata's (2014) study, which showed that the average ability level of participants in answering 50 UASBN questions for the Indonesian Language subject was high, with a value of 1.92 (>1.00).

Meanwhile, in the Mathematics subject, 32 students (17%) are in the high ability category, 114 students (61%) are in the medium ability category, and 41 students (22%) are in the low ability category. This aligns with Mutakin et al.'s (2022) study, which shows that ability levels are divided into eight levels: $X > M + 3SD$ (Special), $M + 3SD \leq X < M + 2SD$ (Very High), $M + 2SD \leq X < M + 1SD$ (High), $M + 1SD \leq X < 0$ (Medium), $0 \leq X < M - 1SD$ (Low), $M - 1SD \leq X < M - 2SD$ (Very Low), and $X < M - 3SD$ (Minimum). The study found four students with special abilities: six very high, 12 high, 77 medium, 102 low, 29 very low, and three minimum. The studies by Rahmayani et al. (2022) and Hanna & Retnawati (2022) categorized ability into three parts: high (estimate >1.00), medium (estimate between -1.00 and 1.00), and low (estimate < -1.00). Rahmayani et al.'s (2022) study showed that the math ability levels from 10 items and 152 students were 23% high ability, 64% medium ability, and 13% low ability.

Based on this ability data, it can be concluded that the majority of students fall into the medium ability category across all three subjects. This is also supported by several pieces of literature that show a large number of students with medium ability levels. However, the related literature has different classifications of ability levels.

The data reveals a significant variation in students' abilities across the three subjects. This implies a need for a more flexible teaching approach that can accommodate diverse levels of ability. In particular, mathematics requires special attention, as a considerable number of students still struggle with understanding fundamental mathematical concepts.

On the other hand, the data also suggests a great potential for improvement in students' academic achievement. The significant proportion of students with average abilities in all subjects indicates that with the right teaching strategies, these students can reach higher levels of understanding. Therefore, it is essential to develop learning programs that can cater to a wide range of abilities and provide appropriate challenges to push students to reach their full potential.

## CONCLUSION

Based on the data presented, the quality of the test items in this study can be categorized into four main aspects. First, the consistency of student responses showed weak values, but the overall quality of the test items was good, as indicated by the low person reliability and high item reliability. Second, the fit with the Rasch model revealed that the Science subject met the Rasch model criteria for all items. However, in the Indonesian Language subject, one item (item 19) did not fit the model, and in Mathematics, three items (items 21, 27, and 28) were found to


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v28i2.75448

be misaligned. Therefore, these items require revision or replacement. Third, the difficulty level of the questions was predominantly medium, suggesting that the composition of the items is not yet optimal. Finally, the student ability levels also leaned toward the medium range, indicating that many students achieved less than 50% correct answers. This calls for additional tutoring for students with medium ability levels to improve their performance.

## ACKNOWLEDGMENTS

This research was possible because of research grant support obtained from the Research and Development Institute, Universitas Muhammadiyah Prof.Dr.HAMKA. The author would like to thank the schools who were willing to be our research subjects. We would also like to thank the Postgraduate School of Universitas Muhammadiyah Prof.Dr.HAMKA, especially Educational Research and Evaluation Study Program department, for supporting and facilitating this research.

### Conflict of interests

There are no known conflicts of interest associated with this publication.

## REFERENCES

Allen, M. J., & Yen, W. M. (1979). Introduction to Measurement Theory. Brooks/Cole.

Adams, R. J., & Khoo, S.-T. (1996). Acer Quest: The Interactive Test Analysis System. In L. J. P. Service (Ed.), Australian Council for Educational Research. The Australian Council for Educational Research.

Anggraini, D., & Suyata, P. (2014). Karakteristik soal UASBN mata pelajaran bahasa indonesia di Daerah Istimewa Yogyakarta pada tahun pelajaran 2008/2009. Jurnal Prima Edukasia, 2(1), 57–65. https://journal.uny.ac.id/index.php/jpe/article/view/2644/2199

Awaluddin, T., Wahyuni, L. D., & Afriadi, B. (2023). Konstruksi Alat Ukur dalam Pendidikan. UNJ Press.

Azizah, I., & Supahar. (2023). Analisis kualitas butir soal penilaian harian bersama Ifisika kelas X SMA Negeri 1 Patikraja. Jurnal Pendidikan Fisika, 10(02), 90–104. https://journal.student.uny.ac.id/ojs/index.php/pfisika/index

Azizah, N., Suseno, M., & Hayat, B. (2021). Item Analysis of the Rasch Model Items in the Final Semester Exam Indonesian Language Lesson. World Journal of English Language, 12(1), 15. https://doi.org/10.5430/wjel.v12n1p15

Dewi, H. H., Damio, S. M., & Sukarno, S. (2023). Item analysis of reading comprehension questions for English proficiency test using Rasch model. REID (Research and Evaluation in Education), 9(1), 24–36. https://doi.org/10.21831/reid.v9i1.53514

Ernawati, E., Manik, F. Y., Trisnawati, R. D., Emiliana, E., & Yuliawati, S. (2022). Understanding and quality of minimum competency assessment (AKM) questions made by Integrated Science teachers in junior high schools. Jurnal Penelitian Dan Evaluasi Pendidikan, 26(2). https://doi.org/10.21831/pep.v26i2.48670

Hanna, W. F., & Retnawati, H. (2022). Analisis kualitas butir soal matematika menggunakan model rasch dengan bantuan software QUEST. AKSIOMA: Jurnal Program Studi Pendidikan Matematika, 11(4), 3695. https://doi.org/10.24127/ajpm.v11i4.5908


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v28i2.75448

Hartini, P., Setiadi, H., & Ernawati, E. (2021). Cognitive domain analysis (LOTS and HOTS) assessment instruments made by primary school teachers. Jurnal Penelitian Dan Evaluasi Pendidikan, 25(1). https://doi.org/10.21831/pep.v25i1.34411

Isitiyono, E., Setiawan, R., & Harun. (2020). Pelatihan Penyusunan Instrumen Tes dan Analisisnya Secara Modern Bagi Guru-Guru IPA SMP Training of Test Instrument Development and Its Analysis for Modern Teachers of SMP. J. Pengabdian Masyarakat MIPA Dan Pendidikan MIPA, 4(1), 102–108. http://journal.uny.ac.id/index.php/jpmmp

Kunandar. (2015). Penilaian Autentik (Penilaian Hasil Belajar Peserta Didik Berdasarkan Kurikulum 2013) : Suatu Pendekatan Praktis Disertai dengan Contoh. Rajawali Press.

Malonisio, M. O., & Malonisio, C. C. (2023). Validation of the teacher education institution's entrance test using the Rasch model. International Journal of Innovative Research and Scientific Studies, 6(3), 644–655. https://doi.org/10.53894/ijirss.v6i3.1726

Mardapi, D. (2012). Pengukuran Penilaian dan Evaluasi Pendidikan. Nuha Medika.

Maryani, I., Prasetyo, Z. K., Wilujeng, I., Purwanti, S., & Fitrianawati, M. (2021). HOTs Multiple Choice and Essay Questions: A Validated Instrument to Measure Higher-order Thinking Skills of Prospective Teachers. Turkish Journal of Science Education, 18(4), 674–690. https://doi.org/10.36681/tused.2021.97

Maulana, S., Rusilowati, A., Nugroho, S. E., & Susilaningsih, E. (2023). Implementasi Rasch Model dalam Pengembangan Instrumen Tes Diagnostik. Prosiding Seminar Nasional Pascasarjana Universitas Negeri Semarang, 748–757. http://pps.unnes.ac.id/pps2/prodi/prosiding-pascasarjana-unnes748

Menteri Pendidikan Kebudayaan Riset dan Teknologi. (2022). Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia tentang Standar Penilaian Pendidikan pada Pendidikan Anak Usia Dini, Jenjang Pendidikan Dasar, dan Jenjang Pendidikan Menengah. https://jdih.kemdikbud.go.id/detail_peraturan?main=3104

Mutakin, T. Z., Tola, B., & Hayat, B. (2022). Rasch model to analyze item quality and ability of fourth elementary school students. International Journal of Innovation, Creativity and Change, 16(2), 2022. https://www.ijicc.net/images/Vol_16/Iss2/16172_Mutakin_2022_E1_R.pdf

Ngadi. (2023). Analisis model rasch untuk mengukur kompetensi pengetahuan siswa SMKN 1 Kalianget pada mata pelajaran perawatan sistem kelistrikan sepeda motor. Jurnal Pendidikan Vokasi Otomotif, 6(1), 1–20. https://journal.uny.ac.id/index.php/jpvo/article/view/63479

Pratama, D. (2020). Analisis kualitas tes buatan guru melalui pendekatan item response theory (IRT) Model Rasch. Tarbawy : Jurnal Pendidikan Islam, 7(1), 61–70. https://doi.org/10.32923/tarbawy.v7i1.1187

Pusat Penilaian Pendidikan. (2019). Panduan penilaian tes tertulis. Kementerian Pendidikan dan Kebudayaan Republik Indonesia. https://repositori.kemdikbud.go.id/18344/1/PANDUAN%20PENILAIAN%20TERTULIS%202019.pdf

Rahmayani, A., Tiurlina, & Alfarisa, F. (2022). Analisis kualitas butir soal ulangan harian matematika di kelas IV MI Al-Islamiyah menggunakan Rasch Model. Jurnal PERSEDA, V(3), 170–177. https://doi.org/https://doi.org/10.37150/perseda.v5i3.1716


ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v28i2.75448

Retnawati, H. (2014). Teori Respons Butir dan Penerapannya : Untuk Peneliti, Praktisi Pengukuran dan Pengujian Mahasiswa dan Pascasarjana. Nuha Medika.

Retnawati, H. (2016). Analisis Kuantitatif Instrumen Penelitian (Panduan Penelitian, Mahasiswa, dan Psikometrian). Parama Publishing.

Setiadi, H. (2016). Pelaksanaan penilaian pada Kurikulum 2013. Jurnal Penelitian Dan Evaluasi Pendidikan, 20(2), 166–178. https://doi.org/10.21831/pep.v20i2.7173

Setyawarno, D. (2017). Upaya Peningkatan Kualitas Butir Soal dengan Analisis Aplikasi QUEST. https://www.scribd.com/document/619903922/PPM-Panduan-Quest

Sumardi. (2020). Teknik Pengukuran dan Penilaian. Deepublishing.

Sumintono, B., & Widhiarso, W. (2014). Aplikasi Model Rasch untuk Penelitian Ilmu-Ilmu Sosial (B. Trim (ed.)). Tri Komunikata Publishing House.

Suseno, E., & Sasongko, P. (2021). Mengukur Validitas Tes (E. Suseno (ed.)). Pemeral Edukreatif.

Suyata, P., Hidayanto, N., & Widyantoro, A. (2014). Standarisasi instrumen integrated assessment hasil belajar bahasa dengan program QUEST. LITERA, 13(2). https://doi.org/10.21831/ltr.v13i2.2588

Taylor, R. T., Bishop, P. R., Lenhart, S., Gross, L. J., & Sturner, K. (2020). Development of the BioCalculus Assessment (BCA). CBE—Life Sciences Education, 19(1), ar6. https://doi.org/10.1187/cbe.18-10-0216

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Care & Research, 57(8), 1358–1362. https://doi.org/10.1002/art.23108

Tim Pusat Penilaian Pendidikan. (2019). Panduan Penilaian Tertulis. Pusat Penilaian Pendidikan. https://repositori.kemdikbud.go.id/18344/

van de Grift, W. J. C. M., Houtveen, T. A. M., van den Hurk, H. T. G., & Terpstra, O. (2019). Measuring teaching skills in elementary education using the Rasch model. School Effectiveness and School Improvement, 30(4), 455–486. https://doi.org/10.1080/09243453.2019.1577743

Wolfs, Z. G., Brand-Gruwel, S., & Boshuizen, H. P. A. (Els). (2023). Assessing tonal abilities in elementary school children: testing reliability and validity of the implicit tonal ability test using rasch measurement model. SAGE Open, 13(3). https://doi.org/10.1177/21582440231199041


ISSN (print) 2685-7111 | ISSN (online) 2338-6061