Jurnal Pendidikan Matematika
Volume 17, No. 3, September 2023, pp. 397-416
P-ISSN: 1978-0044, E-ISSN: 2549-1040, DOI: https://doi.org/10.22342/jpm.17.3.20627.397-416
Website: https://ejournal.unsri.ac.id/index.php/jpm
Accredited by SINTA 2: http://sinta2.ristekdikti.go.id/journals/detail?id=1811

Quality of Mathematics Even Semester Final Assessment Test in Class VIII Using R Program

Nadilla Rahmadani[1,*], Kana Hidayati[2]

[1]Master of Mathematics Education Program, Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Jl. Colombo No. 1, Yogyakarta, Indonesia
[2]Department of Mathematics Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Jl. Colombo No. 1, Yogyakarta, Indonesia
[*]Email: nadillarahmadani.2021@student.uny.ac.id

Abstract

This study was conducted using the Item Responses Theory (IRT) method with R program to comprehensively analyze quality of Mathematics Even Semester Final Assessment test in class VIII for 2021/2022 Academic Year. This assessment was created through a collaborative effort involving mathematics teachers from a Public Junior High School in Binjai. It consists of 20 multiple-choice questions, each with four alternative answers. Furthermore, the study followed a descriptive framework carried out by a quantitative methodology. The data source was the responses of 189 students in class VIII who took part in Mathematics Even Semester Final Assessment for the 2021/2022 Academic Year, collected using documentation methods. The results showed that the items developed by the teacher: (1) were most appropriate for analysis using a two-parameter logistic model (2-PL), (2) distribution of material achieved during the even semester on the item tested was uneven, (3) eight of the 20 item was acceptable and kept in the question bank, while the remaining 12 were of poor quality; (4) the item fell into the category of easy to moderate difficulty, dominated by item in the moderate category, and (5) Mathematics Even Semester Final Assessment in class VIII provided accurate information regarding students' mathematics ability at moderate ability levels (-2 to $+1.5$).

Keywords: Final Semester Assessment, Item Analysis, Item Responses Theory, Mathematics, R Program

Abstrak

Penelitian ini menggunakan pendekatan Teori Respon Butir dengan program R untuk mendeskripsikan kualitas tes Penilaian Akhir Semester Genap Matematika kelas VIII tahun pelajaran 2021/2022. Tes yang dikembangkan oleh kelompok guru matematika kelas VIII di salah satu SMP Negeri di Binjai ini terdiri dari 20 soal pilihan ganda dengan empat alternatif jawaban. Penelitian ini bersifat deskriptif dengan pendekatan kuantitatif. Sumber data penelitian ini adalah lembar jawaban dari 189 siswa kelas VIII yang mengikuti Penilaian Akhir Semester Genap Matematika tahun ajaran 2021/2022 yang dikumpulkan dengan menggunakan teknik dokumentasi. Temuan penelitian ini menunjukkan bahwa butir-butir soal Penilaian Akhir Semester Genap Matematika di kelas VIII yang dikembangkan oleh guru: (1) paling tepat dianalisis menggunakan model logistik dua parameter (2-PL); (2) persebaran materi yang dicapai selama semester genap pada butir-butir soal yang diujikan masih belum merata; (3) delapan dari 20 butir soal dapat diterima dan disimpan di bank soal untuk digunakan dalam Penilaian Akhir Semester Genap Matematika tahun berikutnya, sedangkan 12 item lainnya berkualitas tidak baik; (4) butir-butir soal termasuk dalam kategori kesukaran mudah sampai sedang, didominasi oleh soal-soal dengan kategori sedang; dan (5) Penilaian Akhir Semester Genap Matematika di kelas VIII akan memberikan informasi yang akurat mengenai kemampuan matematika siswa pada tingkat kemampuan sedang (-2 sampai $+1{,}5$).

Kata kunci: Penilaian Akhir Semester, Analisis Butir Soal, Teori Respon Butir, Matematika, Program R

How to Cite: Rahmadani, N., & Hidayati, K. (2023). The quality of mathematics even semester final assessment test in class VIII using R program. *Jurnal Pendidikan Matematika, 17*(3), 397-416.

INTRODUCTION

The government is actively engaged in enhancing quality of its human resources and preparing for the future, primarily through education. In this context, the most prevalent channel for acquiring an

Received February 23, 2023; Revised October 14, 2023; Accepted October 31, 2023

education is the formal route of schooling. These educational institutions serve as hubs where students can access, accumulate, and cultivate knowledge and skills during their learning process. According to Purwanto (2019), while learning can theoretically transpire in various settings, systematic practice can only be conducted within the structured environment of a school. This distinction arises from the explicit specification of learning characteristics and structures in a formal setting, including the establishment of clear learning objectives (Johnson & Majewska, 2022). Furthermore, the process of learning in a school is nurtured through direct instructional methods, discernible academic progress, and rigorous assessment of outcomes.

Johnson & Majewska (2022) stated that informal learning occurred outside the classroom and might be unconscious. The concept covers the learning conducted by individuals daily throughout their lives. Therefore, the only difference between learning in schools and other environments is that educational goals are planned to change behavior and the achievement can be measured. Purwanto (2019) reported that learning outcomes were changes in behavior achieved after children participate in the teaching and learning process. Similarly, Pratama & Pinayani (2019) stated that learning outcomes were behavioral changes in a person as shown by adjustments in knowledge and abilities coupled with processes learned through specific experiences. This means that the assessment of learning outcomes serves as a direct reflection of the teaching objectives (Gronlund, 1985).

In this context, it is worth noting that from elementary school through to university, mathematics is a subject consistently integrated into the curriculum. A method to maintain the efficacy of mathematics instruction and classroom learning is through the evaluation of student learning outcomes (Mardapi, 2012; Retnawati, 2013). After engaging in various educational activities, assessment is the most important and efficient way to determine students' comprehension (Winarno et al., 2019). Learning outcomes must also be subjected to measurement and evaluation to ascertain the attainment of the set goals, confirming the successful fruition in obtaining the desired results. Therefore, it is crucial to make an appropriate assessment, including preparing the tools to be used.

Test is instruments used to evaluate learning outcomes. These instruments of measurement (Kaplan & Saccuzzo, 2017) are used to measure the behavior of the sample in an "objective" and "standardized" way (Anastasi, 1988). Test serves as an instrument commonly used by teachers to assess student achievement and capabilities within specific academic domains (Zainuddin, 2018). A prominent example frequently used to evaluate student achievement within a given time frame is the final semester assessment using multiple-choice tests, which are considered objective measurement tools (Dehnad et al., 2014). Multiple-choice questions are used due to their simplicity in analysis and capacity to facilitate prompt feedback (Ulitzsch et al., 2020; Yamamoto et al., 2018). According to Mardapi (2012), the test's purpose, the sample size, the time allotted for checking the answers, the breadth of the material, and the characteristics of tested subjects are typically considered when selecting the format.

Considering that giving tests as part of the learning process is crucial for achieving learning goals (Anggoro et al., 2019), a quality assessment tool is needed because this device should be able to show the extent of students’ cognitive abilities (Hutabarat, 2009). This is consistent with Miller, Linn, & Gronlund (2012), where learning assessment is a crucial measuring tool for educational success, enhancing change within the educational system by ensuring that the analysis of assessment items is in line with the intended evaluation criteria. This confirms the appropriateness of the instruments used for measuring the desired competencies and the success of assessing learning outcomes is closely related to the test used (Tilaar & Hasriyanti, 2019). Since dependable tests can precisely show students' learning outcomes, a high-quality analysis must fulfill the fundamental criteria of assessment instrument. This instrument should ensure that the obtained information exceeds the margin of measurement error, thereby attaining a superior level of accuracy (Retnawati, 2013). High information value will provide an overview of the actual measurement results. A high-quality instrument must pay attention to the characteristics of the item described (Santoso et al., 2019). Furthermore, Ali & Istiyono (2022) stated that a good test must have good questions and this understanding is crucial for bettering instruction. Therefore, a good instrument can be described through the item’s characteristics, and analysis is required to ensure the questions are high quality.

Analysis of item characteristics is instrumental in assessing quality and this includes examining the effectiveness in evaluating the abilities of test-takers and their appropriateness for the intended audience (Christian et al., 2017). Through the insights offered, teachers can discern the suitability of test items for the proficiency levels of test takers, as well as the capacity to obtain precise information related to the abilities of the students (Lin, 2018). Item analysis serves several purposes, including evaluating test quality during the process of developing and manufacturing tests (Kaplan & Saccuzzo, 2017), assisting with the investigation of item characteristics and improving the quality of subject exams (Talebi et al., 2013), assisting with the identification of test-related flaws (Anastasi & Urbina, 1997), and confirming students’ understanding of the material used during the learning process (Aiken, 1994). The achievement test serves as an important measuring instrument within the evaluation process, showing the critical importance of conducting a comprehensive analysis of the characteristics (Himelfarb, 2019). Therefore, test’s characteristics must be considered when creating a compatible test. Item analysis can also predict several criteria based on test scores, showing the reliability and determining the increase in test characteristics (Murphy & Davidshofer, 2005). The characteristics of the item are related to the behavior of test taker's answers. Diverse responses indirectly report the attributes and tendencies of the item and these characteristics are connected to test takers. An effective assessment instrument should not be affected by excessively challenging items. This item fails to discriminate between capable and less capable test takers. However, the presence of item that are either too easy or too difficult does not necessarily render the instrument ineffective or validate their exclusion. This means that a comprehensive empirical item analysis must be conducted to ascertain the specific attributes and behaviors of each item (Retnawati, 2013).

Item Response Theory (IRT) and Classical Test Theory (CTT) are two methods used for empirical item analysis. This can be seen from the many previous studies using both methods in analyzing the item (Ali & Istiyono, 2022; Awopeju & Afolabi, 2016; Jabrayilov et al., 2016; Moreta-Herrera et al., 2023; Primi et al., 2015; Sarea & Ruslan, 2019; Sudaryono, 2011). CTT is very dependent on the characteristics of the participants being measured and measurement errors estimated collectively or in groups rather than individually (Ali & Istiyono, 2022; Kaplan & Saccuzzo, 2017; Shanti et al., 2020). Furthermore, test items must be responded to by the same test takers to prevent changing the characteristics (Sudaryono, 2011). To overcome this issue, IRT was developed as an update of CTT, focusing on how each test-taker interacts with the item, making the concept question-oriented rather than test-oriented (Retnawati, 2014).

IRT represents a question-oriented method that maintains the consistency of test-takers' relationships with the item, ensuring their responses remain stable. Additionally, the inherent characteristics remain constant across various participants, despite differences in responses. According to Sudaryono (2011), modern measurements differ fundamentally from classical ones because scoring is invariant (unchanged or fixed) to test items and participants. IRT uses relativity and probability principles to develop a logistic model that connects a person's chances of answering correctly with a scale of ability (θ), level of difficulty (b), distinguishing power (a), and pseudo-guessing (c) (Hambleton et al., 1991; Keeves & Alagumalai, 1999). The relationship between these parameters is expressed in the three logistic parameter (PL) models used to estimate the characteristics, namely the 1-PL or Rasch, 2-PL, and 3-PL models (DeMars, 2010; Hambleton & Swaminathan, 2013). The selection of the appropriate model is important, showing the genuine characteristics of test as an outcome of measurement.

The 1-PL or Rasch model estimates the relationship between difficulty level (b) and ability, where the difference is that Rasch has a discriminant value of one (DeMars, 2010; Finch & French, 2015). The 2-PL model estimates the item's difficulty level (b) and distinguishing power (a), while the 3-PL model describes the characteristics using the level of difficulty (b), distinguishing power (a), and guessing (c) of the item. The distinguishing power parameter (a) shows the ability to distinguish between students with high and low abilities. Meanwhile, the possibility of providing correct responses to an issue at a particular ability level is shown by the difficulty level parameter (b). The difficulty of the questions is said to meet the various levels proportionally when the questions have good quality (Kaya & Tan, 2014). In addition, the pseudo-guessing parameter shows the possibility for test takers with a low ability to answer questions correctly (Hambleton & Swaminathan, 2013).

Item analysis of the IRT method can be carried out using several computer programs as conducted in previous studies, such as MULTILOG and TESTFACT (Baker, 2001), QUEST (Rizbudiani et al., 2021), PARSCALE (Herosian et al., 2022), Winsteps (Azizah & Wahyuningsih, 2020; Palimbong et al., 2018), IRTPROV3.0 (Essen et al., 2017), BILOG-MG V3.0 (Amelia & Kriswantoro, 2017), and TAP (Mahanani, 2015). Furthermore, the R program is open-source software for data processing and statistical analysis based on programming language. Data can be analyzed using the R program, from basic to complex (Sarvina, 2017), and the software has various functions used in mathematics, including being applied to IRT models. However, study using the R program is limited to several results, such as Shanti et al. (2020), using R for Rasch model analysis, Muchlisin, Mardapi, & Setiawati (2019) analyzing the level of difficulty, ICC and IIC curves, Tirta (2015) who analyzed only using 1-PL, 2-PL and 3-PL models, Thepsathit et al. (2022) who analyzed using 1-PL and 2-PL models on the data polytomous, and Ali & Istiyono (2022) who tested the three logistic parameter models, and looked at the ICC and IIC curves on the national mathematics test item.

Even though numerous results have applied IRT to analyze test items, there remains a scarcity of studies regarding item characteristics. This in-depth analysis includes the determination of item difficulty levels, distinguishing power, and guessing, as well as the derivation of item characteristic curves (ICC), item information curves (IIC), and the calculation of the number of items retained in the question bank for subsequent assessment. The choice to select the IRT method was informed by insights from an interview with a mathematics teacher in a public junior high school in Binjai for the 2021/2022 Academic Year. This interview showed an absence of item analysis during the Mid and final-semester assessments, despite the integral components of summative evaluation. The primary purpose is to measure students' learning achievements by assessing their progress toward competencies within a specific time frame (Ismail et al., 2022). Therefore, the quality of these assessments remains uncertain, necessitating the teacher's need to ascertain whether test has effectively fulfilled its intended function.

Based on the description, considering that the test item was tested on all eighth graders, the data sources were obtained from respondents, teachers, learning environments, and differences in methods and strategies used. The IRT method is more suitable for analyzing items based on various student abilities, considering the limitations of the CTT method. Therefore, the primary aim of this study was to assess the quality of the Mathematics Final Semester Assessment instrument, which was collaboratively developed by a group of Grade VIII teachers at a Public Middle School in Binjai during the 2021/2022 Academic Year. This assessment was achieved through the identification of the most suitable model for analyzing item characteristics and estimating the abilities of test takers using IRT with R program.

METHODS

This descriptive study conducted with a quantitative method described the quality of Mathematics Even Semester Final Assessment instrument in class VIII for the 2021/2022 Academic Year. Furthermore, content analysis was used to identify the various specific characteristics of a message objectively, systematically, and generally contained in the questions and the pattern of answer sheets (responses) of test participants (Santoso et al., 2019). Data was collected using documentation methods in class VIII at one of the Public Middle Schools in Binjai in the form of responses of test takers who took Mathematics Even Semester Final Assessment in the 2021/2022 academic year. The number of respondent data obtained was 189 students and the instrument consists of 20 multiple-choice questions with four alternative answer choices.

Student responses data were analyzed using the IRT method with the LTM package in the R program. The LTM package was used to analyze IRT with a logistic parameter model (Tirta, 2015). At the analysis stage (data that only contains 0 and 1), the R program was used to convert the student responses (A, B, C, and D) to create dichotomous data (Paek & Cole, 2020). Several stages were carried out to determine the quality of the instrument, namely the model selection stage, the item parameter estimation stage, the analysis stage through ICC and IIC, and the item quality categorization stage. At the model selection stage with the IRT method, the analysis model was selected. First, item analysis was carried out based on the 1-PL, 2-PL, and 3-PL models. Second, the chi-square value and its significance in the analysis of each model were collected. Third, the chi-square significance value in each model was compared with $\alpha = 0.05$. The item was said to be fit when the model significance value was < 0.05. Fourth, the accumulation of the number of items that fit each model was stated. The model with the highest number of fit items was selected to estimate test takers' ability.

The results of item parameter estimation and ability based on the selected model criteria were described in the following stage. Table 1 provides the criteria for the 1-PL, 2-PL, and 3-PL models adopted by Hambleton et al. (1991).

Table 1. Criteria for IRT model with good category

Model	Parameter		
	a	b	c
1-PL	-	-2 to $+2$	-
2-PL	0 to $+2$	-2 to $+2$	-
3-PL	0 to $+2$	-2 to $+2$	0 to $\frac{1}{k}$

Description:

a: distinguishing power; b: difficulty level; c: pseudo-guessing; and k: number of answer choices

The values obtained for the parameter b are categorized to classify the level of difficulty of the item based on Table 2.

Table 2. Category of item difficulty level in IRT (Retnawati, 2014)

Value Range	Difficulty Level
$b > 2$	Very difficult
$1 < b \leq 2$	Difficult
$-1 \leq b \leq 1$	Moderate
$-2 \leq b < -1$	Easy
$b > -2$	Very easy

The item analysis results on the logistic parameter model are used to determine ICC and IIC as a follow-up analysis. In addition, analysis results are also used to categorize each item's quality through the criteria adopted by Ali & Istiyono (2022), as shown in Table 3.

Table 3. Item quality category in the logistics parameter model

Category	Logistics Parameter Model		
	1-PL	2-PL	3-PL
Good (G)	$-2 \leq b \leq 2$ Fit on the model (FM)	$-2 \leq b \leq 2$ $0 \leq a \leq 2$ Fit on the model (FM)	$-2 \leq b \leq 2$ $0 \leq a \leq 2$ $c \leq 0.25$ Fit on the model (FM)
Not Good (NG)	$b > 2$ atau $b < 2$ Not fit on the model (NFM)	$b > 2$ atau $b < 2$ $a > 2$ atau $a < 0$ Not fit on the model (NFM)	$b > 2$ atau $b < 2$ $a > 2$ atau $a < 0$ $c > 0.25$ Not fit on the model (NFM)

RESULTS AND DISCUSSION

The item for Mathematics Even Semester Final Assessment of grade VIII at one of the Public Middle Schools in Binjai was made based on achievement indicators taught to students for one semester. The material under test comprises three primary topics, namely the Pythagorean theorem, circles, and flat-sided solid figures. There are 20 multiple-choice items in the instrument, with the proportion of 60% Pythagorean theorem, 25% circle, and 15% flat-sided solid figure material. Table 4 shows the results of the distribution of the topics on the item and the uneven distributions of the item.

Table 4. Description of the topic in the item

Topic	Achievement Indicator	Item number	Total
Pythagorean Theorem	Apply the Pythagorean theorem to solve problems	1, 2, 3, 4, 5, 6, 10, 12	8
	Solve a problem that includes Pythagorean triples	7, 11, 14	3
	Find the lengths of the missing sides of a triangle with a $30° - 60° - 90°$ triangle sides ratio	8	1
Circle	Determine the formula of the area and circumference of a circle	15	1
	Apply the area of a circle formula to solve a problem	9	1
	Apply the circle circumference formula to solve a problem	13, 16, 17	3
3-D Geometric Shapes (Flat-Sided Solid Figure)	Predict a 3-D object that can be created from a net (Prism)	18	1
	Mention the body or space diagonal of the cuboid	20	1
	Determine the surface area of a cube	19	1

The distribution is not eve, where the flat-sided solid figure material has a minor proportion, even though it has the most achievement indicators in the syllabus. Therefore, the instruments are only partially capable of measuring the item based on the lesson taught by the indicators of learning achievement in class VIII material.

Selection of the IRT Analysis Model

The IRT method offers three analytical models, namely the 1-PL, 2-PL, and 3-PL. A model that fits the data from the three analyses is selected by comparing the significance value of the chi-square with $\alpha = 0.05$. According to Amelia & Kriswantoro (2017), tthe chi-square value of an item can be used to determine the fitness of the model, with an item being deemed not to fit when the probability value (significance) is less than 0.05 (Retnawati, 2014). Table 5 shows the analysis results of the model fit test based on the chi-square value of each logistic parameter model in item responses theory and based on Table 5, the 2-PL model can be used with the IRT method to analyze the data.

Table 5. Item fit with the model

Analysis Model	Number of items	
	Fit the model	Not fit the model
1-PL	9	11
2- PL	16	4
3- PL	8	12

Item Parameter Estimation

The item parameters were estimated using the 2-PL model based on the model fit test analysis. The parameters of difficulty level (b) and distinguishing power (a) were used in the analysis of the 2-PL model to determine the characteristics of the item. Table 6 shows the analysis results of the 20 items in the 2-PL model using the R program for each item.

Table 6. Item analysis results with the 2-PL model using the R program

Item	Distinguishing power		Difficulty level			Chi-square	
	a	Category	b	Category	Classification	Probability	Category
1	1.122	Good	−0.929	Good	Moderate	0.032	NFM
2	1.416	Good	−0.921	Good	Moderate	0.482	FM
3	1.756	Good	−0.525	Good	Moderate	0.377	FM
4	0.010	Good	57.814	Not Good	Very difficult	0.407	FM
5	0.571	Good	−0.834	Good	Moderate	0.053	FM
6	2.688	Not Good	−0.462	Good	Moderate	0.378	FM
7	2.601	Not Good	−0.598	Good	Moderate	0.525	FM
8	1.623	Good	0.066	Good	Moderate	0.077	FM
9	−0.137	Not Good	−9.956	Not Good	Very easy	0.626	FM
10	2.589	Not Good	−0.461	Good	Moderate	0.384	FM
11	3.836	Not Good	−0.251	Good	Moderate	0.346	FM
12	1.617	Good	−0.150	Good	Moderate	0.558	FM
13	2.562	Not Good	−0.007	Good	Moderate	0.357	FM
14	1.618	Good	−0.130	Good	Moderate	0.050	FM
15	0.380	Good	−3.081	Not Good	Very easy	0.715	FM
16	0.637	Good	−0.473	Good	Moderate	0.039	NFM
17	0.386	Good	0.061	Good	Moderate	0.012	NFM
18	0.712	Good	−1.359	Good	Easy	0.717	FM
19	1.019	Good	−0.685	Good	Moderate	0.001	NFM
20	0.759	Good	−0.491	Good	Moderate	0.063	FM

From the estimation of distinguishing power (a), 14 item had good distinguishing power in the range of 0.010 to 1.756. Based on this value, item number three has the highest estimated value of $1{,}756$. This item facilitates the identification of students who have not attained a comprehensive mastery of the subject. The other six items, namely six, seven, nine, ten, eleven, and thirteen, are outside the good category because the value is not in the range of $0 - 2$. Therefore, the six items have a low level of discrimination and poor distinguishing power (Hamidah & Istiyono, 2022). A total of 17 items fell into the good category for the difficulty level parameter, with estimates ranging from -1.359 to 0.066. Item numbers 18 and 8 have the lowest and highest estimated values ($b = -1.359$ and 0.066). This means that item numbers 18 and 8 have easy and most difficult levels. Due to the estimated value not falling within the range of $-2 \leq b \leq 2$, the difficulty level for the remaining three, specifically 4, 9, and 15, is not good. Analysis results of the difficulty level are categorized by each group, presented in Table 7.

Table 7. Item difficulty level

Difficulty level	Item number	Total	Percentage
Easy	18	1	5%
Moderate	1, 2, 3, 5, 6, 7, 8, 10, 11, 12, 13, 14, 16, 17, 19, 20	16	80%

Table 7 shows that most of Mathematics Even Semester Final Assessment questions belong to the moderate category, with one and three questions going into the easy and bad categories. Theoretically, these results show a proportion not good enough for the test instrument's difficulty level. This is because a good test will naturally include questions with varying levels, such as easy, moderate, and difficult items. Based on the analysis results in Table 6, a percentage is obtained for each item parameter, as shown in Table 8 below.

Table 8. Characteristics of items based on the 2-PL model

Category	Parameter	
	a	b
Good	70% (14 items)	85% (17 items)
Not good	30% (6 items)	15% (3 items)

Item Characteristic Curve (ICC)

ICC is generated through R program for further analysis. This curve explains the examinees' characteristics, the relationship between their ability level, and the probability of correct responses. Furthermore, it is possible to discern which test item is the least challenging and the most demanding. The 2-PL model's ICC is shown in Figure 1.

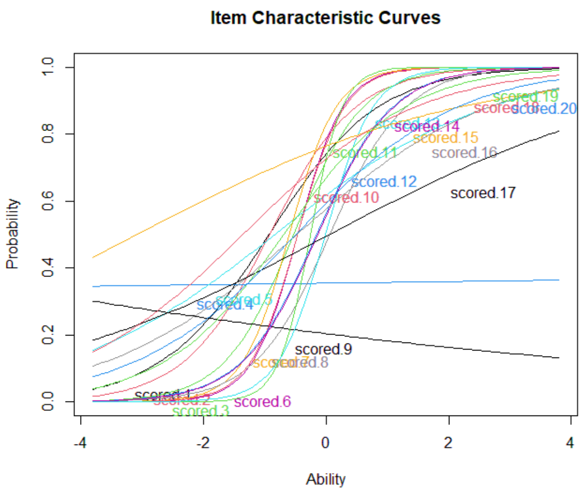
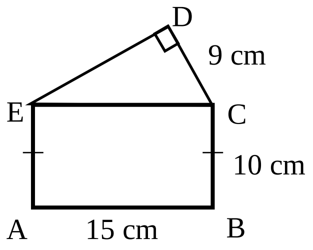


Figure 1. ICC for the 2-PL model using R program

The ICC show of the 2-PL model in Figure 1 shows that items 4, 9, and 15 have distinct curve shapes from the other curves. This shows a weak correlation between the examinees' ability level and the probability of giving accurate responses. In item number 4 (blue curve), the curve forms a horizontal line, showing that all students have an equal chance of correctly answering the question. This is supported by the estimation of parameters b and a of 57,814 (bad category) and 0,010 (good category). (see Table 6). The estimated parameter b shows that the item have a poor difficulty level far from the range of -2 to 2. Even though the estimated parameter a (distinguishing power) is categorized as good between $0-2$, the estimation value is close to 0 and this item can barely differentiate high and low-ability students. This is consistent with the categorization of item discrimination by Hamidah & Istiyono (2022), where a value < 0.20 has poor discrimination or distinguishing power. Item number 4 was designed to assess students' comprehension of determining the circumference of a composite shape formed by combining two plane figures. This includes the application of the Pythagorean theorem to ascertain the necessary side lengths, as shown in Figure 2.

Perhatikan gambar di bawah ini!



Keliling bangun $ABCDE$ adalah …

a. $56\ cm$	b. $59\ cm$
c. $74\ cm$	d. $86\ cm$

Translated to English :

Look at the picture below!

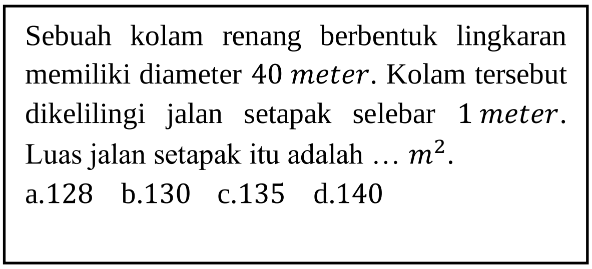
The perimeter of $ABCDE$ is...

a. $56\ cm$	b. $59\ cm$
c. $74\ cm$	d. $86\ cm$

Figure 2. Item number four

In item 9 (black curve), a downward curve is formed, implying that low-ability students are more likely to respond correctly. As a result, this item has low distinguishing power because there is a negative correlation between the examinees' ability level and the probability of providing accurate responses. This is also supported by the estimation of parameters b and a at -9.956 (bad category) and

-0.137 (bad category) (see Table 6). These parameters are unable to distinguish between students with high and low abilities in mastering the material. Negative distinguishing power shows that students with high and low abilities answered the item incorrectly and correctly (Hamidah & Istiyono, 2022). Item number 9 was made to know students' understanding of determining the area of a shape from the circle section with another plane shape, as presented in Figure 3.



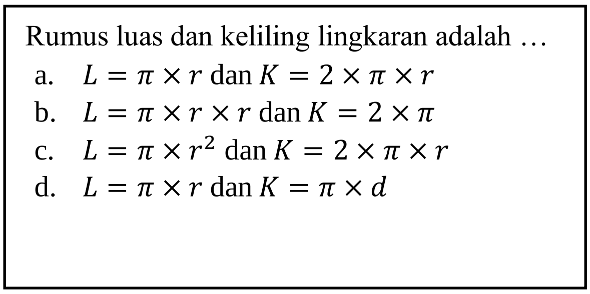
Sebuah kolam renang berbentuk lingkaran memiliki diameter $40\ meter$. Kolam tersebut dikelilingi jalan setapak selebar $1\ meter$. Luas jalan setapak itu adalah $\dots\ m^2$.
a.128 b.130 c.135 d.140

Translated to English :

A circular swimming pool has a diameter of $40\ meters$. The pool is surrounded by a $1\ meter$ wide path. The area of the path is ... m^2.
a.128 b.130 c.135 d.140

Figure 3. Item number nine

The subsequent item is number 15, as shown by the yellow curve. According to the curve, around 50% of answers from students with low abilities are accurate, and students' abilities increase with their chances. Therefore, individuals with low abilities are likely to provide the correct responses, suggesting the simplicity of the item. This is evident from parameter b estimation results (see Table 6), which show a value of $-3{,}081$ (bad category) with very low item's difficulty level. The observation can be shown from the content of question 15, where students are tasked with deriving the formulas for the area and circumference of a circle, as depicted in Figure 4. A conclusion can be drawn from the ICC that the easier the question, the higher the proficiency level of the student with the likelihood of the answer being accurate.



Rumus luas dan keliling lingkaran adalah …
a. $L = \pi \times r$ dan $K = 2 \times \pi \times r$
b. $L = \pi \times r \times r$ dan $K = 2 \times \pi$
c. $L = \pi \times r^2$ dan $K = 2 \times \pi \times r$
d. $L = \pi \times r$ dan $K = \pi \times d$

Translated to English :

The formula for the area and circumference of a circle is...
a. $L = \pi \times r$ and $K = 2 \times \pi \times r$
b. $L = \pi \times r \times r$ and $K = 2 \times \pi$
c. $L = \pi \times r^2$ and $K = 2 \times \pi \times r$
d. $L = \pi \times r$ and $K = \pi \times d$

Figure 4. Item number 15

Item Information Curve (IIC) and Test Information Function

Further analysis is also shown through IIC to discover more information about the item (Ali & Istiyono, 2022). The function will also be substantial when the individual item within the assessment shows a strong function. These results can manifest as test information functions and individual item characteristics. The IIC graph depicted in the image serves as a useful tool for visually representing the item information function. Figure 5 shows that item number 11, with a curve height of 3.7 and an θ ability close to 0, is the highest in offering the most information relative to others.

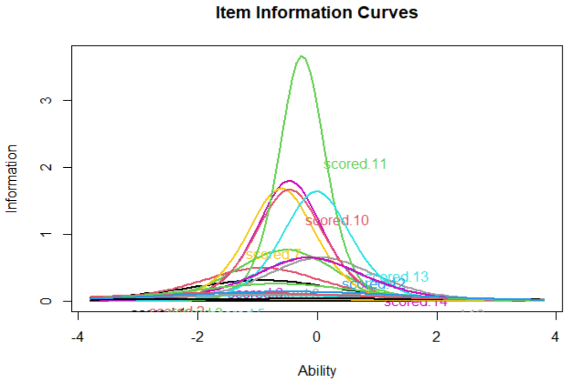


Figure 5. IIC for the 2-PL model

Maximum information on the test is also apparent through the graph of the test information function which discovers the information as shown in Figure 6.

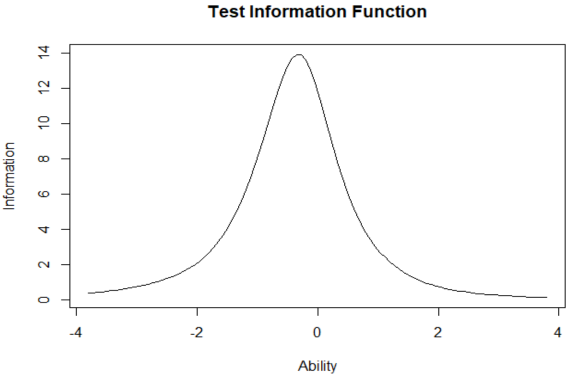
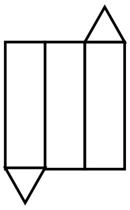


Figure 6. TIF for the 2-PL model

Figure 6 shows that Mathematics Even Semester Final Assessment instrument in class VIII has a higher information value in the ability range of around -2 to $+1.5$. Testing a question on individuals whose abilities lie below the given range leads to a measurement error that exceeds the value of the information function. Mathematics Even Semester Final Assessment test is perfect for measuring test takers with moderate abilities.

Following the model selection phase, an item parameter estimation was conducted, as shown in Table 6. Subsequently, a comprehensive analysis was performed using ICC and IIC. Referring to the criteria in Table 3, items 2, 3, 5, 8, 12, 14, 18, and 20 show good quality. These eight items are characterized by favorable parameter estimates for both b and a and showing a strong consistency with the 2-PL model. However, none of the questions related to circle material has good quality. The following shows a good-quality question suitable for the question bank for testing at the final assessment of the semester (Figures 7, 8, 9).

Gambar dibawah ini merupakan jaring-jaring …



a. Kubus
b. Prisma
a. Limas
b. Kerucut

Translated to English :

The image below is a net of ...

a. Cube
b. Prism
c. Pyramid
d. Cone

Figure 7. Item number 18 (easy difficulty level)

Sebuah tangga panjangnya $2{,}5\ m$ disandarkan pada tembok. Jika jarak ujung bawah tangga ke tembok $0{,}7\ m$, tinggi tangga diukur dari tanah adalah…

a. $1{,}5\ m$	c. $2{,}4\ m$
b. $2\ m$	d. $3{,}75\ m$

Translated to English :

A ladder 2.5 m long is leaning against a wall. If the distance from the bottom of the ladder to the wall is 0.7 m, the ladder's height measured from the ground is...

a. $1.5\ m$	c. $2.4\ m$
b. $2\ m$	d. $3.75\ m$

Figure 8. Item number 12 (moderate difficulty level)

Pada sebuah segitiga PQR diketahui sisi-sisinya $p, q,$ dan r. Dari pernyataan berikut yang benar adalah …

a. Jika $q^2 = p^2 + r^2, \angle P = 90°$
b. Jika $r^2 = q^2 - p^2, \angle R = 90°$
c. Jika $r^2 = p^2 - q^2, \angle Q = 90°$
d. Jika $p^2 = q^2 + r^2, \angle P = 90°$

Translated to English :

In a triangle PQR, the sides p, q, and r are known. Which of the following statements is true...

a. If $q^2 = p^2 + r^2, \angle P = 90°$
b. If $r^2 = q^2 - p^2, \angle R = 90°$
c. If $r^2 = p^2 - q^2, \angle Q = 90°$
d. If $p^2 = q^2 + r^2, \angle P = 90°$

Figure 9. Item number five (moderate difficulty level)

The results provide information that Mathematics Even Semester Final Assessment instrument in class VIII for the 2021/2022 school year developed by a group of class VIII math teachers still needs to be improved. This can be seen from the material distribution to the item still uneven (Table 4). Furthermore, the best model for analyzing tests is determined by model fit analysis. The fit of this item is crucial, considering that the application of IRT can be justified when the data is by the model (KÖSE, 2014). The items on Mathematics Even Semester Final Assessment are very suitable for analysis using 2-PL, according to Table 5, which presents the outcomes of the fit test model of IRT analysis for 20 items with three logistic parameters using R program. This is because the 2PL model offers more items that are better fit. According to Amelia & Kriswantoro (2017), the logistic model used for parameter estimation accepts many fit items. This is supported by Ali & Istiyono (2022), where the number of items in the “fit with the model” (FM) category affects the suitability of analysis. Subali et al. (2019) stated that CTT and IRT (2PL) models could examine the item’s parameters, including difficulty, distinguishing power, and student responses. Therefore, teachers should be encouraged to use IRT when creating test items, and simulating students' responses (Esomonu & Okeke, 2021). An item's fit analysis can determine the best model for creating questions, leading to the highest quality.

According to Almaleki (2021), the IRT model was used to analyze student responses. The results show that the multiple-choice test item is impacted by participant diversity regarding the accuracy of parameter estimates and individual ability. The capability of an item to differentiate between test takers who have mastered the subject being taught and those who have not can be determined by using distinguishing power parameters (Sudaryono, 2011). The distinguishing power parameter (a) describes how well the item can separate students with high and low abilities. According to IRT, a good item's distinguishing power ranges from 0 to $+2$ on a logit scale, meaning that the higher the value, the better the item (Hambleton et al., 1991; Linden & Hambleton, 1997). Based on the estimation using the 2PL IRT model (see Table 8), six items had a poor distinguishing power category while those with less than zero had low distinguishing power. The item cannot distinguish between capable and incapable test takers. This means students who perform well on tests, with high-ability students can provide accurate responses to questions with good distinguishing power. In addition, items with low power will provide inaccurate information (Wu et al., 2016). An item with a low power shows the presence of ambiguous word usage when the student's abilities can be distinguished. Ashraf & Jaseem (2020) state that further inspection is necessary to determine the cause of negative item index values. In this context, several potential issues may arise, including the possibility of an incorrect solution key, the existence of multiple answer keys, ambiguity in defining the competency being measured, the ineffectiveness of the distractor options, or excessively challenging nature of the material (Sudaryono, 2011). According to Wu et al. (2016), there are three potential factors contributing to the item's diminished ability to differentiate between test-takers compared to other items. Firstly, this item may be assessing a different construct or competency. Secondly, the item may be improperly formulated, leading to confusion among test-takers. Lastly, the level of difficulty can fall significantly above (high difficulty) or below (low difficulty) the desired range. Wu et al. (2016) stated that the removal or replacement of items with low distinguishing power should be conducted to enhance test reliability and minimize measurement error, rendering the results more meaningful and interpretable.

The item difficulty level parameter is the opportunity to respond correctly to a question at a particular ability. The score resulting from responses of numerous test-takers determines the difficulty level (b). Furthermore, the difficulty of test decreases with the number of test-takers who can correctly answer the questions. In IRT, a good item's difficulty level on a logit scale ranges from -2 to $+2$ (Hambleton et al., 1991). A b value near -2 and $+2$ shows that the item is becoming easier and more challenging (Amelia & Kriswantoro, 2017). Based on the estimation using the 2-PL (Table 8), three items had a poor difficulty level category, including those in the very difficult and easy categories. The items in the good category are dominated by moderate difficulty levels (Table 7). Meanwhile, low distinguishing power is impacted by very easy or difficult items (Mardapi, 2012). Wu et al. (2016) stated that very easy or difficult items should be maintained because the instrument measures test takers with low and high abilities. The addition of items with a very low and high level of difficulty to the first few questions can lower test-taker anxiety and increase the distinguishing power.

According to the ICC (Figure 1), there is a negative correlation between test takers' ability levels and their chances of correctly answering items 9. The test's information function states that the teacher's instrument provides accurate information when used to evaluate mathematics ability in the range of about −2 to +1.5 (Figure 6). Istiyono et al., (2014) reported that the moderate ability level ranged from −2 to +1.5. Furthermore, the final semester assessment instrument can accurately obtain information on mathematics ability with moderate levels. This can be seen from the highest information value produced by this instrument, which is 13.7 when the ability of test takers or students is −0.5 (including the moderate ability level). According to Ramos et al., (2013), test takers with high abilities have a deep conceptual understanding and can apply their knowledge effectively when addressing problems. This means test takers with moderate abilities have sufficient conceptual understanding to overcome mathematic problems. A total of 8 items in class VIII had good quality based on the results using the 2-PL model in the IRT method. Therefore, the items with good quality are stored in the question bank to be used in the next semester's assessment; as stated by Retnawati & Hadi (2014), good-quality items should be kept in the question bank due to their good distinguishing power and difficulty levels.

CONCLUSION

In conclusion, Mathematics Even Semester Final Assessment instrument in class VIII for the 2021/2022 academic year was suitable for analysis using a two-parameter logistic model (2-PL). Furthermore, the distribution of material achieved during the even semester on the item tested was uneven. A total of the eight questions were acceptable and kept in the bank for use in the subsequent year. In contrast, the remaining 12 items were replaced or unsuitable for storage in the bank because of the failure to fit the model, poor distinguishing power, and the difficulty level was not good. The moderate category of the difficulty level dominated the questions, and one item had a negative correlation between the examinees' ability level and the probability of providing accurate responses. In addition, Mathematics Even Semester Final Assessment instrument in class VIII provided information accurately related to students' mathematics ability at moderate levels (−2 to +1.5).

The research findings had several applications in both practice and future analyses. To consider the item's characteristics used in terms of IRT, this study prepared Mathematics Even Semester Final Assessment instrument for the following year. In addition, test developers should pay attention to the rules in preparing the instruments to produce quality items. These individuals were field experts in developing test instruments and mastering the basic rules of compiling the item. Furthermore, recommendations for compiling test items is a trial examination was conducted with several respondents before using test questions to ensure the effectiveness and reliability of the questions. Concerning the limitations of this study, the sample used was relatively small. Therefore, future analyses were expected to analyze the quality of test instruments developed by other institutions. This study was limited to analyzing the quality of final semester assessment questions through multiple choices. Meanwhile, analysis related to testing quality was not limited to objective instruments and was carried out on descriptive tests and other forms. Further study recommends using additional programs to implement IRT, such as BILOG-MG, QUEST, TAP, and different applications to compare the results between the R program and others.

ACKNOWLEDGMENTS

The authors are grateful to the teachers at one of the Public Middle Schools in Binjai for facilitating the data needed in this study. The authors are also thankful to Mr. Dr. Ariyadi Wijaya, S.Pd.Si., M.Sc, and Mrs. Wahyu Setyaningrum, S.Pd., M.Ed., Ph.D for providing constructive suggestions. Furthermore, the authors would like to thank the anonymous expert reviewers and the editor for their assistance in reviewing this study.

REFERENCES

Aiken, L. R. (1994). *Psychological testing and assessment* (8th ed.). Boston: Allyn and Bacon.

Ali, & Istiyono, E. (2022). An analysis of item response theory using program R. *Al-Jabar: Jurnal Pendidikan Matematika*, *13*(1), 109–122. https://doi.org/http://dx.doi.org/10.24042/ajpm.v13i1.11252

Almaleki, D. (2021). Examinee Characteristics and their Impact on the Psychometric Properties of a Multiple Choice Test According to the Item Response Theory (IRT). *Engineering, Technology & Applied Science Research*, *11*(2), 6889–6901. https://doi.org/https://doi.org/10.48084/etasr.4056

Amelia, R. N., & Kriswantoro. (2017). Implementation of Item Response Theory as a Base for Analysis of Question Item Quality and Chemistry Ability of Yogyakarta City Students [in Bahasa*]. JKPK (Jurnal Kimia Dan Pendidikan Kimia)*, *2*(1), 1–12. https://doi.org/10.20961/jkpk.v2i1.8512

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Mc Millan.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.). New Jersey: Prentice-Hall Inc.

Anggoro, B. S., Agustina, S., Komala, R., Komarudin, Jermsittiparsert, K., & Widyastuti. (2019). An Analysis of Students' Learning Style, Mathematical Disposition, and Mathematical Anxiety toward Metacognitive Reconstruction in Mathematics Learning Process Abstract. *Al-Jabar: Jurnal Pendidikan Matematika*, *10*(2), 187–200. https://doi.org/10.24042/ajpm.v10i2.3541

Ashraf, Z. A., & Jaseem, K. (2020). Classical and Modern Methods in Item Analysis of Test Tools. *International Journal of Research and Review*, *7*(5), 397–403. https://doi.org/10.5281/zenodo.3938796

Awopeju, O. A., & Afolabi, E. R. I. (2016). Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics

Examination. *European Scientific Journal*, *12*(28). https://doi.org/10.19044/esj.2016.v12n28p263

Azizah, & Wahyuningsih, S. (2020). Use of the Rasch Model for Analysis of Test Instruments in Actuarial Mathematics Courses [in Bahasa]. *JUPITEK: Jurnal Pendidikan Matematika*, *3*(1), 45–50. https://doi.org/10.30598/jupitekvol3iss1ppx45-50

Baker, F. B. (2001). *The basics of item response theory* (Second). USA: ERIC.

Christian, D. S., Prajapati, A. C., Rana, B. M., & Dave, V. R. (2017). Evaluation of multiple choice questions using item analysis tool : a study from a medical institute of Ahmedabad , Gujarat. *International Journal of Community Medicine and Public Health*, *4*(6), 1876–1881. https://doi.org/10.18203/2394-6040.ijcmph20172004

Dehnad, A., Nasser, H., & Hosseine, A. F. (2014). A Comparison between Three-and Four-Option Multiple Choice Questions. *Procedia - Social and Behavioral Sciences*, *98*, 398–403. https://doi.org/10.1016/j.sbspro.2014.03.432

DeMars, C. E. (2010). *Item response theory: Understandings statistics measurement*. United Kingdom: Oxford University Press.

Esomonu, N., & Okeke, O. J. (2021). French Language Diagnostic Writing Skill Test For Junior Secondary School Students: Construction And Validation Using Item Response Theory. *International Journal of Education and Social Science Research ISSN*, *4*(02), 334–350. https://doi.org/10.37500/IJESSR.2021.42

Essen, C. B., Idaka, I. E., & Metibemu, M. A. (2017). Item level diagnostics and model-data fit in item response theory (IRT) using BILOG-MG v3. 0 and IRTPRO v3. 0 programmes. *Global Journal of Educational Research*, *16*(2), 87–94. https://doi.org/10.4314/gjedr.v16i2.2

Finch, W. H., & French, B. F. (2015). *Latent variable modeling with R*. United Kingdom: Routledge.

Gronlund, N. E. (1985). *Constructing Achievement Test*. New Jersey: Prentice-Hall Inc.

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. New York: Springer Science & Business.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. London: Sage Publication.

Hamidah, N., & Istiyono, E. (2022). The quality of test on National Examination of Natural science in the level of elementary school. *International Journal of Evaluation and Research in Education*, *11*(2), 604–616. https://doi.org/10.11591/ijere.v11i2.22233

Herosian, M. Y., Sihombing, Y. R., & Pulungan, D. A. (2022). Item Response Theory Analysis on Student Statistical Literacy Tests. *Pedagogik: Jurnal Pendidikan*, *9*(2), 203–215.

Himelfarb, I. (2019). A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating. *Journal of Chiropractic Education*, *33*(2), 151–163. https://doi.org/10.7899/JCE-18-22

Hutabarat, I. M. (2009). Analysis of Question Items with Classical Test Theory and Item Response Theory [in Bahasa]. *Pythagoras: Jurnal Pendidikan Matematika*, *5*(2), 1–13. https://doi.org/10.21831/pg.v5i2.536

Ismail, S. M., Rahul, D. R., Patra, I., & Rezvani, E. (2022). Formative vs summative assessment: impacts on academic motivation, attitude toward learning, test anxiety, and self-regulation skill. *Language Testing in Asia*, *12*(40), 1–23. https://doi.org/10.1186/s40468-022-00191-4

Istiyono, E., Mardapi, D., & Suparno, S. (2014). Development of a high-level physics thinking ability test (pysthots) for high school students [in Bahasa]. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *18*(1), 1–12. https://doi.org/10.21831/pep.v18i1.2120

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. *Applied Psychological Measurement*, *40*(8). https://doi.org/https://doi.org/10.1177/0146621616664046

Johnson, M., & Majewska, D. (2022). *Formal, non-formal, and informal learning: What are they, and how can we research them?* Cambridge University Press & Assessment Research Report.

Kaplan, R. M., & Saccuzzo, D. P. (2017). *Psychological testing: Principles, applications and issues*. Boston: Cengage Learning.

Kaya, Z., & Tan, S. (2014). New trends of measurement and assessment in distance education. *Turkish Online Journal of Distance Education*, *15*(1), 206–217. https://doi.org/10.17718/tojde.30398

Keeves, J. P., & Alagumalai, S. (1999). New approaches to measurement. *Advances in Measurement in Educational Research and Assessment*, 23–42. https://doi.org/10.1016/B978-008043348-6%2F50002-4

KÖSE, İ. A. (2014). Assessing model data fit of unidimensional item response theory models in simulated data. *Educational Research and Reviews*, *9*(17), 642–649. https://doi.org/10.5897/ERR2014.1729

Lin, S. (2018). Item Analysis of English Grammar Achievement Test. *Mandalay University of Foreign Languages Research Journal*, *9*(1), 13–20.

Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In *In Handbook of modern item response theory* (pp. 1–28). New York: Springer.

Mahanani. (2015). *Analysis of international competitions and assessments for schools (ICAS) questions using item response theory (IRT) and classical test theory (CTT)* methods [in Bahasa]. Skripsi Diterbitkan. Semarang: Jurusan Biologi FMIPA Universitas Negeri Semarang.

Mardapi, D. (2012). *Educational Measurement, Assessment and Evaluation* [in Bahasa]. Yogyakarta: Nuha Litera.

Miller, D., Linn, R., & Gronlund, N. (2012). *Measurement and Assessment Teaching*. New York: Pearson Education Limited.

Moreta-Herrera, R., Perdomo-Pérez, M., Reyes-Valenzuela, C., Gavilanes-Gómez, D., Rodas, J. A., & Rodríguez-Lorenzana, A. (2023). Analysis from the classical test theory and item response theory of the Satisfaction with Life Scale (SWLS) in an Ecuadorian and Colombian sample. *Journal of Human Behavior in the Social Environment*. https://doi.org/https://doi.org/10.1080/10911359.2023.2187915

Muchlisin, M., Mardapi, D., & Setiawati, F. A. (2019). An analysis of Javanese language test characteristic using the Rasch model in R program. *Research and Evaluation in Education*, *5*(1), 61–74. https://doi.org/10.21831/reid.v5i1.23773

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing principles and applications* (Fouth, Ed.). New Jersey: Prentice-Hall Inc.

Paek, I., & Cole, K. (2020). *Using R for Item Response Theory Model Applications*. United Kingdom: Routledge.

Palimbong, J., Mujasam, & Allo, A. Y. T. (2018). Item Analysis Using Rasch Model in Semester Final Exam Evaluation Study Subject in Physics Class X TKJ SMK Negeri 2 Manokwari. *Kasuari: Physics Education Journal*, *1*(1), 43–51. https://doi.org/10.37891/kpej.v1i1.40

Pratama, G. P., & Pinayani, A. (2019). Effect of Learning Style on Learning Outcomes with Mediator Variable Learning Motivation. *3rd ICEEBA International Conference on Economics, Education, Business and Accounting, KnE Social Sciences*, 2019, 808–819. https://doi.org/10.18502/kss.v3i11.4052

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2015). The Development and Testing of a New Version of the Cognitive Reflection Test Applying Item Response Theory (IRT). *Journal of Behavioral Decision Making*, *29*(5), 453–469. https://doi.org/https://doi.org/10.1002/bdm.1883

Purwanto. (2019). Educational Goals and Learning Outcomes: Domains and Taxonomies [in Bahasa]. *Jurnal Teknodik*, *9*(16), 146–164. https://doi.org/10.32550/teknodik.v0i0.541

Ramos, J. L. S., Dolipas, B. B., & Villamor, B. B. (2013). Higher order thinking skills and academic performance in physics of college students: A regression analysis. *International Journal of Innovative Interdisciplinary Research*, *4*, 48–60.

Retnawati, H. (2013). *Evaluation of educational programs* [in Bahasa]. Jakarta: Universitas Terbuka.

Retnawati, H. (2014). *Item response theory and its applications for researchers, measurement and testing practitioners, graduate students* [in Bahasa]. Yogyakarta: Nuha Medika.

Retnawati, H., & Hadi, S. (2014). The regional bank system is calibrated to welcome the era of decentralization [in Bahasa]. *Jurnal Ilmu Pendidikan*, *20*(2), 183–193. https://doi.org/10.17977/jip.v20i2.4615

Rizbudiani, A. D., Jaedun, A., Rahim, A., & Nurrahman, A. (2021). Rasch model item response theory (IRT) to analyze the quality of mathematics final semester exam test on system of linear equations in two variables (SLETV). *Al-Jabar: Jurnal Pendidikan Matematika*, *12*(2), 399–412. https://doi.org/10.24042/ajpm.v12i2.9939

Santoso, A., Kartianom, K., & Kassymova, G. K. (2019). Quality of statistics question bank items (Case study: Open University statistics course final exam instrument) [in Bahasa]. *Jurnal Riset Pendidikan Matematika*, *6*(2), 165–176. https://doi.org/10.21831/jrpm.v6i2.28900

Sarea, M. S., & Ruslan, R. (2019). Characteristics of Question Items: Classical Test Theory vs Item Response Theory? [in Bahasa]. *Didaktika : Jurnal Kependidikan*, *13*(1), 1–16. https://doi.org/10.30863/didaktika.v13i1.296

Sarvina, Y. (2017). Utilization of Open Source "R" Software for Agroclimate Research [in Bahasa]. *Informatika Pertanian*, *26*(1), 23–30. https://doi.org/10.21082/ip.v26n1.2017.p23-30

Shanti, M. R. S., Istiyono, E., Munadi, S., Permadi, C., Patiserlihun, A., & Sudjipto, D. N. (2020). Physics Question Assessment Analysis Using the Rasch Model with the R Program [in Bahasa]. *Jurnal Sains Dan Edukasi Sains*, *3*(2), 46–52. https://doi.org/10.24246/juses.v3i2p46-52

Subali, B., Kumaidi, Aminah, N. S., & Sumintono, B. (2019). Student Achievement Based On The Use Of Scientific Method In The Natural Science Subject In Elementary School. *Jurnal Pendidikan IPA Indonesia*, *8*(1), 39–51. https://doi.org/10.15294/jpii.v8i1.16010

Sudaryono. (2011). Implementation of Item Response Theory in the Assessment of Final Learning Outcomes in Schools [in Bahasa]. *Jurnal Pendidikan Dan Kebudayaan*, *17*(6), 719–732. https://doi.org/10.24832/jpnk.v17i6.62

Talebi, G. A., Ghaffari, R., Eskandarzadeh, E., & Oskouei, A. E. (2013). Item Analysis an Effective Tool for Assessing Exam Quality, Designing Appropriate Exam and Determining Weakness in Teaching. *Res Dev Med Educ*, *2*(2), 69–72. https://doi.org/10.5681/rdme.2013.016

Thepsathit, P., Jongsooksai, P., Bowornthammarat, P., Saejea, V., & Chaimongkol, N. (2022). Data Analysis in Polytomous Item Response Theory Using R. *Journal of Educational Measurement*, *28*(2), 13–26.

Tilaar, A. L. F., & Hasriyanti. (2019). Analysis of Odd Semester Question Items in Mathematics Subjects in Junior High Schools [in Bahasa]. *JP3I (Jurnal Pengukuran Psikologi Dan Pendidikan Di Indonesia)*, *8*(1), 57–68. https://doi.org/10.15408/jp3i.v8i1.13068

Tirta, I. M. (2015). Development of online interactive item response analysis using R for dichotomous responses with logistic models (1-PL, 2-PL, 3-PL) [in Bahasa]. *Seminar Nasional Pendidikan Matematika*. FKIP Universitas Jember.

Ulitzsch, E., Davier, M. von, & Pohl, S. (2020). A Multiprocess Item Response Model for Not- Reached Items due to Time Limits and Quitting. *Educational and Psychological Measurement*, *80*(3), 522–547. https://doi.org/10.1177/0013164419878241

Winarno, Zuhri, M., Mansur, Sutomo, I., & Widhyahrini, K. (2019). Development of Assessment for the Learning of the Humanistic Model to Improve Evaluation of Elementary School Mathematics. *International Journal of Instruction*, *12*(4), 49–64. https://doi.org/10.29333/iji.2019.1244a

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers*. Singapore: Springer Singapore.

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage Adaptive Testing Design in International Large-Scale Assessments. *Educational Measurement: Issues and Practice*, *0*(0), 1–12. https://doi.org/10.1111/emip.12226

Zainuddin, Z. (2018). Students' learning performance and perceived motivation in gamified flipped-class instruction. *Computers & Education*, *126*, 75–88. https://doi.org/10.1016/j.compedu.2018.07.003