RESEARCH ARTICLE

Analysis of biology midterm exam items using a comparison of the classical theory test and the Rasch model

Tanti Priyani [a,1], **Bowo Sugiharto**[a,2,*]

[a] Biology Education Masters Program, Faculty of Teacher Training and Education, Universitas Sebelas Maret, Jl. Ir. Sutami No. 36A Kentingan Surakarta, Central Java 57126, Indonesia

[1]tantipriyani@student.uns.ac.id; [2]bowo@fkip.uns.ac.id*

Abstract: In biology learning, test instruments are essential for assessing students' understanding of complex concepts. A test instrument is a crucial factor in learning evaluation; however, its implementation remains minimal. This descriptive quantitative study aims to analyze the quality of test items using the classical approach in terms of validity, reliability, difficulty index, discrimination power, distractor effectiveness, and the Rasch model analysis. The data consists of 30 multiple-choice questions from a biology midterm exam administered to 40 students. Classical test data analysis uses Microsoft Excel, while Rasch model analysis uses Winsteps software. The validity test results from both approaches show 14 valid questions and 16 invalid ones. The reliability scores are 0.619 (adequate) for the classical approach's Cronbach's Alpha, 0.85 (good) for the Rasch model, and 0.65 (weak) for personal reliability. The classical test theory and the Rasch model categorize item difficulty into four levels. The classical approach produces five categories for item discrimination, while the Rasch model identifies three groups based on the item separation index (H=3.45) and two groups based on respondent ability (H=1.96). Distractor effectiveness shows 93.3% functional distractors in the classical test and 80% in the Rasch model. The Rasch model offers greater precision in measuring student ability and detecting bias. Both models should be integrated for comprehensive item analysis. Future tests should focus on improving invalid items and the quality of distractors.

Keywords: difficulty level; distractor effectiveness; item discrimination; reliability; validity

For correspondence:
bowo@fkip.uns.ac.id

Article history:
Received: 17 June 2024
Revised: 1 October 2024
Accepted: 29 October 2024
Published: 18 November 2024

10.22219/jpbi.v10i3.34345

p-ISSN: 2442-3750
e-ISSN: 2537-6204

How to cite:
Priyani, T., & Sugiharto, B. (2024). Analysis of biology midterm exam items using a comparison of the classical theory test and the Rasch model. *JPBI (Jurnal Pendidikan Biologi Indonesia), 10*(3), 939-958
https://doi.org/10.22219/jpbi.v10i3.34345

Introduction

The quality of education is very important for the progress of a country. The quality of education is not only determined by the role of educators but also by the learning process. Educators play an important role in designing and implementing the learning process, as well as conducting evaluations and assessments to achieve learning objectives. The terms evaluation and assessment are often used interchangeably, although they have significant differences. Assessment involves collecting data on student performance through various methods, such as quizzes and reflective questions, to inform teaching and learning strategies (Vorwerk & Engenhart-Cabillic, 2022). Evaluation, on the other hand, synthesizes assessment data to make high-stakes decisions about student competence. This process requires high-quality narrative feedback to ensure robust evaluation (de Jong et al., 2022).
An educator must have the ability to develop assessment tools that are aligned with indicators of learning goal achievement (Milania & Murniati, 2022). In biology learning, midterm exams are critical for assessing students' understanding of complex biological concepts. However, in practice, the quality of the test items often does not reflect the intended learning outcomes, leading to an inaccurate measurement of student abilities. Many assessments do not adequately reflect the core concepts of biology, which is critical for integrated learning (Cliff, 2023). Educators must have assessment literacy to distinguish between Assessment for Learning (AfL) and Assessment of Learning (AoL), which affects their teaching practices and student engagement (Wu et al., 2021). To achieve this, educators must be

trained in the implementation of assessments, from tool creation to detailed analysis and decision-making based on the results obtained.

However, in reality, the lack of teachers' ability to create evaluation instruments causes the quality of the tested questions to be unknown. A major problem in the biology midterm exam is the lack of item quality analysis, which causes questions to be unable to accurately assess students' knowledge or skills. Poorly structured questions may not reflect the cognitive processes students use, leading to an inaccurate assessment of their knowledge and skills (Steiner & Frey, 2021). Inadequate evaluation tools can lead to significant differences in the assessment of student performance, thus undermining the validity of educational outcomes (Kok & Priemer, 2023). Such issues in biology midterm exams may hinder effective learning, as students may not receive accurate feedback. Inadequate feedback from flawed assessments may hinder students' understanding of their strengths and weaknesses. As a result, educators may face challenges in tailoring instruction to meet the needs of their students (Angell et al., 2024).

Evaluation is used to determine the quality of instruments and to ascertain whether learning objectives have been achieved. Therefore, analyzing the quality of biology midterm exam items is essential to ensure that the assessments accurately measure student learning and provide valuable insights for educators to improve instruction. Conducting item analysis is essential to improve the quality of the instrument to ensure that items have high validity and reliability. This is especially important for teachers who aim to accurately assess student understanding. Validity refers to how well an instrument measures what it is intended to measure (Jones et al., 2019). Reliability ensures consistent results across multiple contexts. The development of validated instruments demonstrates the importance of reliability in educational assessment (Batista et al., 2021). Other elements that support the quality of test items include difficulty level and discrimination power. Distractors in multiple-choice tests can also be used to measure students' understanding (Sumantri & Retni Satriani, 2016).

Item quality can be analyzed using Classical Test Theory (CTT) and Modern Test Theory (Retnawati, 2016). Modern Test Theory employs Item Response Theory (IRT), developed by Georg Rash, making modern theory known as the Rasch model (Retnawati, 2014). Many testing programs still refer to CTT in test design and result assessment due to several advantages of CTT over IRT. For example, CTT models can be implemented with basic statistical tools, unlike Item Response Theory (IRT), which often requires sophisticated software and expertise (Fischer & Rose, 2016). CTT can produce reliable estimates even with smaller sample sizes, which is beneficial in educational settings where data may be limited (Prenovost et al., 2018). CTT designs typically require minimal resources, making them accessible to a wide range of institutions, especially in resource-limited environments (Gray et al., 2020). CTT produces total scores that are easy to understand and interpret, making them easier to use in context practice (Morgan-López et al., 2020).

However, one of the disadvantages of CTT is that it assumes all items contribute equally to the total score, which can misrepresent the data, especially in heterogeneous populations (Sen et al., 2016). This does not allow test items to match ability levels. IRT models, such as the Rasch model, can handle various item characteristics and respondents' abilities, leading to more accurate assessments (R. Liu & Jiang, 2020). Using IRT, the ability level and difficulty index of items can be mapped on a line using a logit scale (Subali et al., 2019).

The Rasch model transforms raw scores into logit values, facilitating a direct comparison between item difficulty and person ability (Van Zile-Tamsen, 2017). This transformation is important for maintaining the integrity of parameter estimates, as it ensures that items and individuals are evaluated on the same scale (Finch & Edwards, 2016). The model allows for the independent estimation of person and item parameters, ensuring that ability measures are independent of the specific items used (Baghaei et al., 2017). Infit and outfit MNsQ are important fit statistics in Rasch analysis that assess how well each item fits the Rasch model. Values close to 1 indicate a good fit, while values outside the range of 0.5 to 1.5 indicate a mismatch (Poorebrahim et al., 2021). The Rasch model also allows for various ways to analyze differential item functioning (DIF) to detect item bias that may unfairly favor or disadvantage certain groups, thus increasing the likelihood of developing fair measurement scales (Hope et al., 2018). With the concepts of unidimensionality, local independence, reliability, and fit, Rasch measurement has contributed to research (Aryadoust et al., 2021).

Based on the above issues, research on the results of CTT and Rasch model analysis is essential to determine the extent to which these two approaches are effectively used in learning evaluation. This research aims to compare the analysis of validity, reliability, difficulty level, discrimination power of questions, and distractor effectiveness through both CTT and Rasch model approaches, and to identify the advantages of the Rasch model over classical test theory, including the detection of biased measurements and test takers' abilities. The instrument used in this research is the Mid-Semester Assessment test for Grade XII. An important effort to improve and develop evaluation instruments is the item analysis of the Mid-Semester Assessment instrument. This analysis also enhances the objectivity of tests to determine the achievement of learning objectives.

CTT and the Rasch Model each offer significant advantages in the analysis and evaluation of test

instruments. CTT enables simpler and faster data analysis, as the results can be easily calculated and interpreted (Fujimoto et al., 2019). In addition, CTT facilitates detailed item analysis, allowing educators to evaluate item quality based on difficulty and discrimination power, which contributes to the overall improvement of test quality (Saat, 2020). CTT also supports the validation of assessment instruments and promotes fairness and consistency in assessment, which are crucial on a broader educational scale (Jimam et al., 2019). Another advantage is its ability to provide useful information for educational decision-making through simple statistical method mapping (Aaij et al., 2022).
On the other hand, the Rasch Model offers a high level of reliability by enabling differential item functioning analysis, ensuring that test items function consistently across different groups (Bejerholm & Lundgren-Nilsson, 2015). Rasch also provides more accurate interval-level data, which supports more precise statistical analyses and interpretations (Robinson et al., 2019). In addition, the Rasch Model ensures that items measure one underlying trait (unidimensionality), which is critical to the validity of assessments and the interpretation of test results (Wilberforce et al., 2019). With its ability to identify items that function differently across different groups, the model increases fairness in scoring and optimizes response categories, resulting in more reliable measurements (Murphy et al., 2019).

Method

This research was conducted from February to March 2024, involving students from Class XII IPA 1 at Madrasah Aliyah Negeri 2 Kota Tangerang. The research population consisted of 40 students, comprising 28 female and 12 male participants. The study employed a descriptive quantitative method to analyze the quality of the assessment instrument based on several key elements: validity, reliability, difficulty level, discrimination power of questions, and distractor effectiveness. The analysis utilized both Classical Test Theory (CTT) and the Rasch model, facilitating a comprehensive evaluation of the assessment tool. The assessment instrument utilized in this study was the Mid-Semester Assessment for the Biology subject, consisting of 30 multiple-choice questions. Each question was carefully designed to assess specific biological concepts based on the curriculum objectives for Class XII students. These objectives covered a range of topics, from plant growth to metabolic processes, and the questions were grouped based on corresponding indicators. Table 1 summarizes the indicators and corresponding questions.

Table 1. Summary of Indicators and Corresponding Questions

Indicator	Questions Number
Understanding the Characteristics of Plant Growth	1, 2, 3
Identifying Types of Seed Germination	4, 5
Understanding Plant Hormones and Their Effects on Growth	6, 10
Analyzing the Role of Light and Hormones in Plant Growth	7, 8, 9, 11
Understanding Catabolic Reactions and Metabolic Processes	12, 14
Analyzing the Properties and Functions of Enzymes in Metabolism	13, 15, 16
Understanding the Stages of Aerobic and Anaerobic Respiration	17, 18, 20, 21, 22
Understanding Photosynthesis and Its Reactions	19, 23, 24, 25
Comparing Different Types of Fermentation	28
Analyzing the Glycolysis Pathway and Subsequent Reactions	29
Understanding the Role of CO_2 and ATP in Photosynthesis and Respiration	26, 27, 30

The data collected from the assessment were analyzed using both Classical Test Theory (CTT) and the Rasch model. CTT analysis was conducted using Microsoft Excel, focusing on the validity, reliability, difficulty level, discrimination power, and distractor effectiveness of the questions. The Rasch model analysis was performed using Winsteps software version 4.5.2, enabling a more sophisticated evaluation of item characteristics and providing insights into the advantages of Rasch analysis over CTT, including its ability to detect biased measurements and accurately assess the abilities of test-takers.
In the Rasch Model, item discrimination is analyzed based on the individual ability levels of test takers. This discrimination reflects how well each item can distinguish between high and low-ability participants. In addition, the respondent separation index is used to identify different groups of respondents based on their ability levels. To determine this separation, the strata equation (H) is used, which is calculated by the Formula 1.

$$H = \frac{[(4\ x\ separation\ )+1]}{3} \quad (1)$$

The H value provides information about the number of ability groups of respondents that can be identified based on the Rasch analysis results. Thus, the Rasch Model provides a more in-depth view of the ability distribution of test takers and the suitability of items in the assessment instrument.

Results and Discussion

Validity

Instrument validity is a crucial element in research to ensure that the measurement tool accurately assesses the intended construct. In educational instrument analysis, validity takes several main forms, including content validity, construct validity and criterion validity. Content validity assesses the extent to which an instrument covers relevant and comprehensive content areas, ensuring that every important aspect of the theoretical construct it seeks to measure is represented. Instruments that have good content validity reflect the learning objectives appropriately (Echevarría-Guanilo et al., 2019).

Construct validity, on the other hand, evaluates whether the measuring instrument assesses the theoretical construct it is intended to measure, as described by Mazurek et al. (2020). Construct validity can be evaluated through factor analysis or other statistical methods to ensure that the items in the instrument measure the expected dimensions. In the Rasch model, construct validity is evaluated using Item Outfit Mean Square (MNSQ) and Item Outfit Z-Standard (ZSTD), which compare the expected response with the observed response (Köhler & Hartig, 2017). The MNSQ Infit and Outfit statistics help assess how well items fit the expected model, ensuring that each item contributes meaningfully to the overall construct being measured (Stanley & Edwards, 2016). An item is considered valid if it meets the following three criteria:

a) Outfit MNSQ (Mean Square) Value : 0,5 < Outfit - MNSQ < 1,5
b) Outfit ZSTD (Z-Standard) Value : -2,0 < ZSTD < +2,0
c) Point Measure Correlation Value : 0,4 < Point Measure Corr < 0,85 (Sumintono & Widhiarso, 2015). This indicates that the item measures the same construct across student groups, ensuring consistency of measurement.

Criterion validity involves the correlation between the results of a measurement tool and another measure that has been recognized as a standard. Criterion validity can be obtained by predicting the results of one test with the results of another test that is considered relevant (Ayres et al., 2021). In Classical Test Theory (CTT), criterion validity is often assessed through Pearson's product-moment correlation (r). An item is considered valid if the r value obtained exceeds the critical value of the table, indicating a significant correlation with the total score (Ronk et al., 2016). On the other hand, the Rasch model provides a more detailed approach through Point Measure Correlation analysis, which assesses the relationship between student ability and items, strengthening the overall validity of the instrument. The results of the Rasch model analysis can be seen in Table 2 and Table 3.

From the 30 test items analyzed using CTT and the Rasch model, 14 items were found to be valid, while 16 items were categorized as invalid. Although both models identified a similar number of valid and invalid items, the underlying reasons for invalidity varied, providing insight into the strengths and weaknesses of each measurement model.

CTT identifies the validity of test items primarily based on the correlation between students' performance on individual items and their total test scores. While this method is straightforward and provides a general overview of item validity, it is sensitive to the sample used and may not account for bias in individual items. For example, in items 17 and 18, CTT could not provide meaningful results, as these items had a division by zero error (marked as #DIV/0!), indicating that all students answered them correctly or incorrectly, preventing further analysis. This limitation in CTT means that it cannot handle items where performance is homogeneous across test-takers (Zlatkin-Troitschanskaia et al., 2017).

In contrast, the Rasch model offers several advantages over CTT by focusing on each test item independently of the test-taker's total score. This model can detect biased items and identify the measurement error associated with individual test items (Eden, 2018). For items 17 and 18, the Rasch model identified these as "item-free person" items, meaning they did not contribute to distinguishing between high- and low-ability students. The Rasch model was able to indicate that these items should either be revised or discarded, as they provided no meaningful differentiation in terms of student ability. Moreover, Rasch's analysis provides detailed information about the difficulty level of each item and how well each item fits within the overall test framework.

Each question in the assessment instrument was designed to evaluate specific indicators related to biological concepts. The valid questions effectively measured students' understanding of the corresponding indicators, while the invalid questions failed to provide meaningful results due to various factors, including their inability to discriminate between different levels of student ability (Tzafilkou et al., 2022).

Table 2. Output Table for Validity Testing Using Winsteps Application

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL	INFIT		OUTFIT		PTMEASUR		EXACT MATCH		Item
					MNSQ	ZSTD	MNSQ	ZSTD	AL CORR	EXP	OBS%	EXP%	
3	7	40	3.36	.44	.92	-.21	.71	-.62	.45	.33	85.0	83.4	S3
28	14	40	2.28	.36	1.31	1.97	1.83	3.23	-.02	.38	60.0	70.4	8
22	15	40	2.15	.35	1.35	2.30	1.38	1.81	.03	.38	57.5	69.4	S22
21	18	40	1.79	.34	1.16	1.35	1.31	1.75	.19	.38	60.0	66.8	S21
23	19	40	1.67	.34	.85	-1.30	.81	-1.17	.53	.38	80.0	66.2	S23
26	23	40	1.20	.34	1.02	.21	.97	-.12	.36	.37	62.5	66.1	S26
16	29	40	.44	.38	.9	-.56	.78	-.72	.46	.33	80.0	74.4	S16
20	29	40	.44	.38	1.05	.33	1.12	.48	.27	.33	75.0	74.4	S20
24	29	40	.44	.38	1.20	1.12	1.21	.77	.14	.33	65.0	74.4	S24
1	30	40	.29	.39	.98	-.05	.89	-.25	.36	.32	75.0	76.5	S1
7	30	40	.29	.39	1.03	.20	1.02	.12	.30	.32	75.0	76.5	S7
29	30	40	.29	.39	.78	-1.21	.64	-1.19	.57	.32	80.0	76.5	S29
27	31	40	.13	.40	.87	-.56	.74	-.68	.46	.31	82.5	78.6	S27
30	31	40	.13	.40	.87	-.55	.73	-.73	.46	.31	82.5	78.6	S30
13	32	40	-.03	.42	.77	-.97	.60	-1.06	.55	.30	85.0	80.7	S13
2	34	40	-.42	.46	.82	-.53	.59	-.81	.48	.27	87.5	85.3	S2
10	34	40	-.42	.46	1.17	.65	.62	1.21	.04	.27	82.5	85.3	S10
9	35	40	-.65	.50	.98	.06	.83	-.12	.28	.26	87.5	87.5	S9
19	35	40	-.65	.50	.87	-.29	.75	-.29	.38	.26	87.5	87.5	S19
6	36	40	-.92	.55	1.10	.37	.90	.07	.17	.24	90.0	90.0	S6
15	36	40	-.92	.55	.89	-.15	.66	-.36	.36	.24	90.0	90.0	S15
25	36	40	-.92	.55	1.15	.48	1.22	.53	.08	.24	90.0	90.0	S25
4	37	40	-1.26	.62	1.10	.35	.93	.17	.14	.21	92.5	92.5	S4
8	37	40	-1.26	.62	1.0	.20	.64	-.25	.26	.21	92.5	92.5	S8
12	37	40	-1.26	.62	.91	-.04	.72	-.13	.30	.21	92.5	92.5	S12
14	37	40	-1.26	.62	.87	-.11	.64	-.26	.35	.21	92.5	92.5	S14
5	39	40	-2.46	1.03	1.02	.33	.64	.14	.14	.13	97.5	97.5	S5
11	39	40	-2.46	1.03	1.00	.31	.52	.01	.18	.13	97.5	97.5	S11
17	40	40	-3.69	1.83	MINIMUM MEASURE				.00	.00	100.0	100.0	S17
18	40	40	-3.69	1.83	MINIMUM MEASURE				.00	.00	100.0	100.O	S18
MEAN	30.6	40.0	-.25	.58	1.00	.1	.91	.1			81.6	81.9	
P. SD	8.3	.0	1.60	.37	.15	.8	.32	1.0			11.4	9.5	

Table 3. Results of Analysis Using CTT and Rasch Model

No.	Result	Question Number	
		Classical Test Theory	Model Rasch
1.	Valid	1, 2, 3, 12, 13, 14, 15, 16, 19, 23, 26, 27, 29, 30	1, 2, 3, 13, 14, 15, 16, 19, 23, 24, 26, 27, 29, 30
2	Not Valid	4, 5, 6, 7, 8, 9, 10, 11, 17, 18, 20, 21, 22, 24, 25, 28	4, 5, 6, 7, 8, 9, 10, 11, 12, 17, 18, 20, 21, 22, 25, 28

The valid questions, such as 1, 2, and 3, measured students' understanding of plant growth characteristics, successfully differentiating between students with varying levels of comprehension. Questions 13 through 16, which focused on enzyme functions in metabolism, also demonstrated high validity by providing clear insights into students' grasp of complex biological processes.

Many of the invalid questions, such as 4 through 11, failed to discriminate between students of different abilities. For example, question 6, related to plant hormone functions, was invalid in both models due to the homogeneity of student responses, indicating a lack of complexity or ambiguity in the question design. These items may require revision to introduce more nuanced distractors or to cover more challenging aspects of the concepts being tested.

Invalid questions, particularly items 17 and 18, provided critical insights into the limitations of both CTT and Rasch analysis. In CTT, these items resulted in undefined values due to the lack of variance in student responses (all students answered correctly). In the Rasch model, the same items were flagged as problematic because they failed to discriminate between high and low-ability students, as all participants performed uniformly (De Sá et al., 2019). This lack of differentiation means the items did not effectively measure the intended construct. The Rasch model, however, was able to identify these issues more precisely, suggesting that the items did not contribute meaningful information to the overall test and should be revised or excluded from future assessments (Korbee et al., 2022).

Additionally, the Rasch model revealed that some questions, such as item 12, were valid according to CTT but borderline in terms of Rasch's item-fit analysis. This discrepancy suggests that while the question correlated well with the total score in CTT, it did not perform as well in differentiating ability levels according to the Rasch model. Therefore, these items should be carefully reviewed to determine whether minor revisions could enhance their effectiveness.

Based on these findings, invalid items could be revised to improve their ability to differentiate between different levels of student understanding. This may include: 1) revising distractors to make incorrect answers more plausible; 2) adjusting question-wording to avoid ambiguity or unexpected clues; and 3)

ensuring that items cover a range of difficulty levels to differentiate between students of different abilities. In addition, incorporating more challenging aspects of biological concepts can help improve the validity and reliability of test items (Blacquiere & Hoese, 2016).

Reliability

Instrument reliability is an important aspect in the evaluation of measuring instruments to ensure the consistency and stability of the results obtained. Reliability, defined as the ability of an instrument to produce consistent results across a range of conditions, is crucial in ensuring that the data generated is reliable in scientific research (Babu & Kohli, 2023). Good reliability ensures that the measurement tool can be used repeatedly in various situations without compromising the validity of the findings obtained. Strong reliability allows researchers to draw more accurate and trustworthy conclusions, contributing significantly to the quality of research and reliable data-based decision-making (Mohajan, 2017).

The reliability of testing instruments is often assessed through Cronbach's Alpha, a method in Classical Test Theory (CTT) that measures internal consistency between items in an instrument. A test item is considered reliable if it has a correlation coefficient value of $-1.00 \leq \rho \leq +1.00$ (Retnawati, 2016). An alpha coefficient above 0.7 is generally considered to indicate acceptable consistency, with coefficients between 0.79 and 0.84 reflecting good reliability (Jacobs et al., 2017). Reliability is categorized as weak if < 0.67, while > 0.94 is considered excellent (Sumintono & Widhiarso, 2015).

Strong reliability indicates that the instrument can capture stable traits in individuals, even when they are faced with different challenges or conditions (Baldan et al., 2021). In the Rasch model, reliability is assessed through two main indicators: person reliability and item reliability, which provide insight into measurement reliability at both the student and item levels. The Rasch model also uses the person separation index to group respondents based on their ability, providing an additional dimension in assessing measurement consistency. In this analysis, higher Rasch reliability values compared to CTT indicate superiority in instrument measurement consistency (Królikowska et al., 2023).

Reliability testing in the Rasch model can be analyzed from Table 4 output in the Winsteps application. The Table 4 was the results of the reliability analysis of the PTS instrument using CTT and the Rasch Model.

Table 4. Results of Reliability Analysis via CTT and Rasch Model

Reliability CTT		**Reliability of the Rasch Model**			
Alpha Value Cronbach	Category	Person Reliability	Category	Item Reliability	Category
0.619	Adequate	0.63	Weak	0.85	Good

The reliability analysis provides crucial insights into the consistency of the test instrument. In Classical Test Theory (CTT), a Cronbach's alpha value of 0.619 was obtained, which falls under the "adequate" category according to commonly accepted reliability standards (Retnawati, 2016). Cronbach's alpha of 0.619, while adequate, indicates that test items may not consistently measure the intended concept, requiring refinement (Elliott et al., 2020). This indicates that while the test demonstrates some degree of internal consistency, there is still room for improvement to ensure that the items consistently measure the intended biological concepts. Higher Cronbach's alpha values, closer to 1.0, would indicate higher reliability, Cronbach's values above 0.9 indicate excellent reliability (Van Vliet et al., 2021). However, the current value indicates that the test can provide a satisfactory measure of student knowledge.

In contrast, the Rasch model offers a more detailed view of test reliability. The item reliability value obtained is 0.85, the Rasch model item reliability value of 0.85 is classified as “good”. This indicates that the items used in the midterm exam are strong in measuring the targeted biology skills. The Rasch model's ability to evaluate each item individually provides valuable information about which items perform well and which need revision. In addition, the person reliability score of 0.63, which is classified as “weak”, indicates that students' answers lacked consistency. An individual reliability score of 0.63 is considered “weak”, indicating variability in students' answers (Lewis et al., 2020). This may suggest that the biology exam questions may be too difficult or too easy for some students, leading to inconsistencies in their answers.

The difference in reliability results between CTT and the Rasch model can be attributed to their distinct focuses. While CTT evaluates the overall test, the Rasch model offers a more granular analysis down to the performance of individual items, providing a clearer picture of which items contribute to or detract from the test's reliability. This is one of the Rasch model's main advantages, as it allows for more targeted revisions and improvements in item quality (Gaitán-Rossi et al., 2021).

The biology test items included in this study appear to be realistic in covering understanding biological concepts. However, the relatively low reliability values from the Rasch analysis suggest that some items may not be optimal in differentiating between students of different ability levels. This may indicate that some questions were too easy or difficult, or some distractors were ineffective, leading students to guess the correct answer instead of choosing the correct one based on their understanding.

To address this, further item-level analysis, particularly using the Rasch model, can be conducted to revise questions that do not contribute meaningfully to the accurate measurement of students' knowledge and skills. By adjusting the difficulty level of specific items and ensuring that all distractors are valid, the reliability and validity of the test as a whole can be improved (Sumintono & Widhiarso, 2015).

Difficulty Level

The difficulty index (p-value) is calculated as the ratio of correct responses to the total number of responses. This index ranges from 0-100% or 0.0-1.0 (Musa et al., 2021), as in Table 5. A higher index (closer to 1) indicates that more participants answered the item correctly (Seide et al., 2019). A higher difficulty index correlates with a greater percentage of correct answers, indicating that items with higher values are easier for participants. This is important for creating a balanced assessment that can effectively differentiate between different levels of knowledge (Orozco et al., 2022). In CTT, the criteria for the item difficulty index are as follows (Lestari & Yudhanegara, 2015).

Table 5. Item Difficulty Index in Classical Test Theory (CTT)

Difficulty Index	Interpretation
IK = 0.00	Very difficult
0.00 < IK ≤ 0,30	Difficult
0.30 < IK ≤ 0.70	Moderate
0.70 < IK ≤ 1.00	Easy
IK = 1.00	Very easy

Meanwhile, in the Rasch model, the difficulty level of an item is based on its size value. Items with a measure < -1 are considered very easy, indicating that most respondents can answer them correctly with little effort (Gay et al., 2016). Items that range from -1 to 0 are classified as easy, indicating that they are accessible to most respondents (Nielsen et al., 2017). Items with a measure between 0.1 and 1 are considered difficult, reflecting a higher level of challenge for respondents. Items that exceed a measure of 1 are categorized as very difficult, indicating that only a few respondents can answer them correctly (Rodrigo et al., 2019). Difficulty level analysis in the Rasch model can be observed from the measure values listed in Table 6 (discussed under validity). Based on the CTT and Rasch Model analysis results, the data for difficulty levels are as follows.

Table 6. Difficulty Level Analysis Results via CTT and Rasch Model

CTT Difficulty Level			Rasch Model Difficulty Level		
Category	Number of Questions	Total Number of Questions	Category	Number of Questions	Total Number of Questions
Very difficult			Very difficult	3, 21, 22, 23, 26, 28	6
Difficult	3	1	Difficult	1, 7, 16, 20, 24, 27, 29, 30,	8
Moderate	21, 22, 23, 26, 28	5	Moderate		
Easy	1, 2, 4, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 19, 20, 24, 25, 27, 29, 30	20	Easy	2, 6, 9, 10, 13, 15, 19, 25	8
Very easy	5, 11, 17, 18	4	Very easy	4, 5, 8, 11, 12, 14, 17, 18	8

The analysis showed significant differences between the CTT and Rasch models in categorizing item difficulty. In CTT, the majority of items (20) were classified as easy, whereas in the Rasch model, only 8 items fell into this category. In addition, the Rasch model identified 6 items as very difficult, while CTT did not identify items in this category.

The Rasch model's identification of 6 very difficult items suggests that these items may require a higher level of cognitive ability than other items, potentially making them unsuitable for the intended population. These items, particularly item numbers 3, 21, 22, 23, 26, and 28, should be reviewed to ensure that they are appropriate for the knowledge and skills expected of students. Similarly, the identification of 8 very easy items (items 4, 5, 8, 11, 12, 14, 17, and 18) by the Rasch model raises concerns about whether these items effectively measure the targeted concepts or are too easy for most students.

CTT and the Rasch model show that the two models provide different perspectives on item difficulty, CTT relies on aggregate data, which can obscure individual item performance and lead to biased

interpretations of item difficulty (O'Brien et al., 2021). Whereas the Rasch Model evaluates items based on their ability to differentiate between different ability levels of respondents, offering a more nuanced understanding of item difficulty (Kassim et al., 2023). The model also allows for the identification of differential item functioning, which CTT does not adequately address.

Item Discrimination

Item discriminating power measures how effectively an item can distinguish between high and low-ability students on a test. High discriminating power indicates that the item is effective in identifying differences in ability levels among test-takers (McKeigue, 2019). The discrimination index ranges from 0.00 to 1.00, where higher values indicate better discriminating power. The discrimination index is essential to ensure that the assessment effectively distinguishes different levels of ability among test takers, which is important for accurate measurement in an educational setting (Teresi et al., 2021).

The results of the discrimination index analysis using CTT, as shown in Table 7, indicate that most of the items in the test exhibited low discriminating power, with 15 items falling into the 'Low' category (D ≤ 0.20). Items that are too easy or difficult do not provide meaningful information about the ability of test takers, resulting in a low discriminating power index (Hagquist, 2019). Meanwhile, 8 items fell into the 'Medium' category (0.20 < D ≤ 0.40), 5 items fell into the 'High' category (0.40 < D ≤ 0.70), and only 1 item showed 'Very High' discriminating power (D > 0.70). High discriminability in test items enables accurate measurement of the underlying construct, ensuring that judgments reflect true differences in ability or symptoms rather than demographic bias (De Sá et al., 2019). High item discrimination contributes to test reliability and validity, as it ensures that items are not biased towards certain groups, thus increasing fairness in assessment (Liuzza et al., 2021).

Table 7. Criteria for Differentiating Questions and Analysis Results Using CTT

The magnitude of the D grade	Differentiating Power Category	Number of Questions	Total Number of Questions
D ≤ 0	Very low	22	1
0 < D ≤ 0.2	Low	4, 5, 6, 9, 10, 11, 12, 14, 5, 17, 18, 21, 24, 25, 28	15
0.2 < D ≤ 0.4	Fair / Medium	1, 7, 8, 19, 20, 26, 27, 30	8
0.4 < D ≤ 0.7	Tall	2, 3, 13, 16, 23	5
0.7 < D ≤ 1	Very high	29	1

In the Rasch Model, item discrimination is analyzed based on the ability levels of individual test-takers. Additionally, respondent separation indices can be used to identify respondent groups. To determine grouping, the strata equation (H) is utilized, as Formula 2.

$$H = \frac{[(4\ x\ separation\ )+1]}{3} \quad (2)$$

The Rasch model provides a more sophisticated item discrimination analysis. The model calculates the item and respondent separation index to determine how well test items differentiate between different ability groups (Garrido et al., 2019). The item separation index measures the ability of test items to differentiate between groups, while respondent separation indicates the degree of differentiation among test takers. The item separation index measures item discrimination, indicating how well test items can differentiate between individuals of different ability levels (Jin & Jeon, 2019).

The item separation value was 2.34, resulting in a stratum (H) value of 3.45, indicating that the items could categorize students into three different ability levels. In addition, the respondent separation was calculated as 1.22, with H = 1.96, indicating that there are two groups of respondents with different ability levels. The separation values are observed from the output table in the Winsteps application, as in Table 8.

The findings of Li et al. (2016) highlight that high item separation values indicate that test items are well-directed and sufficiently challenging for the population tested. High item separation values indicate that items are effective in distinguishing between high- and low-ability respondents, improving the overall reliability of the assessment. The item reliability index of 0.85 in Table 8 further confirms the quality and consistency of the test items in differentiating student ability.

The Rasch model offers a powerful approach to analyzing item discrimination, enabling a deeper understanding of student performance. Compared to CTT, which relies solely on percentage-based item indices, the Rasch model incorporates a probabilistic framework that adapts to test-takers individual abilities, ensuring a more accurate and meaningful interpretation of test items (Nielsen et al., 2017). The results show that while there are some areas for improvement in test design (e.g., items with low discrimination in the CTT), the Rasch model provides deeper insights into how these items function in

differentiating groups of students, which is critical to ensuring the validity and reliability of assessments.

Table 8. Problem Differentiation Analysis Using the Rasch Model

Summary of 40 Measured Person								
	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	23.0	30.0	1.57	.55	.99	.07	.91	.06
SEM	.5	.0	.15	.02	.04	.13	.07	.14
P.SD	3.2	.0	.92	.12	.23	.83	.46	.85
S.SD	3.3	.0	.93	.12	.24	.84	.47	.86
MAX.	29.0	30.0	4.19	1.08	1.55	1.88	2.04	2.47
MIN.	13.0	30.0	-.63	.44	.52	-1.40	.20	-1.2
REAL RMSE	.58	.58	TRUE SD	.71	SEPARATION		Person REABILITY	.60
MODEL RMSE	.56	.56	TRUE SD	.72	SEPARATION		Person REABILITY	.62
S.E. Of Person MEAN =		= .15						
Person RAW SCORE-TO-MEASURE CORRELATION = .97								
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" REABILITY = .63 SEM = 1.97								

Summary of 28 Measured Items								
	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	30.0	40.0	.00	.49	1.00	.13	.91	.06
SEM	1.6	.0	.26	.03	.03	.16	.06	.19
P.SD	8.2	.0	1.36	.17	.15	.81	.32	.97
S.SD	8.3	.0	1.38	.18	.15	.83	.33	.99
MAX.	39.0	40.0	3.36	1.03	1.35	2.30	1.83	3.23
MIN.	7.0	40.0	-2.46	.34	.77	-1.30	.52	-1.19
REAL RMSE	.53	TRUE SD	1.25	SEPARATION	2. 34	Item REABILITY		.85
MODEL RMSE	.52	TRUE SD	1.25	SEPARATION	2.40	Item REABILITY		.85
S.E. Of Person MEAN		= .26						

Effectiveness of Distractors

Multiple-choice tests are widely used in the learning assessment process due to their ability to streamline the evaluation process and provide rapid assessment of a wide range of knowledge and skills (Trampush et al., 2017). The objective nature of multiple-choice tests also minimizes bias in scoring, resulting in a clearer assessment of students' cognitive abilities (Barbic et al., 2018). In multiple-choice tests, the use of good stems and plausible answers (options) along with incorrect choices (distractors) can reveal underlying misconceptions among test-takers (F. Wang et al., 2020). Therefore, effective distractors become an important element in assessing students' conceptual understanding. Well-constructed distractors not only challenge test-takers to think critically but also provide a clear framework for evaluating knowledge, both through recall and recognition (Ye et al., 2022).

Distractor efficiency has a significant impact on the Difficulty Index and Discrimination Power of multiple-choice items. Recent research has shown that items with efficient distractors tend to have lower difficulty and higher discrimination power, indicating their effectiveness in differentiating between students' diverse levels of understanding (Rezigalla et al., 2024). Thus, in both Rasch analysis and the Classical Theory of Tests, effective distractors not only serve as a diagnostic tool to evaluate student understanding but also improve item quality in terms of more valid and reliable measurement.

The effectiveness of distractors in the Rasch model can be observed in the Table 9.

Table 9. The Snippet of the Results of the Analysis of the Rasch Model Representing the Effectiveness of the Distractor

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA		ABILITY		S.E. MEAN	INFT MNSQ	OUTF MNSQ	PTMA CORR.	Item
			COUNT	%	MEAN	P.SD					
3	E	0	2	5	.99	.45	.45	.4	.5	-.15	S3
	C	0	7	18	1.07	.74	.30	.6	.6	-.25	
	B	0	3	8	1.46	.35	.25	.6	.7	-.03	
	A	0	21	53	1.51	.88	.20	1.1	1.1	-.07	
	D	1	7	18	2.46	.81	.33	.8	.7	.45	
28	B	0	13	33	1.34	.67	.19	1.0	.9	-.18	S28
	A	0	1	3	1.44	.00		.7	.8	-.02	
	C	0	8	20	1.84	1.06	.40	1.7	2.5	.15	
	D	0	4	10	1.91	46	.27	1.3	1.4	.12	
	E	1	14	35	1.54*	1.07	.30	1.5	2.0	-.02	
22	A	0	1	3	-.63	.00		.1	.1	-.38	S28
	D	0	24	60	1.64	.76	.16	1.3	1.3	.09	
	B	1	15	38	1.60*	.98	.26	1.5	1.4	.03	
11	A	0	1	3	.54	.00		1.0	.5	-.18	S11
	D	1	39	98	1.60	.91	.15	1.0	1.0	.18	
17	E	1	40	100	1.57	.92	.15	.0	.0	.00	S17

Table 10 is a comparison table of the effectiveness of the distractor of PTS questions that have been

analyzed using CTT and the Rasch Model.

Table 10. Comparison of Effectiveness of Distractors in PTS Items Analyzed Using CTT and Rasch Model

Distractor Effectiveness	Classical Test Theory			Rasch Model		
	Number of Questions	Total of Number Questions	Percentage	Number of Questions	Total of Number Questions	Percentage
Functions	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30	28	93,3%	1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 19, 20, 21, 23, 24, 26, 27, 29, 30	24	80%
Not functioning	17, 18	2	6.7%	10, 17, 18, 22, 25, 28	6	20%

In CTT, distractors are considered effective based solely on student selection, suggesting a direct measure of distractor performance (Gao et al., 2019). The Rasch model evaluates distractors not only by student choice but also by their impact on overall ability estimates, providing a more nuanced understanding of distractor effectiveness (Cecilio-Fernandes et al., 2017). This model allows for detailed analyses of item functioning, revealing how distractors can affect the measurement of students' true abilities (Goh et al., 2017). High-quality distractors play an important role in improving test diagnostics (Blanco et al., 2023), and the use of overly obvious distractors can reduce test performance by increasing correct responses higher than they should be (Gao et al., 2019).

For example, in question 3, 7 students (18%) answered correctly by selecting option D, while distractor A was selected by 21 students (53%), B by 3 students (8%), C by 7 students (18%), and E by 2 students (5%). This shows that the distractors are functioning well (effective). According to Yow & Priyashri (2019), well-constructed distractors encourage students to think critically by evaluating each option in more depth, rather than relying solely on memorization.

In the Rasch model analysis, the effectiveness of distractors is also seen from the increase in average ability. In question 3, the average proficiency for each option was E = 0.99, C = 1.07, B = 1.46, A = 1.51, and D (correct answer) = 2.46, indicating an increase in average proficiency and that the distractor was effective. In contrast, in question number 28, the average proficiency for each option was 1.34, 1.44, 1.84, 1.91, and 1.54, indicating a decrease in average proficiency, so the distractors were considered ineffective. Papenberg & Musch (2017) emphasize that tests with few high-quality distractors are more effective than many low-quality distractors, as they can produce more reliable scores.

Furthermore, in question 17, all test takers chose option E, which indicates that the distractors were not functioning properly. Distractors that are too similar to the correct answer can cause increased anxiety and confusion in students, negatively impacting their performance (Rogowska et al., 2022).

Overall, analyzing the effectiveness of distractors using the Rasch model is more comprehensive because it considers changes in student ability more accurately, thus being able to distinguish between students who have mastered the material and those who have not. Therefore, the use of this model can improve the diagnostic ability of a test (Blanco et al., 2023).

Differential Item Functioning (DIF) or Measurement Bias

Differential Item Functioning (DIF) refers to a situation where individuals from different groups, but having the same underlying ability, do not have the same probability of answering an item correctly. This signals a bias in measurement and may result in unfairness in the assessment results. DIF is an important issue in measurement using Rasch models and Item Response Theory (IRT), which assume that the probability of answering an item correctly depends only on individual ability, not on group characteristics (Goel & Gross, 2019). If there is a violation of this measurement invariance assumption, there will be misinterpretation of scores and potential bias toward certain groups (Y. Liu et al., 2019). Therefore, DIF detection is essential to maintain validity and fairness in tests, especially in the context of educational assessment.

DIF can be caused by various factors, including differences in culture, language, or group characteristics such as gender and socioeconomic status (Cheema, 2019; Runge et al., 2019). For example, on some items, biological or cultural differences can cause certain groups to have different responses to the same item (Hagquist, 2019). Detection and treatment of DIF can be done using statistical analyses such as the Rasch model, which allows for the identification of items that function differently among such groups (Chen et al., 2019). In the Winsteps application, bias is observed through Item: DIF, between/within in

the output Table 11 (Sumintono, 2016).

Table 11. Results of DIF Analysis Using the Rasch Model

Person Classes	SUMMARY DIF	D.F.	PROB.	BETWEEN-CLASS/GROUP		Item	Name
				UNWTD	MNSQ		
2	.0231	1	.8791	.0239	-1.04	1	S1
2	1.1041	1	.2934	1.1971	.60	2	S2
2	.4937	1	.4823	.5319	.07	3	S3
2	1.7433	1	.1867	1.9850	1.02	4	S4
2	.1414	1	.7069	.4685	.00	5	S5
2	1.3214	1	.2503	2.1741	1.10	6	S6
2	.0231	1	.8791	.0239	-1.04	7	S7
2	.8855	1	.3467	1.5561	.81	8	S8
2	2.0448	1	.1527	2.3078	1.15	9	S9
2	.0107	1	.9175	0.101	-1.19	10	S10
2	.1414	1	.7069	.4685	.00	11	S11
2	.8855	1	.3467	1.5561	.81	12	S12
2	.2241	1	.6359	.2349	-.34	13	S13
2	.8855	1	.3467	1.5561	.81	14	S14
2	.0876	1	.7673	.0914	-.69	15	S15
2	.1612	1	.6880	.1684	-.48	16	S16
2	.3591	1	.5490	.3813	-.11	19	S19
2	.1612	1	.6880	.1684	-.48	20	S20
2	1.9648	1	.1610	2.2077	1.11	21	S21
2	1.9601	1	.1615	2.1852	1.10	22	S22
2	2.5445	1	.1107	2.9175	1.38	23	S23
2	.1516	1	.6970	.1583	-.50	24	S24
2	.7060	1	.4008	.7566	.28	25	S25
2	.0969	1	.7556	.1003	-.66	26	S26
2	5.6410	1	.0175	7.2570	2.46	27	S27
2	2.6467	1	.1038	3.0276	1.42	28	S28
2	.0231	1	.8791	.0239	-1.04	29	S29
2	.5190	1	.4713	.5512	.09	30	S30

The results of the DIF analysis using the Rasch model, as shown in Table 11, identified that item S27 had a probability value of 0.0175 (smaller than 5%). A common threshold for identifying significant DIF is a probability value of 0.05, which indicates a potential bias in item responses (Wyse & Mapuranga, 2009). This indicates a significant difference in the item responses between the tested groups, which may suggest the presence of sex-related bias in item S27. To maintain measurement fairness, this item needs to be revised so as not to disadvantage one group, in this case, a particular gender group.

In general, the other items analyzed in this table showed no significant DIF, as indicated by probability values greater than 0.05 for most items. This indicates that the items function equally among the groups tested and do not require further revision.

DIF is an important indicator in ensuring the validity of assessment instruments. In this context, DIF analysis helps in maintaining that the multiple-choice test used can provide a fair measurement without bias towards a particular group. Previous studies support the importance of considering demographic factors such as gender and age in DIF analysis (De Sá et al., 2019), which was also found in this study on item S27. Detection and management of DIF will improve the overall validity and reliability of the test, as well as ensure fairness to all test takers.

Individual Abilities

Individual abilities in the Rasch Model refer to participants' abilities or skills measured based on their responses to test items. Each participant is mapped on a logit scale, which describes the individual's level of ability relative to the difficulty of the test item. This logit scale allows for a more accurate analysis of how much ability a participant has to solve a given item, with participants who have higher logit values considered to have a higher ability to solve the item than participants with lower logit values. The Wright Map is a visual tool that helps map individual ability against item difficulty, thus facilitating a clearer interpretation of the interaction between participant ability and item characteristics (Le et al., 2022; Nielsen et al., 2017; Parra-Anguita et al., 2019).

This Rasch Model provides the advantage of separating individual ability and item difficulty on the same scale, allowing visualization of the distribution of test-taker ability and item difficulty simultaneously (Tesio et al., 2024). This makes the Rasch Model superior to classical test theory (CTT), which tends not to account for item and participant variability independently (L. Wang et al., 2022).

The application of the Rasch Model can be seen in the construct map in output Table 1.0 in the Winsteps application. In the Figure 1, individual abilities are identified on the left side (in logit units), while item difficulty levels are on the right side.

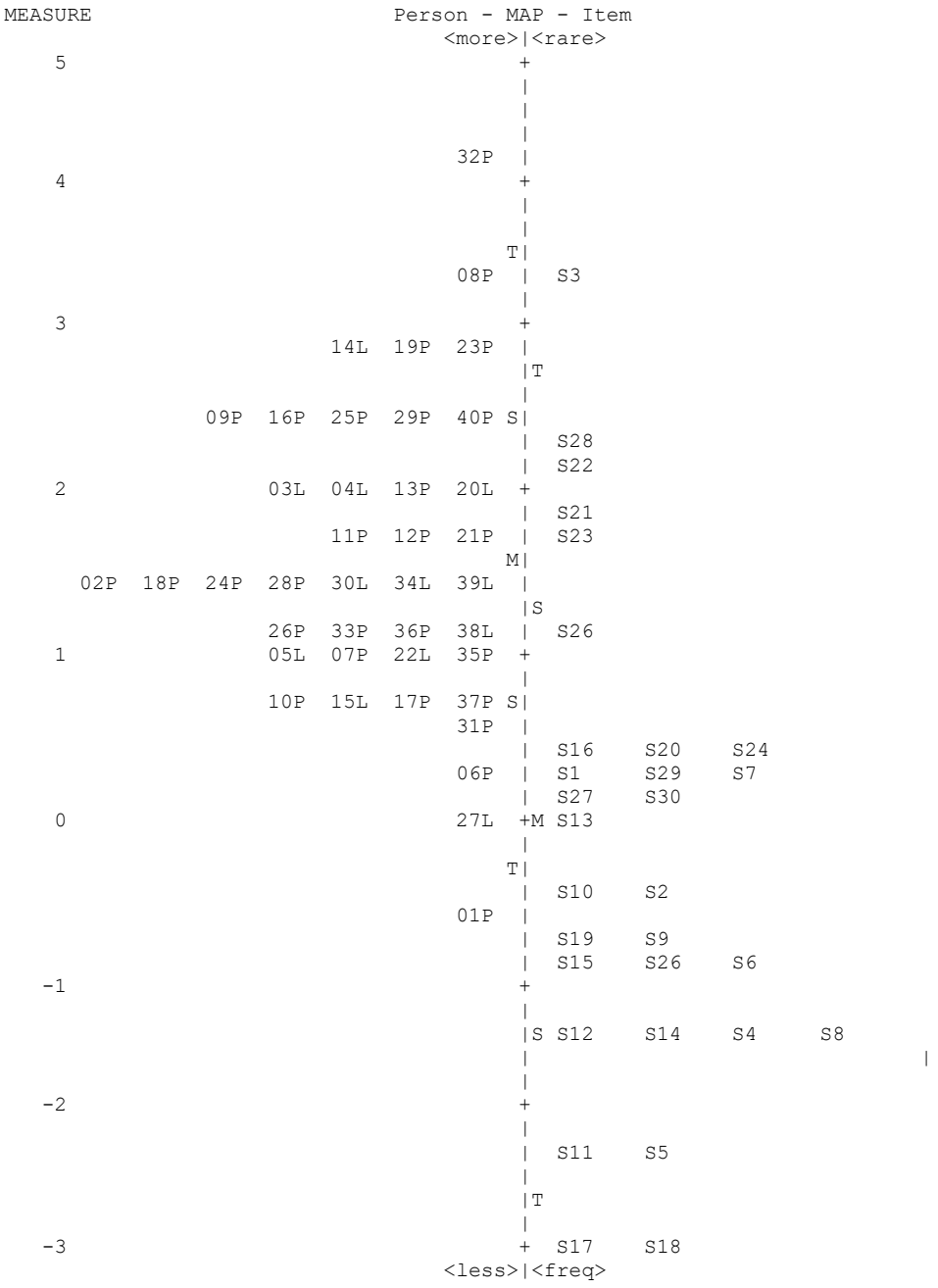

Figure 1. Wright's Map of the Rasch Model Individualization Ability

Based on the Wright Map in Figure 1, the distribution of test taker ability and item difficulty can be seen clearly. The map depicts test takers' ability on the logit scale on the left side and item difficulty on the right side. As can be seen, student 32P has the highest ability, which is positioned at the top of the logit scale. This student can cope with all items, including the most difficult item, S3, which is at the top of the item difficulty scale. This indicates that student 32P's ability is higher than the difficulty of the items available.

In contrast, student 01P showed the lowest ability among all participants, positioned at the bottom of the logit scale. This student was unable to complete even the easiest items, such as S17 and S18, which are at the bottom of the item difficulty scale. Thus, this student performed below the difficulty level of the items in the test.

The use of the Wright Map in this analysis is very useful for identifying gaps or differences in ability between test takers and item difficulty. For example, the large gap between students 32P and 01P

indicates a wide range of abilities among participants, which demands adjustments in item difficulty to accommodate the entire spectrum of test takers' abilities (Vaccarino et al., 2023). The Rasch model facilitates more accurate and focused measurement by mapping individual ability more precisely to item difficulty (Bradley & Massof, 2017; Jacob et al., 2019).

Thus, test takers' abilities can be mapped more granularly and accurately using the Wright Map. This allows educators and researchers to assess how precisely the items are aligned with the ability of the individual being tested, as well as providing insights for further interventions that may be needed at different ability levels (Le et al., 2022).

An advantage of the Rasch model over CTT is its ability to accurately measure individual abilities and analyze the alignment of abilities with response patterns among test participants and items using a scalogram. Vertically, from top to bottom, the scalogram indicates high to low abilities, while horizontally, from left to right, it shows items from easy to difficult. The scalogram can be observed in Figure 2.

```
GUTTMAN SCALOGRAM OF RESPONSES:
Person |Item
       |11 1  11 12 1 1123  212222222
       |78514824655992037O179604631283
       |------------------------------
    32 +11111111111111111111111111101  32P
     8 +11111111111111111111111111010  08P
    14 +11111111111111111111101110110  14L
    19 +11111111111111111111110101110  19P
    23 +11111111111111011111111111001  23P
     9 +11111111111111111111011110110  09P
    16 +11111111111111111111101111110001  16P
    25 +11111111111111101111111111011001  25P
    29 +11111111111111111111111111110000  29P
    40 +11111111111111111111111111110000  40P
     3 +11111111111111111110111110100001  03L
     4 +11111111111111111110111011111000  04L
    13 +11111111110111111111111010100011  13P
    20 +11111111111111111111111110001010  20L
    11 +11111111111111111111111100111110000  11P
    12 +11111111111101110111011111010100  12P
    21 +11111111111111111110111110011000  21P
     2 +11111101111111111111011101110000  02P
    18 +11111111001111111111111110001010  18P
    24 +11111110011111001111101111001  24P
    28 +11111111111101111111111001100100  28P
    30 +11111111110011110111111111100100  30L
    34 +11111111110011110111111111101000  34L
    39 +11110111111111111111111010000110  39L
    26 +11111111111111010101001110111010  26P
    33 +11111111111111111111011101010000  33P
    36 +11111111111111111110001110101000  36P
    38 +11110111111011010011111111100110  38L
     5 +11111111101111111101001110000110  05L
     7 +11111111111011111111010011000100  07P
    22 +11111111111111010111100111000000  22L
    35 +11111001111110111100111011111000  35P
    10 +11111111011111111111101000000100  10P
    15 +11111111111111001011001101001000  15L
    17 +11111011011101111111001100110000  17P
    37 +11010111111111111011000101100110  37P
    31 +11101010100111101111111111000100  31P
     6 +11111111111111100000100110010000  06P
    27 +11111111111110010001000010000010  27L
     1 +11111100101000100001000100101O  01P
       |------------------------------
       |11 1  11 12 1 1123  212222222
       |78514824655992037O179604631283
```

Figure 2. Analysis of Abilities Using a Scalogram with Rasch Model

In the analysis of individual ability using the Rasch model, one of its main advantages is the ability to convert raw scores into linear measures of ability. This allows for a clearer understanding of student ability relative to the difficulty level of a given item (Tesio et al., 2024). Based on the analysis results of the Guttman Scalogram, several things stand out regarding the discrepancy in students' response patterns, which may indicate unintentional guesses or errors.

For example, student 39L appeared to be inattentive because he was unable to answer the easiest item (item 4) but managed to answer several more difficult items such as item 7 and item 29. This pattern indicates a mismatch that may occur due to a lack of understanding of the concept or guessing behavior (Ibrahim et al., 2015). According to Tesio et al. (2024), this can be identified as part of the Rasch model's ability to detect inconsistent patterns and signal potential guessing.

In addition, students 29P and 40P showed identical response patterns, suggesting the possibility of collusion. As described by (Bramley, 2015), the Rasch model allows educators to identify non-genuine or cheating response patterns, especially if there are striking similarities between two test takers. This analysis increases the validity of the assessment, as it can separate genuine ability from potential cheating or random guessing (Darmana et al., 2021).

In the case of student 37P, there was a pattern where the student failed to answer an easy question (item 5) but was able to answer a difficult question (item 28), which exceeded his logit ability. A similar pattern was seen for student 31P, where the easy item (item 11) was not answered correctly, but the difficult item (item 22) was answered correctly. This may indicate luck or guessing, as explained by Goh et al. (2017), that the Rasch model is able to detect responses that indicate guessing behavior and distinguish between genuine ability and unintentional guessing (Pretz et al., 2016).

Further research has also shown that Rasch analysis can effectively identify response patterns that do not match true ability, which can help in distinguishing genuine concept understanding from responses generated by chance (Seamon et al., 2019). In cases like this, Rasch analysis helps in maintaining the accuracy of the assessment, ensuring that students' scores reflect their genuine ability and are not the result of random responses (Andrich et al., 2016).

Conclusion

This study demonstrates that both Classical Test Theory (CTT) and the Rasch model provide valuable insights into the characterization of biology test items, though they offer different perspectives. Specifically, CTT highlighted that while 14 items were valid, 16 were deemed invalid, and the reliability (Cronbach's Alpha) was at a sufficient level of 0.619. In contrast, the Rasch model produced a more detailed analysis, indicating a higher overall reliability of 0.85, with person reliability at 0.65, suggesting areas for improvement in the consistency of student responses. In terms of item difficulty, CTT categorized most items as easy (20 items), while the Rasch model provided a broader classification, identifying 6 very difficult and 8 difficult items, which underscores the model's capacity for more nuanced item calibration. Item discrimination revealed that CTT showed more variability in categorization, with most items rated as having low to moderate discrimination. The Rasch model, however, consolidated discrimination into fewer categories but with clearer item separation indices (H = 3.45), suggesting greater precision in distinguishing item difficulty across the respondent population. Regarding distractor effectiveness, CTT indicated that 93.3% of distractors were effective, while the Rasch model showed that 80% were effective, highlighting a potential gap in distractor performance that could benefit from further refinement using Rasch analysis.

CTT offers a more general evaluation of test reliability and item difficulty, but its analysis of item discrimination and distractor effectiveness may lack the depth provided by the Rasch model. The Rasch model provides a more refined understanding of item difficulty and respondent ability, detecting finer distinctions in both item and person parameters. The Rasch model identifies patterns of guessing and potential item bias that were not as evident in the CTT analysis, making it a more robust tool for identifying test anomalies and improving test quality.

This research is limited by the small sample size of 40 students from one class, which may affect the generalizability of the findings. Future studies should consider larger and more diverse sample sizes to strengthen the conclusions and applicability of the results. Future research should explore larger samples to verify the generalizability of these findings across different contexts. Additionally, combining CTT and Rasch model analyses can provide a more comprehensive evaluation of test instruments. Educators should consider utilizing the Rasch model for deeper insights into test validity and reliability, especially in identifying inconsistencies in item performance and student responses. Further exploration of item bias detection in high-stakes assessments is also recommended.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author Contributions

T. Priyani: conceptualization, methodology, design, analysis, writing original draft preparation, review, and editing. **B. Sugiharto:** conceptualization, methodology, analysis, evaluation, writing original draft

preparation, review, and editing.

References

Aaij, R., Abdelmotteleb, A. S. W., Beteta, C. A., Gallego, F. J. A., Ackernley, T., Adeva, B., Adinolfi, M., Afsharnia, H., Agapopoulou, C., Aidala, C. A., Aiola, S., Ajaltouni, Z., Akar, S., Albrecht, J., Alessio, F., Alexander, M., Albero, A. A., Aliouche, Z., Alkhazov, G., … Zunica, G. (2022). Study of the doubly charmed tetraquark Tcc+. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-30206-w

Andrich, D., Marais, I., & Humphry, S. M. (2016). Controlling Guessing Bias in the Dichotomous Rasch Model Applied to a Large-Scale, Vertically Scaled Testing Program. *Educational and Psychological Measurement*, *76*(3), 412–435. https://doi.org/10.1177/0013164415594202

Angell, D. K., Lane-Getaz, S., Okonek, T., & Smith, S. (2024). Metacognitive Exam Preparation Assignments in an Introductory Biology Course Improve Exam Scores for Lower ACT Students Compared with Assignments that Focus on Terms. *CBE Life Sciences Education*, *23*(1). https://doi.org/10.1187/cbe.22-10-0212

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Ayres, P., Lee, J. Y., Paas, F., & van Merriënboer, J. J. G. (2021). The Validity of Physiological Measures to Identify Differences in Intrinsic Cognitive Load. In *Frontiers in Psychology* (Vol. 12). Frontiers Media S.A. https://doi.org/10.3389/fpsyg.2021.702538

Babu, N., & Kohli, P. (2023). Commentary: Reliability in research. In *Indian Journal of Ophthalmology* (Vol. 71, Issue 2, pp. 400–401). Wolters Kluwer Medknow Publications. https://doi.org/10.4103/ijo.IJO_2016_22

Baghaei, P., Yanagida, T., & Heene, M. (2017). Development of a Descriptive Fit Statistic for the Rasch Model. In *North American Journal of Psychology* (Vol. 19, Issue 1).

Baldan, D., Negash, M., & Ouyang, J. Q. (2021). Are individuals consistent? Endocrine reaction norms under different ecological challenges. *Journal of Experimental Biology*, *224*(12). https://doi.org/10.1242/jeb.240499

Barbic, D., Kim, B., Salehmohamed, Q., Kemplin, K., Carpenter, C. R., & Barbic, S. P. (2018). Diagnostic accuracy of the Ottawa 3DY and Short Blessed Test to detect cognitive dysfunction in geriatric patients presenting to the emergency department. *BMJ Open*, *8*(3). https://doi.org/10.1136/bmjopen-2017-019652

Batista, S. A., Stedefeldt, E., Nakano, E. Y., De Oliveira Cortes, M., Assunção Botelho, R. B., Zandonadi, R. P., Raposo, A., Han, H., & Ginani, V. C. (2021). Design and development of an instrument on knowledge of food safety, practices, and risk perception addressed to children and adolescents from low-income families. *Sustainability (Switzerland)*, *13*(4), 1–20. https://doi.org/10.3390/su13042324

Bejerholm, U., & Lundgren-Nilsson, Å. (2015). Rasch Analysis of the Profiles of Occupational Engagement in people with Severe mental illness (POES) instrument. *Health and Quality of Life Outcomes*, *13*(1). https://doi.org/10.1186/s12955-015-0327-0

Blacquiere, L. D., & Hoese, W. J. (2016). A valid assessment of students' skill in determining relationships on evolutionary trees. *Evolution: Education and Outreach*, *9*(1). https://doi.org/10.1186/s12052-016-0056-9

Blanco, I., Boemo, T., Martin-Garcia, O., Koster, E. H. W., De Raedt, R., & Sanchez-Lopez, A. (2023a). Online Contingent Attention Training (OCAT): transfer effects to cognitive biases, rumination, and anxiety symptoms from two proof-of-principle studies. *Cognitive Research: Principles and Implications*, *8*(1). https://doi.org/10.1186/s41235-023-00480-3

Bradley, C., & Massof, R. W. (2017). Validating Translations of Rating Scale Questionnaires Using Rasch Analysis. In *Ophthalmic Epidemiology* (Vol. 24, Issue 1, pp. 1–2). Taylor and Francis Ltd. https://doi.org/10.1080/09286586.2016.1246667

Bramley, T. (2015). Rasch Measurement in the Social Sciences and Quality of Life Research. *Europe's Journal of Psychology*, *11*(1), 169–171. https://doi.org/10.5964/ejop.v11i1.913

Cecilio-Fernandes, D., Medema, H., Collares, C. F., Schuwirth, L., Cohen-Schotanus, J., & Tio, R. A. (2017). Comparison of formula and number-right scoring in undergraduate medical training: A Rasch model analysis. *BMC Medical Education*, *17*(1). https://doi.org/10.1186/s12909-017-1051-8

Cheema, J. R. (2019). Cross-country gender DIF in PISA science literacy items. *European Journal of Developmental Psychology*, *16*(2), 152–166. https://doi.org/10.1080/17405629.2017.1358607

Chen, P.-Y., Wu, W., Garnier-Villarreal, M., Kite, B. A., & Jia, F. (2019). *Testing Measurement Invariance with Ordinal Missing Data: A Testing Measurement Invariance with Ordinal Missing*

Data: A Comparison of Estimators and Missing Data Techniques Comparison of Estimators and Missing Data Techniques. https://epublications.marquette.edu/nursing_fac

Cliff, W. H. (2023). Teaching with core concepts to facilitate the integrated learning of introductory organismal biology. *Advances in Physiology Education*, *47*(3), 562–572. https://doi.org/10.1152/ADVAN.00134.2022

Darmana, A., Sutiani, A., Nasution, H. A., Ismanisa*, I., & Nurhaswinda, N. (2021). Analysis of Rasch Model for the Validation of Chemistry National Exam Instruments. *Jurnal Pendidikan Sains Indonesia*, *9*(3), 329–345. https://doi.org/10.24815/jpsi.v9i3.19618

de Jong, L. H., Bok, H. G. J., Schellekens, L. H., Kremer, W. D. J., Jonker, F. H., & van der Vleuten, C. P. M. (2022). Shaping the right conditions in programmatic assessment: how quality of narrative information affects the quality of high-stakes decision-making. *BMC Medical Education*, *22*(1). https://doi.org/10.1186/s12909-022-03257-2

De Sá, A. R., Liebel, G., De Andrade, A. G., Andrade, L. H., Gorenstein, C., & Wang, Y. P. (2019a). Can gender and age impact on response pattern of depressive symptoms among college students? A differential item functioning analysis. *Frontiers in Psychiatry*, *10*(FEB). https://doi.org/10.3389/fpsyt.2019.00050

Echevarría-Guanilo, M. E., Gonçalves, N., & Juceli Romanoski, P. (2019). Psychometric properties of measurement instruments: Conceptual basis and evaluation methods- Part II. *Texto e Contexto Enfermagem*, *28*. https://doi.org/10.1590/1980-265X-TCE-2017-0311

Eden, M. M. (2018). *Shoulder-Specific Patient Reported Outcome Measures for Use in Patients with Head and Neck Cancer:An Assessment of Reliability, Construct Validity, and Overall Appropriateness of Test Score Interpretation Using Rasch Analysis. Doctoral dissertation. Nova Southeastern University. Retrieved from NSUWorks, College of Health Care Sciences-Physical Therapy Department* (Issue 62). https://nsuworks.nova.edu/hpd_pt_stuetd

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, *31*(7), 792–806. https://doi.org/10.1177/0956797620916786

Finch, H., & Edwards, J. M. (2016). Rasch Model Parameter Estimation in the Presence of a Nonnormal Latent Trait Using a Nonparametric Bayesian Approach. *Educational and Psychological Measurement*, *76*(4), 662–684. https://doi.org/10.1177/0013164415608418

Fischer, H. F., & Rose, M. (2016). Www.common-metrics.org: A web application to estimate scores from different patient-reported outcome measures on a common scale. *BMC Medical Research Methodology*, *16*(1). https://doi.org/10.1186/s12874-016-0241-0

Fujimoto, Y., Chevance, M., Haydon, D. T., Krumholz, M. R., & Kruijssen, J. M. D. (2019). A fundamental test for stellar feedback recipes in galaxy simulations. *Monthly Notices of the Royal Astronomical Society*, *487*(2), 1717–1728. https://doi.org/10.1093/mnras/stz641

Gaitán-Rossi, P., Vilar-Compte, M., Teruel, G., & Pérez-Escamilla, R. (2021). Food insecurity measurement and prevalence estimates during the COVID-19 pandemic in a repeated cross-sectional survey in Mexico. *Public Health Nutrition*, *24*(3), 412–421. https://doi.org/10.1017/S1368980020004000

Gao, Y., Bing, L., Li, P., King, I., & Lyu, M. R. (2019). *Generating Distractors for Reading Comprehension Questions from Real Examinations*. www.aaai.org

Garrido, C. C., González, D. N., Seva, U. L., & Piera, P. J. F. (2019). Multidimensional or essentially unidimensional? A multi-faceted factoranalytic approach for assessing the dimensionality of tests and items. *Psicothema*, *31*(4), 450–457. https://doi.org/10.7334/psicothema2019.153

Gay, C. L., Kottorp, A., Lerdal, A., & Lee, K. A. (2016). Psychometric limitations of the center for epidemiologic studies-depression scale for assessing depressive symptoms among adults with HIV/AIDS: A rasch analysis. *Depression Research and Treatment*, *2016*. https://doi.org/10.1155/2016/2824595

Goel, A., & Gross, A. (2019). Differential item functioning in the cognitive screener used in the Longitudinal Aging Study in India. *International Psychogeriatrics*, *31*(9), 1331–1341. https://doi.org/10.1017/S1041610218001746

Goh, H. E., Marais, I., & Ireland, M. J. (2017). A Rasch Model Analysis of the Mindful Attention Awareness Scale. *Assessment*, *24*(3), 387–398. https://doi.org/10.1177/1073191115607043

Gray, N., Calleja, D., Wimbush, A., Miralles-Dolz, E., Gray, A., De Angelis, M., Derrer-Merk, E., Oparaji, B. U., Stepanov, V., Clearkin, L., & Ferson, S. (2020). Is “no test is better than a bad test”? Impact of diagnostic uncertainty in mass testing on the spread of COVID-19. *PLoS ONE*, *15*(10 October). https://doi.org/10.1371/journal.pone.0240775

Hagquist, C. (2019). Explaining differential item functioning focusing on the crucial role of external information - an example from the measurement of adolescent mental health. *BMC Medical Research Methodology*, *19*(1), 185. https://doi.org/10.1186/s12874-019-0828-3

Hope, D., Adamson, K., McManus, I. C., Chis, L., & Elder, A. (2018). Using differential item functioning to evaluate potential bias in a high stakes postgraduate knowledge based assessment. *BMC*

Medical Education, *18*(1). https://doi.org/10.1186/s12909-018-1143-0
Ibrahim, F. M., Shariff, A. A., & Tahir, R. M. (2015). Using Rasch model to analyze the ability of pre-university students in vector. *AIP Conference Proceedings*, *1682*. https://doi.org/10.1063/1.4932472
Jacob, E. R., Duffield, C., & Jacob, A. M. (2019). Validation of data using RASCH analysis in a tool measuring changes in critical thinking in nursing students. *Nurse Education Today*, *76*, 196–199. https://doi.org/10.1016/j.nedt.2019.02.012
Jacobs, N. W., Berduszek, R. J., Dijkstra, P. U., & van der Sluis, C. K. (2017). Validity and Reliability of the Upper Extremity Work Demands Scale. *Journal of Occupational Rehabilitation*, *27*(4), 520–529. https://doi.org/10.1007/s10926-016-9683-9
Jimam, N. S., Ahmad, S., & Ismail, N. E. (2019). Psychometric Classical Theory Test and Item Response Theory Validation of Patients' Knowledge, Attitudes and Practices of Uncomplicated Malaria Instrument. *Journal of Young Pharmacists*, *11*(2), 186–191. https://doi.org/10.5530/jyp.2019.11.39
Jin, I. H., & Jeon, M. (2019). A Doubly Latent Space Joint Model for Local Item and Person Dependence in the Analysis of Item Response Data. *Psychometrika*, *84*(1), 236–260. https://doi.org/10.1007/s11336-018-9630-0
Jones, I., Bisson, M., Gilmore, C., & Inglis, M. (2019). Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help? *British Educational Research Journal*, *45*(3), 662–680. https://doi.org/10.1002/berj.3519
Kassim, M. A. M., Pang, N. T. P., Kamu, A., Arslan, G., Mohamed, N. H., Zainudin, S. P., Ayu, F., & Ho, C. M. (2023). Psychometric Properties of the Coronavirus Stress Measure with Malaysian Young Adults: Association with Psychological Inflexibility and Psychological Distress. *International Journal of Mental Health and Addiction*, *21*(2), 819–835. https://doi.org/10.1007/s11469-021-00622-y
Köhler, C., & Hartig, J. (2017). Practical Significance of Item Misfit in Educational Assessments. *Applied Psychological Measurement*, *41*(5), 388–400. https://doi.org/10.1177/0146621617692978
Kok, K., & Priemer, B. (2023). Assessment tool to understand how students justify their decisions in data comparison problems. *Physical Review Physics Education Research*, *19*(2). https://doi.org/10.1103/PhysRevPhysEducRes.19.020141
Korbee, S., VAN KEMPEN, R., VAN WENSEN, R., VAN DER STEEN, M., & Liu, W. Y. (2022). Measurement properties of the HOOS-PS in revision total hip arthroplasty: a validation study on validity, interpretability, and responsiveness in 136 revision hip arthroplasty patients. *Acta Orthopaedica*, *93*, 742–749. https://doi.org/10.2340/17453674.2022.4572
Królikowska, A., Reichert, P., Karlsson, J., Mouton, C., Becker, R., & Prill, R. (2023). Improving the reliability of measurements in orthopaedics and sports medicine. *Knee Surgery, Sports Traumatology, Arthroscopy*, *31*(12), 5277–5285. https://doi.org/10.1007/s00167-023-07635-1
Kumar Mohajan, H. (2017). *TWO CRITERIA FOR GOOD MEASUREMENTS IN RESEARCH: VALIDITY AND RELIABILITY*.
Le, C., Guttersrud, Ø., Sørensen, K., & Finbråten, H. S. (2022). Developing the HLS19-YP12 for measuring health literacy in young people: a latent trait analysis using Rasch modelling and confirmatory factor analysis. *BMC Health Services Research*, *22*(1). https://doi.org/10.1186/s12913-022-08831-4
Lestari, E. K., & Yudhanegara, R. M. (2015). *Penelitian Pendidikan Matematika*. Refika Aditama.
Lewis, A. F., Myers, M., Heiser, J., Kolar, M., Baird, J. F., & Stewart, J. C. (2020). Test–retest reliability and minimal detectable change of corticospinal tract integrity in chronic stroke. *Human Brain Mapping*, *41*(9), 2514–2526. https://doi.org/10.1002/hbm.24961
Li, J. J., Reise, S. P., Chronis-Tuscano, A., Mikami, A. Y., & Lee, S. S. (2016). Item Response Theory Analysis of ADHD Symptoms in Children With and Without ADHD. *Assessment*, *23*(6), 655–671. https://doi.org/10.1177/1073191115591595
Liu, R., & Jiang, Z. (2020). A general diagnostic classification model for rating scales. *Behavior Research Methods*, *52*(1), 422–439. https://doi.org/10.3758/s13428-019-01239-9
Liu, Y., Yin, H., Xin, T., Shao, L., & Yuan, L. (2019). A comparison of differential item functioning detection methods in cognitive diagnostic models. *Frontiers in Psychology*, *10*(MAY). https://doi.org/10.3389/fpsyg.2019.01137
Liuzza, M. T., Spagnuolo, R., Antonucci, G., Grembiale, R. D., Cosco, C., Iaquinta, F. S., Funari, V., Dastoli, S., Nistico, S., & Doldo, P. (2021). Psychometric evaluation of an Italian custom 4-item short form of the PROMIS anxiety item bank in immune-mediated inflammatory diseases: An item response theory analysis. *PeerJ*, *9*. https://doi.org/10.7717/peerj.12100
Mazurek, M. O., Carlson, C., Baker-Ericzén, M., Butter, E., Norris, M., & Kanne, S. (2020). Construct Validity of the Autism Impact Measure (AIM). *Journal of Autism and Developmental Disorders*, *50*(7), 2307–2319. https://doi.org/10.1007/s10803-018-3462-8
McKeigue, P. (2019). Quantifying performance of a diagnostic test as the expected information for

discrimination: Relation to the C-statistic. *Statistical Methods in Medical Research*, *28*(6), 1841–1851. https://doi.org/10.1177/0962280218776989

Milania, A. A., & Murniati, W. (2022). Teacher's Pedagogic Competence In Evaluating Learning. *KINDERGARTEN: Journal of Islamic Early Childhood Education*, *5*(2), 245. https://doi.org/10.24014/kjiece.v5i2.20013

Morgan-López, A. A., Saavedra, L. M., Hien, D. A., Killeen, T. K., Back, S. E., Ruglass, L. M., Fitzpatrick, S., López-Castro, T., & Patock-Peckham, J. A. (2020). Estimation of equable scale scores and treatment outcomes from patient-and clinician-reported PTSD measures using item response theory calibration. *Psychological Assessment*, *32*(4), 321–335. https://doi.org/10.1037/pas0000789

Murphy, M., McCloughen, A., & Curtis, K. (2019). Using theories of behaviour change to transition multidisciplinary trauma team training from the training environment to clinical practice. *Implementation Science*, *14*(1). https://doi.org/10.1186/s13012-019-0890-6

Musa, A., Shaheen, S., Elmardi, A., & Ahmed, A. (2021). Item difficulty & item discrimination as quality indicators of physiology MCQ examinations at the Faculty of Medicine Khartoum University. *Khartoum Medical Journal*, *11*(2). https://doi.org/10.53332/kmj.v11i2.610

Nielsen, J. B., Kyvsgaard, J. N., Sildorf, S. M., Kreiner, S., & Svensson, J. (2017). Item analysis using Rasch models confirms that the Danish versions of the DISABKIDS® chronic-generic and diabetes-specific modules are valid and reliable. *Health and Quality of Life Outcomes*, *15*(1). https://doi.org/10.1186/s12955-017-0618-8

O'Brien, K. K., Dzingina, M., Harding, R., Gao, W., Namisango, E., Avery, L., & Davis, A. M. (2021). Developing a short-form version of the HIV Disability Questionnaire (SF-HDQ) for use in clinical practice: a Rasch analysis. *Health and Quality of Life Outcomes*, *19*(1). https://doi.org/10.1186/s12955-020-01643-2

Orozco, T., Segal, E., Hinkamp, C., Olaoye, O., Shell, P., & Shukla, A. M. (2022). Development and validation of an end stage kidney disease awareness survey: Item difficulty and discrimination indices. *PLoS ONE*, *17*(9 September). https://doi.org/10.1371/journal.pone.0269488

Papenberg, M., & Musch, J. (2017). Of Small Beauties and Large Beasts: The Quality of Distractors on Multiple-Choice Tests Is More Important Than Their Quantity. *Applied Measurement in Education*, *30*(4), 273–286. https://doi.org/10.1080/08957347.2017.1353987

Parra-Anguita, L., Sánchez-García, I., Del Pino-Casado, R., & Pancorbo-Hidalgo, P. L. (2019). Measuring knowledge of Alzheimer's: Development and psychometric testing of the UJA Alzheimer's Care Scale. *BMC Geriatrics*, *19*(1). https://doi.org/10.1186/s12877-019-1086-2

Poorebrahim, A., Lin, C. Y., Imani, V., Kolvani, S. S., Alaviyoun, S. A., Ehsani, N., & Pakpour, A. H. (2021). Using Mindful Attention Awareness Scale on male prisoners: Confirmatory factor analysis and Rasch models. *PLoS ONE*, *16*(7 July). https://doi.org/10.1371/journal.pone.0254333

Prenovost, K. M., Fihn, S. D., Maciejewski, M. L., Nelson, K., Vijan, S., & Rosland, A. M. (2018). Using item response theory with health system data to identify latent groups of patients with multiple health conditions. *PLoS ONE*, *13*(11). https://doi.org/10.1371/journal.pone.0206915

Pretz, C. R., Kean, J., Heinemann, A. W., Kozlowski, A. J., Bode, R. K., & Gebhardt, E. (2016). A Multidimensional Rasch Analysis of the Functional Independence Measure Based on the National Institute on Disability, Independent Living, and Rehabilitation Research Traumatic Brain Injury Model Systems National Database. *Journal of Neurotrauma*, *33*(14), 1358–1362. https://doi.org/10.1089/neu.2015.4138

Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Nuha Medika.

Retnawati, H. (2016). *Analisis Kuantitatif Instrumen Penelitian* (Pertama). Parama Publishing. www.nuhamedika.gu.ma

Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaifi, J., ALGhamdi, M. A., Al Ameer, A. Y., Yahia, A. I. O., Mohammed, O. A., & Adam, M. I. E. (2024). Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*, *24*(1). https://doi.org/10.1186/s12909-024-05433-y

Robinson, M., Johnson, A. M., Walton, D. M., & MacDermid, J. C. (2019). A comparison of the polytomous Rasch analysis output of RUMM2030 and R (ltm/eRm/TAM/lordif). *BMC Medical Research Methodology*, *19*(1). https://doi.org/10.1186/s12874-019-0680-5

Rodrigo, M. F., Molina, J. G., Losilla, J. M., Vives, J., & Tomás, J. M. (2019). Method effects associated with negatively and positively worded items on the 12-item General Health Questionnaire (GHQ-12): Results from a cross-sectional survey with a representative sample of Catalonian workers. *BMJ Open*, *9*(11). https://doi.org/10.1136/bmjopen-2019-031859

Rogowska, A. M., Ochnik, D., & Kuśnierz, C. (2022). Revisiting the multidimensional interaction model of stress, anxiety and coping during the COVID-19 pandemic: a longitudinal study. *BMC Psychology*, *10*(1). https://doi.org/10.1186/s40359-022-00950-1

Ronk, F. R., Hooke, G. R., & Page, A. C. (2016). Validity of clinically significant change classifications

yielded by Jacobson-Truax and Hageman-Arrindell methods. *BMC Psychiatry*, *16*(1). https://doi.org/10.1186/s12888-016-0895-5

Runge, J. M., Lang, J. W. B., Chasiotis, A., & Hofer, J. (2019). Improving the Assessment of Implicit Motives Using IRT: Cultural Differences and Differential Item Functioning. *Journal of Personality Assessment*, *101*(4), 414–424. https://doi.org/10.1080/00223891.2017.1418748

Saat, N. A. (2020). *Sains Humanika Humanika Summative Test Items Analysis Using Classical Test Theory (CTT) Analisis Item Kertas Peperiksaan Sumatif Menggunakan Teori Ujian Klasik (TUK)*. www.sainshumanika.utm.my

Seamon, B. A., Kautz, S. A., & Velozo, C. A. (2019). Rasch Analysis of the Activities-Specific Balance Confidence Scale in Individuals Poststroke. *Archives of Rehabilitation Research and Clinical Translation*, *1*(3–4). https://doi.org/10.1016/j.arrct.2019.100028

Seide, S. E., Röver, C., & Friede, T. (2019). Likelihood-based random-effects meta-analysis with few studies: Empirical and simulation studies. *BMC Medical Research Methodology*, *19*(1). https://doi.org/10.1186/s12874-018-0618-3

Sen, S., Cohen, A. S., & Kim, S. H. (2016). The Impact of Non-Normality on Extraction of Spurious Latent Classes in Mixture IRT Models. *Applied Psychological Measurement*, *40*(2), 98–113. https://doi.org/10.1177/0146621615605080

Stanley, L. M., & Edwards, M. C. (2016). Reliability and Model Fit. *Educational and Psychological Measurement*, *76*(6), 976–985. https://doi.org/10.1177/0013164416638900

Steiner, M. D., & Frey, R. (2021). Representative Design in Psychological Assessment: A Case Study Using the Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: General*, *150*(10), 2117–2136. https://doi.org/10.1037/xge0001036

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

Sumantri, S. M., & Retni Satriani. (2016). The Effect of Formative Testing and Self-Directed Learning on Mathematics Learning Outcomes. In *International Electronic Journal of Elementary Education* (Vol. 8, Issue 3). www.simdik.info/hasilun/index.aspx.

Sumintono, B. (2016). *Seminar Nasional Pendidikan IPA Prosiding Seminar Nasional Pendidikan IPA "Mengembangkan Keterampilan Berpikir Tingkat Tinggi Melalui Pembelajaran IPA" Penerbit: S2 IPA UNLAM PRESS PENILAIAN KETERAMPILAN BERPIKIR TINGKAT TINGGI: APLIKASI PEMODELAN RASCH PADA ASESMEN PENDIDIKAN*.

Sumintono, & Widhiarso, W. (2015). *Aplikasi Permodelan Rasch Pada Assessment Pendidikan* (B. Trim, Ed.; Cetakan I). Trim Komunikata. www.trimkomunikata.com

Teresi, J. A., Wang, C., Kleinman, M., Jones, R. N., & Weiss, D. J. (2021). Differential Item Functioning Analyses of the Patient-Reported Outcomes Measurement Information System (PROMIS®) Measures: Methods, Challenges, Advances, and Future Directions. *Psychometrika*, *86*(3), 674–711. https://doi.org/10.1007/s11336-021-09775-0

Tesio, L., Caronni, A., Kumbhare, D., & Scarano, S. (2024). Interpreting results from Rasch analysis 1. The "most likely" measures coming from the model. *Disability and Rehabilitation*, *46*(3), 591–603. https://doi.org/10.1080/09638288.2023.2169771

Trampush, J. W., Yang, M. L. Z., Yu, J., Knowles, E., Davies, G., Liewald, D. C., Starr, J. M., Djurovic, S., Melle, I., Sundet, K., Christoforou, A., Reinvang, I., Derosse, P., Lundervold, A. J., Steen, V. M., Espeseth, T., Räikkönen, K., Widen, E., Palotie, A., … Lencz, T. (2017). GWAS meta-analysis reveals novel loci and genetic correlates for general cognitive function: A report from the COGENT consortium. *Molecular Psychiatry*, *22*(3), 336–345. https://doi.org/10.1038/mp.2016.244

Tzafilkou, K., Perifanou, M., & Economides, A. A. (2022). Development and validation of students' digital competence scale (SDiCoS). *International Journal of Educational Technology in Higher Education*, *19*(1). https://doi.org/10.1186/s41239-022-00330-0

Vaccarino, A. L., Black, S. E., Gilbert Evans, S., Frey, B. N., Javadi, M., Kennedy, S. H., Lam, B., Lam, R. W., Lasalandra, B., Martens, E., Masellis, M., Milev, R., Mitchell, S., Munoz, D. P., Sparks, A., Swartz, R. H., Tan, B., Uher, R., & Evans, K. R. (2023). Rasch analyses of the Quick Inventory of Depressive Symptomatology Self-Report in neurodegenerative and major depressive disorders. *Frontiers in Psychiatry*, *14*. https://doi.org/10.3389/fpsyt.2023.1154519

Van Vliet, M., Doornenbal, B. M., Boerema, S., & Van Den Akker-Van Marle, E. M. (2021). Development and psychometric evaluation of a Positive Health measurement scale: A factor analysis study based on a Dutch population. *BMJ Open*, *11*(2). https://doi.org/10.1136/bmjopen-2020-040816

Van Zile-Tamsen, C. (2017). Using Rasch Analysis to Inform Rating Scale Development. *Research in Higher Education*, *58*(8), 922–933. https://doi.org/10.1007/s11162-017-9448-0

Wang, F., Liu, Q., Chen, E., Huang, Z., Chen, Y., Yin, Y., Huang, Z., & Wang, S. (2020). *Neural Cognitive Diagnosis for Intelligent Education Systems*. www.aaai.org

Wang, L., Wu, Y. X., Lin, Y. Q., Wang, L., Zeng, Z. N., Xie, X. L., Chen, Q. Y., & Wei, S. C. (2022).

Reliability and validity of the Pittsburgh Sleep Quality Index among frontline COVID-19 health care workers using classical test theory and item response theory. *Journal of Clinical Sleep Medicine*, *18*(2), 541–551. https://doi.org/10.5664/jcsm.9658

Wilberforce, M., Sköldunger, A., & Edvardsson, D. (2019). A Rasch analysis of the Person-Centred Climate Questionnaire - Staff version. *BMC Health Services Research*, *19*(1). https://doi.org/10.1186/s12913-019-4803-9

Wu, X., Zhang, L. J., & Liu, Q. (2021). Using Assessment for Learning: Multi-Case Studies of Three Chinese University English as a Foreign Language (EFL) Teachers Engaging Students in Learning and Assessment. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.725132

Wyse, A. E., & Mapuranga, R. (2009). Differential Item Functioning Analysis Using Rasch Item Information Functions. *International Journal of Testing*, *9*(4), 333–357. https://doi.org/10.1080/15305050903352040

Ye, S., Sun, K., Huynh, D., Phi, H. Q., Ko, B., Huang, B., & Ghomi, R. H. (2022). A Computerized Cognitive Test Battery for Detection of Dementia and Mild Cognitive Impairment: Instrument Validation Study. *JMIR Aging*, *5*(2). https://doi.org/10.2196/36825

Yow, W. Q., & Priyashri, S. (2019). Computerized Electronic Features Direct Children's Attention to Print in Single-and Dual-Language e-Books. *AERA Open*, *5*(3). https://doi.org/10.1177/2332858419878126

Zlatkin-Troitschanskaia, O., Pant, H. A., Toepper, M., Lautenbach, C., & Molerov, D. (2017). Valid Competency Assessment in Higher Education: Framework, Results, and Further Perspectives of the German Research Program KoKoHs. *AERA Open*, *3*(1). https://doi.org/10.1177/2332858416686739