Asian Journal of Assessment in Teaching and Learning
Vol 16, Issue 1, 2026 (12-27) /eISSN 2821-2916

Assessing Educational Instrument Validity-Reliability and Student Performance: Rasch Model Insights

Mohd Zaidi Amiruddin*
The student of Doctorate Program of Science Education of Universitas Pendidikan Indonesia, Bandung, Indonesia

*Corresponding author: mohdzaidi@upi.edu

Received: 12 February 2026; Revised: 1 May 2026;
Accepted: 20 May 2026; Published: 3 June 2026

To link to this article: https://doi.org/10.37134/ajatel.vol16.1.2.2026

Abstract

This study aims to assess the validity-reliability and student performance. A cross-sectional research design with a quantitative method was employed in this study. The participants of this research were 34 students divided into 12 males (35.3%) and 22 females (64.7%) with an age range of 16-17 years. The data collection used a Quizizz with 15 questions online but still under teacher supervision. The results demonstrate moderate reliability and validity per Rasch analysis, with a person reliability score of 0.63 and an item reliability score of 0.46, indicating a need for item refinement. The person separation index (1.31) and item separation index (0.93) show moderate differentiation, while a Cronbach alpha of 0.70 and raw variance explained by measures (41.6%) support instrument validity. The characteristics of the items and persons reveal that most items align well with the model expectations, although some exhibit variability or inconsistencies that need addressing.

Keywords: *Assessment, Instruments, Performance, Rasch Analysis, Reliability, Validity*

INTRODUCTION

High-quality instruments in education have always become a crucial issue, which drives the need for accurate and reliable measuring instruments to assess learning achievement. Good measuring tools ensure that assessments are carried out fairly and valid and can appropriately measure students' abilities and knowledge. This is crucial to provide helpful feedback to students and educators and develop more effective learning strategies. According to Bennett, (2011); Wiliam et al. (2004) a good assessment instrument helps accurately measure student achievement and abilities, allows educators to know how much students understand the subject matter, and identifies areas that need improvement. In addition, appropriate instruments also contribute to improving the quality of learning by identifying gaps in students' understanding and adjusting teaching methods (Pianta et al., 2012; Stigler & Hiebert, 2009). Data generated from assessment instruments supports better decision-making in curriculum development, program evaluation, and education policy (Dunn et al., 2013; Hora et al., 2017; Shen et al., 2012).

On the other hand, well-designed instruments help develop students' competencies and skills by measuring certain aspects of learning and providing a clear picture of their progress. Using consistent and reliable instruments allows continuous evaluation of student progress for long-term planning and educational development (Jonsson & Svingby, 2007; Spooren et al., 2013). High-quality instruments are the foundation of an effective and efficient education system, which assesses student learning outcomes and supports the continuous development of the learning process (Care et al., 2018; Chalmers, 2007). In the long run, using good instruments can improve the overall quality of education and ensure that all students have an equal opportunity to succeed. Therefore, investment in developing and

implementing quality assessment instruments must be a priority in improving education quality (Martínez-Caro et al., 2015).

In the present study, we identified whether the items used in measuring learning achievement were appropriate. We selected the Rasch model because it offers advantages over classical test theory for identifying students' ability and item difficulty. Cavanagh and Waugh (2011); Törmäkangas (2011) mentioned some of the advantages of Rasch modeling: (1) generates linear unidimensional scales; (2) demands that the data conform to the measurement model; (3) creates person measures that are independent of the scale; (4) yields item difficulties that are independent of the sample; (5) computes standard errors; (6) determines person measures and item difficulties on the same linear scale using standard units (logits); and (7) verifies the logical and consistent use of the scoring system. These characteristics identify student abilities by considering item and person parameters.

1. Validity-Reliability Instruments Development

Developing high-quality instruments is vital in scientific research because it guarantees the reliability and validity of the collected data. This is essential for reaching accurate conclusions and making well-informed decisions. Reliability pertains to the consistency of measurement, whereas validity concerns the degree to which the instrument accurately measures what it is designed to measure (Cook & Beckman, 2006; Sürücü & Maslakci, 2020). The process of developing and validating a research instrument is a multifaceted endeavor that involves reducing errors in the measurement process (Floyd & Widaman, 1995). Kimberlin and Winterstein (2008), reliability estimates, which assess factors like the stability of measures, internal consistency, and interrater reliability, are crucial for ensuring the instrument's consistency and dependability. On the other hand, validity focuses on the degree to which the interpretations of a test's results are justified, which is contingent on the specific purpose the test is designed to serve (Kimberlin & Winterstein, 2008). Ensuring the reliability and validity of measurement instruments is crucial in qualitative and quantitative research, as these qualities are essential for the accuracy and trustworthiness of the data collected (Behi & Nolan, 1995; Golafshani, 2003). The development of quality instruments involves carefully considering factors that can lead to measurement errors, the assessment of reliability in quantitative terms, and the interpretation of reliability coefficients.

Reliability consists of three types to ensure the consistency of measures: test-retest reliability, internal consistency, and inter-rater reliability (Ahmed et al., 2022; Park et al., 2018). Test-retest reliability assesses whether a measure yields similar results over time, which is crucial for constructs like intelligence and self-esteem, with a Pearson correlation of +.80 or higher indicating good reliability (De Castella & Byrne, 2015). For instance, intelligence tends to remain stable over time. A very intelligent person today will likely be just as intelligent next week. Therefore, a reliable measure of intelligence should yield similar scores for this individual next week as it does today. Internal consistency examines the correlation among items within a measure, ensuring they reflect the same underlying construct, with Cronbach's α of +.80 or greater being desirable (Vaske et al., 2017; Viladrich et al., 2017). For instance, individuals might place a series of bets in a simulated roulette game to gauge their risk-seeking behavior. This measure would be internally consistent if participants consistently made high or low bets across different trials. Similar to test-retest reliability, internal consistency can only be evaluated by gathering and analyzing data. Inter-rater reliability evaluates the consistency of judgments made by different observers, using Cronbach's α for quantitative data or Cohen's κ for categorical data to determine high correlation and reliability in observations. For instance, if you wanted to assess university students' social skills, you could record videos of them interacting with another student they are meeting for the first time. Then, two or more observers could watch the videos and rate each student's social skill level.

Validity refers to how well a measure represents the intended variable, evaluated through various forms of evidence beyond reliability (Fink & Litwin, 1995; Jonsson & Svingby, 2007). Face validity assesses if a measure appears to measure the construct, though it relies on intuition and can be misleading (Zumbo, 2006). Validity is at best, tentative evidence that a measurement method is accurately measuring what it is supposed to (Borsboom et al., 2004; L. Cohen et al., 2017). One reason is that it is based on people’s intuitions about human behavior, which are frequently wrong. Content validity checks if the measure comprehensively covers the construct, ensuring all relevant aspects are

included (Lewis et al., 2005). For instance, if a researcher defines test anxiety as comprising both sympathetic nervous system activation (causing nervous feelings) and negative thoughts, then their measure of test anxiety should include items about both nervous feelings and negative thoughts. Similarly, attitudes generally encompass thoughts, feelings, and actions toward something. Criterion validity examines correlations with related variables, including concurrent, predictive, and convergent validity, where the latter ensures new measures correlate with established ones (Van Iddekinge et al., 2012). Discriminant validity ensures that the measure does not correlate with distinct, unrelated variables, confirming that it measures the intended construct rather than something else (Henseler et al., 2015). For instance, self-esteem is a stable, overall attitude toward oneself that persists over time, unlike mood, which reflects how good or bad one feels at a particular moment. Therefore, scores on a new self-esteem measure should not be strongly correlated with an individual's mood.

In educational research and evaluation, tests and questionnaires are the most commonly used methods for data collection. These tools gather information about knowledge, attitudes, opinions, behavior, facts, and other relevant details (Bradburn et al., 2004; Feldman & Lynch, 1988). Developing valid and reliable questionnaires or tests is a must to reduce measurement error. Groves (1987) defines measurement error as "a discrepancy between respondents' attributes and their survey answers". Developing instruments in the form of valid and reliable questionnaires or tests involves several steps that require quite a long time. The adaptation stages for developing and testing the instrument (Radhakrishna, 2007) are presented in Figure 1.

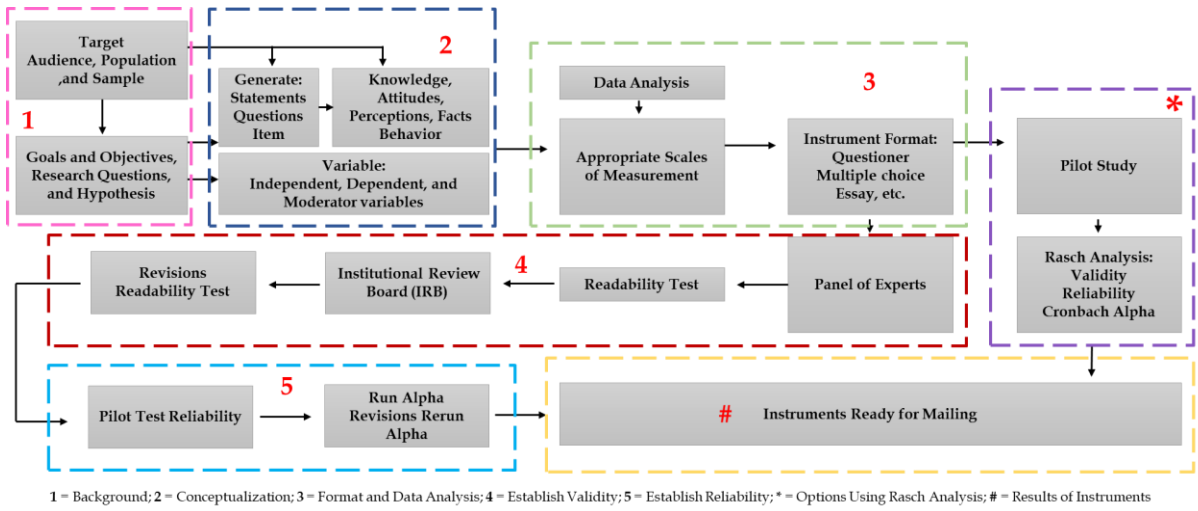

Figure 1 Sequence of instruments developments

2. Assessing Students Performance

Assessing the development of student performance is a critical component in understanding the efficacy of educational approaches and identifying areas for improvement. Researchers have explored a range of instructional strategies and their influence on student achievement, highlighting the importance of creating an effective learning environment, leveraging prior knowledge, and facilitating the application and extension of concepts (Hitt & Tucker, 2016; Kyriakides et al., 2013). One key factor in student performance is the assessment strategies employed by teachers. Assessment informs teachers about individual student knowledge and abilities and guides decisions on content, methods, and skill development for individual students and the class (Brookhart, 2011; Stiggins, 2010; Suskie, 2018). Teachers' approaches and strategies can significantly impact student achievement, underscoring the need to identify effective techniques.

Effective assessment strategies include both formative and summative assessments, each with specific roles in the learning process. Formative assessments—like quizzes, peer reviews, and interactive class activities—offer immediate feedback to both students and teachers, enabling prompt interventions and adjustments to teaching methods (Gikandi et al., 2011; Stiggins, 2010). This ongoing

assessment helps identify learning gaps, misunderstandings, and areas requiring further reinforcement, promoting a more personalized learning experience. Conversely, Summative assessments, such as final exams, standardized tests, and end-of-term projects, assess student learning at the end of an instructional period. They provide a thorough overview of student achievement and evaluate the effectiveness of the educational strategies used (Al-Hawamdeh et al., 2023; Kulasegaram & Rangachari, 2018). Adaptation from the National Forum for the Enhancement of Teaching and Learning in Higher Education in Ireland (2017) provides a valuable perspective on assessment for, of, and as learning. It highlights the interaction between different types of assessments and the important roles of both the assessments and the individuals involved, as illustrated in Figure 2.

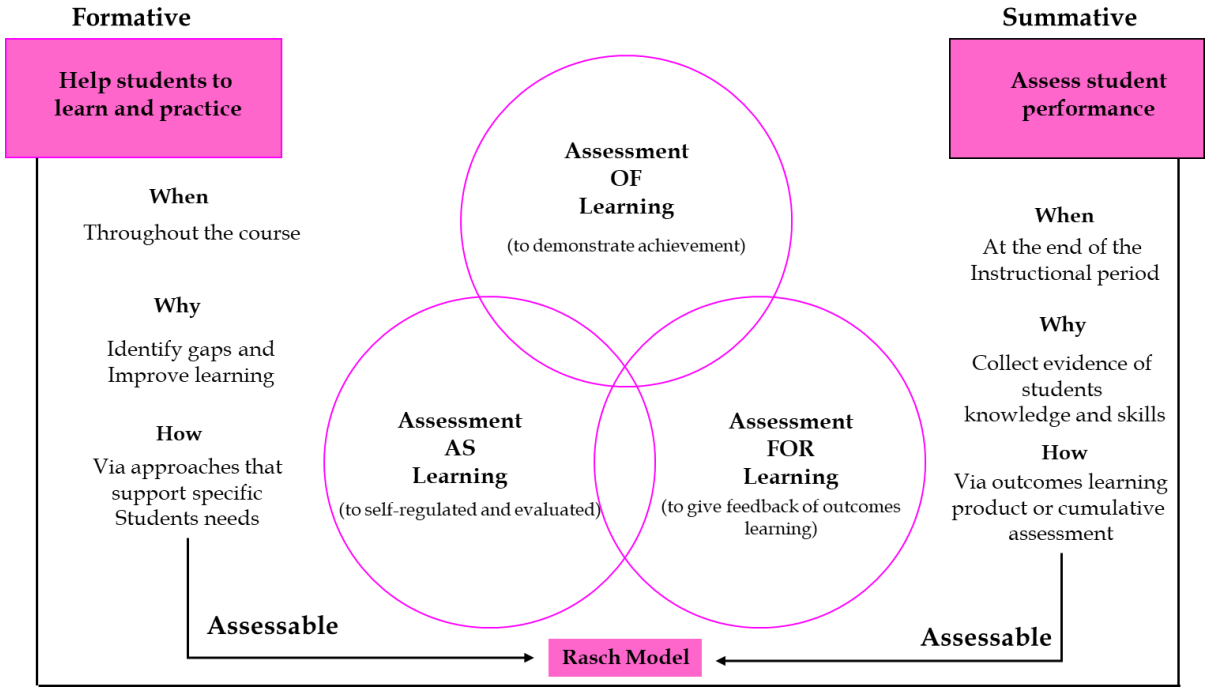

Figure 2 Assessment OF/FOR/AS learning

Technology integration is another critical aspect of student performance assessment (Pellegrino & Quellmalz, 2010). Digital tools and platforms can enhance the assessment process by providing more dynamic, interactive, and accessible means for evaluating student progress. Tools such as online quizzes, learning management systems, and educational software enable teachers to track student performance in real-time, analyze data more efficiently, and tailor instruction to meet individual student needs (Bienkowski et al., 2012; D. Cohen & Sasson, 2016; Galy et al., 2011). Additionally, technology can facilitate more diverse and inclusive assessment methods, accommodating different learning styles and providing opportunities for students to demonstrate their understanding in various formats (Akbulut & Cardak, 2012; Akkoyunlu & Soylu, 2008). As education evolves, leveraging technology in assessment strategies will be essential in fostering an adaptive and responsive learning environment.

3. The Research Questions

This study aimed to assess the quality of educational instruments and student performance. The data was analyzed by employing the Rasch measurement using WINSTEP software. The following questions were formulated as follows:

1. How does the Rasch model help evaluate the reliability and validity of educational assessment instruments?
2. What are the characteristics of an Item and Person?
3. How can the Rasch model be used to identify and address the person's abilities and item difficulty?

METHODS

This research used a cross-sectional research design with a quantitative method to present information on the quality of instruments and student performance. A cross-sectional research design with a quantitative descriptive was chosen because it can provide a reasonable interpretation, description, and appearance of results from the data obtained (Al-Sagarat et al., 2017; Colville et al., 2009). This approach can describe the frequency distribution, percentage, mean, and spread of data, such as standard deviation, to clearly understand the quality of educational instruments and student performance. In addition, quantitative descriptive analysis allows the identification of patterns and trends present in the data, making it easier to present findings that are informative and easily understood by readers (Creswell & Creswell, 2017; Johnson & Christensen, 2019). The results of these analyses can also provide a solid basis for further research or decision-making in an educational context.

1. Research Participants

This research involved 34 students, 12 males (35.3%) and 22 females (64.7%), aged 16-17. The participants were first-year students at a senior high school in East Java Province, Indonesia. Students were selected to provide a representative picture of the existing population to determine the quality of the instrument and simultaneously determine student performance using Rasch analysis. The participants were given 5-15 minutes to complete the test monitored by the teacher during regular class time.

2. Data Collection

The data was collected by giving a multiple-choice test consisting of 15 questions on physics measurements. The work on these questions is carried out with the help of Quizizz online but remains under the teacher's supervision in the classroom (e.g., Arifin & Setiawan, 2022; Lim & Yunus, 2021). This test is given to determine students' understanding and mastery after studying the topic of physical measurements. The online quiz results can be interpreted directly into Microsoft Excel for further analysis. The displays in Microsoft Excel, namely, overview, participant data, time, and quiz details, are presented in Figure 3.

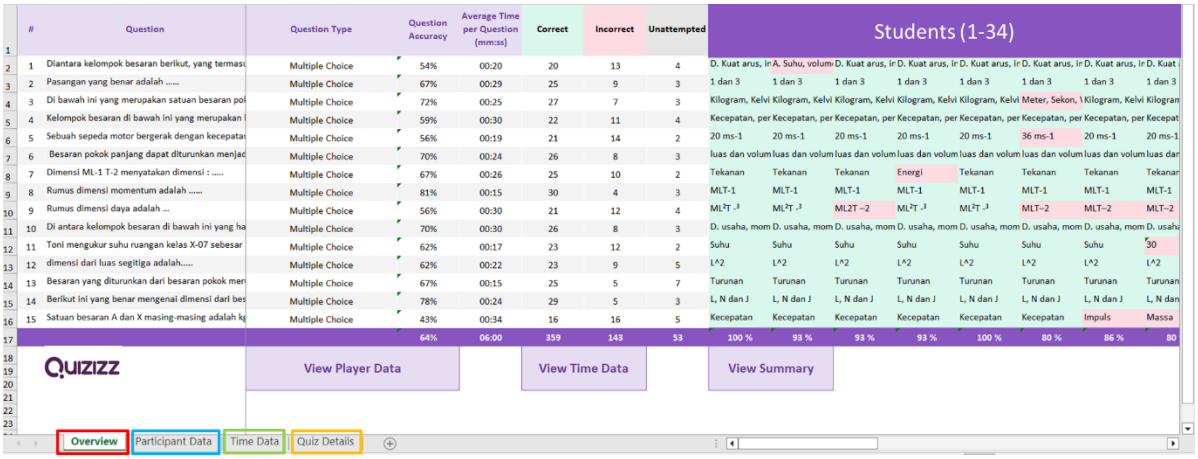

Figure 3 Excel overview Quizizz

3. Data Analysis

The data integrated from Quizizz into Microsoft Excel was then coded to apply the dichotomous scoring (e.g., correct = 1 and incorrect = 0). Then, Rasch analysis was conducted using WINSTEP software version 5.1.4 (Linacre, 2021). The Rasch measurement model estimates a person's ability (parameter a) and item difficulty (parameter b) on a single scale. Rasch analysis involved using Rasch modeling with

joint maximum likelihood estimation (JMLE), which transformed student scores into the logit scale (interval data) ranging from negative to positive infinity. Rasch parameter evaluation was used to assess validity and reliability by examining unidimensionality, and local independence, and by evaluating person and item reliability criteria. The Wright map was provided to verify the alignment between a person's abilities and item difficulty levels, confirming the targeting criteria. Furthermore, item and person fit were assessed using outfit statistics. Item fit was evaluated through Outfit Mean Squares (MNSQ) and ZSTD values. The acceptable range for Outfit MNSQ is between 0.5 and 1.5, while ZSTD values should fall between -2.0 and +2.0 to indicate a reasonable fit. Additionally, Pt. Mean Corr. values ranged from 0.4 to 0.85 (Boone et al., 2014).

RESULTS AND DISCUSSION

1. The Validity and Reliability

The results of the Rasch analysis indicate that the measurement instrument has moderate reliability and reasonably fits the Rasch model (see Table 1). The person reliability is 0.63, suggesting moderate consistency in measuring individual abilities, while the item reliability of 0.46 is relatively low, indicating a need for item refinement. The mean person measure of 0.94 suggests that the test is slightly challenging for the average respondent. The model standard errors for the person (0.68) and item (0.42) measures reflect the precision of these estimates, with person measures showing room for improvement. The fit statistics (MNSQ values close to 1 and ZSTD values near zero) indicate a good fit between the data and the Rasch model (Abdellatif, 2023). However, the person separation index of 1.31 suggests only moderate differentiation between different levels of person abilities, and the item separation index of 0.93 indicates limited ability to distinguish between different item difficulties. The Cronbach alpha of 0.70 signifies acceptable test score reliability, though there is potential for improvement. Additionally, the raw variance explained by the measures is 41.6% more than 40 %, indicating a significant proportion of variance accounted for by the Rasch model.

Table 1 Summary of parameters for instruments

	Count		**Reliability**		**Measure**		**Model S.E.**		**Output MNSQ**		**Output ZSTD**		**Separation**	
	Person	Item	Person	Item	Person	Item	Person	Item	Person	Item	Person	Item	Person	Item
Mean	15.0	34.0	**0.63**	**0.46**	0.94	0.00	**0.68**	**0.42**	0.95	0.95	0.12	-0.05	1.31	0.93
SEM	0.0	0.0	0.0	0.0	0.21	0.16	0.03	0.01	0.04	0.09	0.12	0.27	0.0	0.0
P.SD	0.0	0.0	0.0	0.0	1.18	0.60	0.18	0.04	0.22	0.33	0.70	0.99	0.0	0.0
S.SD	0.0	0.0	0.0	0.0	1.20	0.62	0.18	0.04	0.23	0.34	0.71	1.03	0.0	0.0
MAX.	15.0	34.0	0.63	046.	2.78	1.26	1.05	0.54	1.48	1.72	2.57	1.97	1.31	0.93
MIN.	15.0	34.0	0.63	0.46	-2.79	-1.21	0.54	0.39	0.49	0.56	-1.22	-1.32	1.31	0.93
Cronbach Alpha (KR-20) Person Raw Score "Test" Reliability = **0.70**; SEM = 1.65 Raw variance explained by measures = **41.6%**														

2. Unidimensionality

The unidimensionality assumption was examined, and the "Scree Plot," shown in Figure 4, was analyzed. This analysis helps determine the number of factors to retain in a factor analysis.

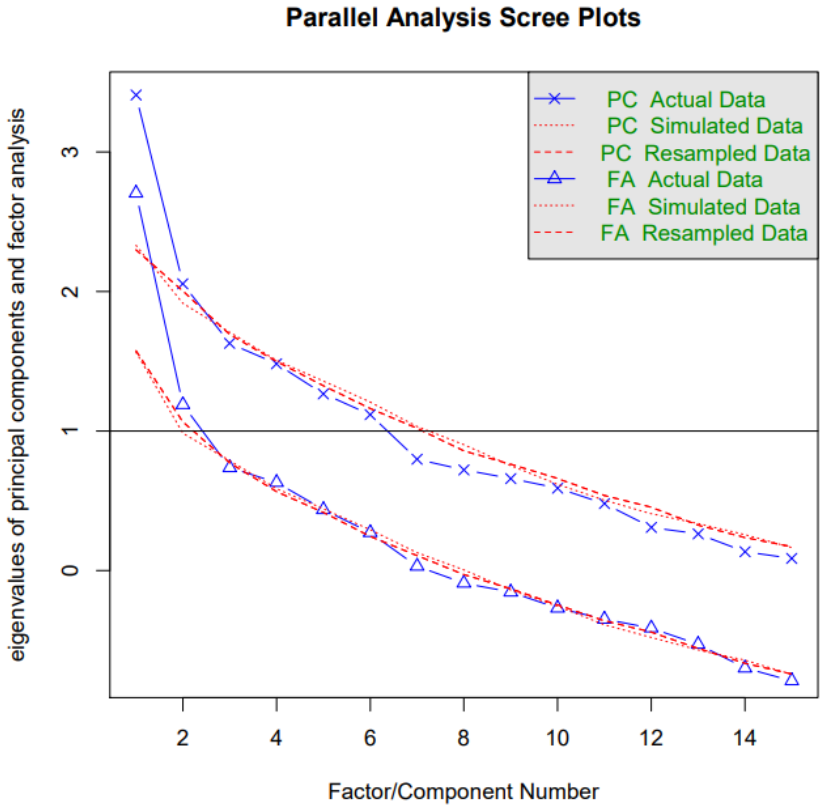

Figure 4 Scree lot

Figure 4 provides a scree plot from a parallel analysis that compares the eigenvalues of actual data with those from simulated and resampled data for both principal components (PC) and factor analysis (FA). The x-axis represents the factor/component number, while the y-axis shows the eigenvalues. The blue crosses and triangles represent the eigenvalues from the actual dataset using PC and FA, respectively. The red dotted and dashed lines show the eigenvalues from simulated and resampled data. The horizontal black line at an eigenvalue of 1 serves as a threshold for retaining components or factors. Components or factors with eigenvalues above this line are typically considered significant. To determine the number of factors/components to retain, we look at where the actual data's eigenvalues exceed those of the simulated and resampled data. For PC, the actual data eigenvalues (blue crosses) drop below the simulated and resampled data (red lines) after about the 2nd or 3rd component, suggesting that 2 to 3 components should be retained.

Similarly, for FA, the actual data eigenvalues (blue triangles) fall below the simulated and resampled data after around the 2nd or 3rd factor. This indicates that the data likely has 2 to 3 underlying dimensions or factors, making it reasonable to retain approximately 2 to 3 factors or components based on this analysis. As a result, it was determined that this assumption was met, and the structure was unidimensional.

3. Characteristics of item and person

Tables 2 and Table 3 shows the characteristics of the item and person. Table 2 shows the most challenging and effortless questions (item), and Table 3 shows the abilities of a student (person) regarding the JMLE measure score (logit).

The JMLE measurement results generally indicate acceptable fit indices for most items. The mean infit and outfit MNSQ values are close to 1.00, suggesting that most items align well with the model expectations. Items such as Q4 and Q5 show strong fit, with infit MNSQ values below 0.50 and outfit MNSQ values below 0.86, indicating consistent responses. Conversely, items like Q1 and Q15 exhibit higher infit and outfit MNSQ values, which might suggest these items have more variability or inconsistencies in responses compared to the model's expectations. The z-values further support this observation, with extreme values indicating potential fit issues in some items. Point measure

correlations reveal variability in how healthy items relate to the underlying construct, ranging from 0.19 to 0.62. Items such as Q1 and Q15 have lower correlations, suggesting weaker relationships with the construct, which may impact their effectiveness. On the other hand, items like Q4 and Q5, with higher correlations, show a stronger association with the measured construct. This shows that while most items fit well within the model, items with significant deviations or lower correlations may need further review or revision to enhance overall measurement quality.

Table 2 Item characteristics

Item	JMLE Measure	Model S.E.	Infit		Outfit		PT. Measure.	
			MNSQ	ZSTD	MNSQ	ZSTD	Corr.	Exp.
Q15	**1.26**	0.39	1.14	0.93	1.04	0.25	0.35	0.43
Q1	**0.52**	0.39	1.34	2.12	1.39	1.40	0.19	0.44
Q9	**0.52**	0.39	1.27	1.74	1.38	1.35	0.22	0.44
Q4	**0.37**	0.39	0.77	-1.57	0.66	-1.32	0.62	0.44
Q5	**0.37**	0.39	0.86	-0.88	0.75	-0.90	0.56	0.44
Q11	**0.21**	0.40	0.86	-0.86	0.72	-0.96	0.56	0.44
Q12	**0.21**	0.40	0.78	-1.37	0.64	-1.31	0.62	0.44
Q7	**0.5**	0.41	1.14	0.83	1.72	1.97	0.25	0.44
Q2	**-0.12**	0.42	1.06	0.38	1.30	0.91	0.36	0.43
Q10	**-0.12**	0.42	1.02	0.16	0.94	-0.06	0.42	0.43
Q13	**-0.12**	0.42	0.91	-0.44	0.78	-0.58	0.51	0.43
Q3	**-0.50**	0.45	1.02	0.15	0.88	-0.15	0.43	0.42
Q6	**-0.50**	0.45	0.94	-0.18	0.74	-0.52	0.49	0.42
Q14	**-0.94**	0.50	1.08	0.37	0.81	-0.18	0.39	0.41
Q8	**-1.21**	0.54	0.80	-0.47	0.56	-0.62	0.55	0.40
Mean	0.00	0.42	1.00	0.06	0.95	-0.05		
P.SD	0.60	0.04	0.17	1.02	0.33	0.99		

Table 3 Person Characteristics

Person	JMLE Measure	Model S.E.	Infit		Outfit		PT. Measure.	
			MNSQ	ZSTD	MNSQ	ZSTD	Corr.	Exp.
01PD	**2.78**	1.04	1.06	0.35	1.10	0.45	0.02	0.14
02PD	**2.78**	1.04	0.98	0.27	0.71	0.06	0.23	0.14
03PD	**2.78**	1.04	0.98	0.27	0.71	0.14	0.23	0.14
04PD	**2.78**	1.04	1.06	0.35	1.10	0.45	0.02	0.14
05PD	**2.78**	1.04	1.08	0.37	1.29	0.61	-0.05	0.14
06PD	**1.48**	0.66	1.07	0.31	1.07	0.31	0.11	0.22
07PD	**1.99**	0.77	0.79	-0.25	-.54	-0.58	0.58	0.19
08LD	**1.48**	0.66	0.81	-0.41	0.67	-.061	0.55	0.22
09PD	**1.48**	0.66	0.95	0.00	0.94	0.03	0.28	0.22
10LD	**1.48**	0.66	0.86	-0.27	0.77	-.036	-.46	0.22
11LD	**1.48**	0.66	0.86	-0.27	0.77	-.036	-.46	0.22
12PD	**1.99**	0.77	0.86	-0.09	0.72	-0.22	0.43	0.19
13PD	**1.48**	0.66	1.03	0.20	0.92	-0.01	0.21	0.22
14LD	**0.75**	0.57	0.91	-0.35	0.86	-0.46	0.41	0.26
15PD	**1.48**	0.66	0.86	-.26	0.75	-0.41	0.47	0.22
16PD	**1.09**	0.60	0.86	-0.43	0.85	-0.32	0.45	0.24
17LD	**0.44**	0.55	1.05	0.38	1.03	0.21	0.19	0.27
18LD	**0.75**	0.57	1.03	0.21	0.99	0.03	0.23	0.26
19PD	**0.75**	0.57	1.00	0.07	1.16	0.66	0.20	0.26
20PD	**1.09**	0.60	1.18	0.72	1.20	0.64	-0.06	0.24
21PD	**0.75**	0.57	1.10	0.52	1.15	0.60	0.09	0.26
22PD	**0.75**	0.57	1.08	0.43	1.04	0.25	0.15	0.26
23PD	**0.44**	0.55	1.02	0.18	0.99	0.01	0.25	0.27

continued

Person	JMLE Measure	Model S.E.	Infit		Outfit		PT. Measure.	
			MNSQ	ZSTD	MNSQ	ZSTD	Corr.	Exp.
24LD	**0.75**	0.57	0.86	-0.61	0.84	-0.53	0.48	0.26
25PD	**0.44**	0.55	1.13	0.79	1.16	0.79	0.06	0.27
26LD	**0.44**	0.55	0.83	-1.05	0.78	-1.07	0.66	0.27
27PD	**0.15**	0.54	1.37	2.45	1.48	2.57	-0.37	0.27
28LD	**-0.14**	0.54	0.97	-0.18	0.94	-0.28	0.34	0.28
29PD	**-0.14**	0.54	1.17	1.20	1.22	1.32	-0.02	0.28
30LD	**-0.14**	0.54	0.82	-1.27	0,80	-1.22	0.57	0.28
31PD	**-0.43**	0.55	1.03	0.23	1.00	0.09	0.24	0.27
32LD	**-0.14**	0.54	1.11	0.76	1.12	0.75	0.10	0.28
33LD	**-1.09**	0.6	1.22	0.80	1.28	0.83	-0.11	0.25
34PD	**-2.79**	1.05	0.90	0.16	0.49	-0.24	0.42	0.15
Mean	0.94	0.68	1.00	0.17	0.95	0.12		
P.SD	1.18	0.18	0.13	0.65	0.22	0.70		

The JMLE measurement results for persons show a diverse range of fit indices and point-measure correlations. Most individuals have infit, and outfit MNSQ values close to 1.00, indicating that their responses generally align well with the model's expectations. However, there are notable exceptions: for instance, persons like 01PD and 27PD show higher infit and outfit MNSQ values, suggesting they may exhibit more significant variability or inconsistency in their responses. The z-values and point measure correlations further highlight individual variability, with extreme values such as those for 27PD indicating significant deviations from the expected model fit. The average point measure correlation across individuals is 0.12, reflecting a moderate relationship between item responses and the underlying construct. Some individuals, like 27PD, exhibit a lack of correlations, indicating weaker associations with the measured construct, while others, such as 26LD and 25PD, show higher correlations and a firmer fit. The match of item and person characteristics can be reviewed through the Wright map of student ability and the corresponding question difficulty level between the two.

4. Person – Item Map

Through the ZSTD statistics outfit, a fit statistic is used to detect unusual patterns of responses. Values close to 0 indicate a good fit; values far from 0 may indicate a misfit (Maydeu-Olivares, 2017). Figure 5 depicts the alley construct related to how the person and item are distributed in ZSTD outfit categories per item and person. Instead, see the balance between the level of student ability and the difficulty level of the questions presented on the Wright map in Figure 6.

Figure 5 provides insights into the performance of both persons and items in a psychometric assessment. For persons, the measures range from high ability levels (2.78) to low ability levels (-2.79), with the standard error values indicating varying levels of precision. Most persons show outfit ZSTD values close to 0, indicating a good fit, although there are a few outliers, like person 27PD with a high outfit ZSTD of 2.57, indicating a potential misfit. Higher ability measures tend to have higher standard errors, indicating less precision in these measures. For items, the difficulty levels vary from 0.52 (high difficulty) to -1.09 (low difficulty), with most items showing a good fit based on their outfit ZSTD values. However, some items, such as Q7 with an outfit ZSTD of 1.97, indicate a potential misfit. The standard error values for items are relatively consistent, ranging from 0.39 to 0.54, suggesting a reasonable level of precision in the item measures. Overall, this analysis highlights areas where both person and item measures could be refined for improved precision and better fit in the assessment tool. The results showed that the instrument quality regarding validity is valid and reliable in line with rash parameters. Rash reliability analysis can be measured by referring to person and item, where item reliability (0.46) is included in the low category and person reliability (0.63) is included in the moderate category. To see how the relationship between two countries can be traced based on the Cronbach alpha value. The Cronbach alpha value between item-person (0.70) is acceptable. However, it is necessary to pay more attention to item-person reliability for future improvements. Planinic et al. (2019); Wright (1977), the Rasch model can measure reliability based on items and people through the results of the

answers given. Then, the quality of the validity of the instruments used can be seen based on the raw variance explained measure, which is 41.6% greater than the minimum limit of 20%. Cook and Beckman (2006); Kimberlin and Winterstein (2008), in the instrument development process, it is crucial to ensure the validity and reliability of the instrument so that errors do not occur in the measurement process. That way, the measurement results measure what should be measured.

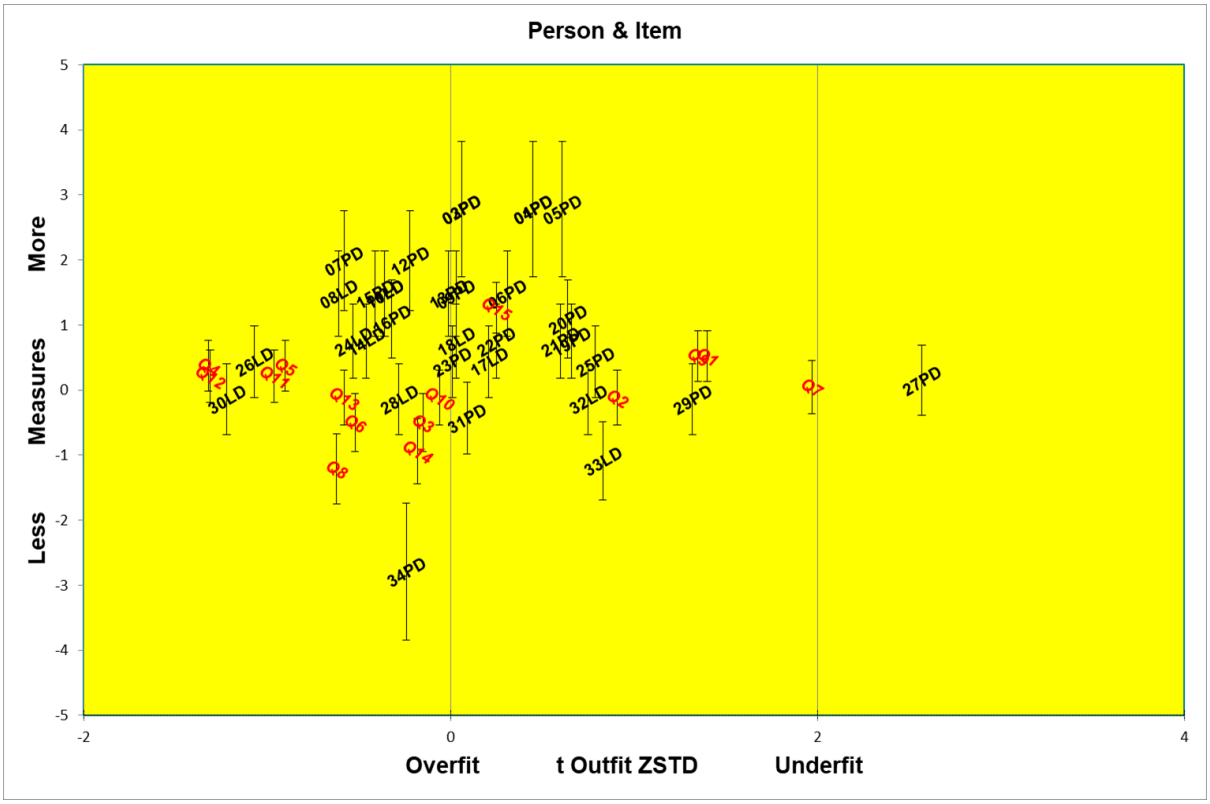

Figure 5 Construct alley

Furthermore, examining the unidimensionality assumption using a scree plot analysis, as presented in Figure 4, provides valuable insights into the dimensionality of the data. A scree plot represents the number of factors or components to retain in a factor or principal component analysis (PCA) (Abdi & Williams, 2010; Jackson, 1993). In this analysis, the scree plot is derived from a parallel analysis that compares the eigenvalues of the actual data with those obtained from simulated and resampled data. In line with research from various statisticians and psychometricians (Garrido et al., 2013; Green et al., 2012), this parallel analysis method determines the number of significant factors in data. The plot shows that for principal components (PC), the eigenvalues from the actual data drop below those of the simulated and resampled data after about the 2nd or 3rd component, suggesting that retaining 2 to 3 components is appropriate. Despite this, the conclusion was made that the unidimensionality assumption is met. This implies that while multiple dimensions or factors are present, they collectively support a primary dimension or factor structure. Therefore, the data's structure can be considered unidimensional for practical purposes, meaning that the factors retained predominantly measure a single construct or latent trait (Bartholomew et al., 2011; Reise et al., 2000, 2010). This approach ensures that the primary construct of interest is adequately represented while acknowledging the underlying complexity of the data.

Continuing from the unidimensionality of the data construct through item and person characteristics by looking at the JMLE measure (logit) value. Amiruddin et al. (20230; Wright (1977), a high logit value indicates the level of difficulty of the question (item) and the level of student ability (person). Of the 15 questions given to students, seven questions had a minus value, and eight questions had a plus value from the standard logit value limit of 0. However, it cannot be concluded that the questions with a logit value (-) were not good because they had to look at the value. logit person, too.

Apart from that, looking at the MNSQ value, all items are in the expectation model. The Rasch model also presents expectation and observation values that can reference the model's fit for items and people (Neumann et al., 2011). Meanwhile, for person characteristics, only seven people have a logit value minus 34 people. Apart from the logit value, you can also see whether the MNSQ outfit is acceptable with a range of 0.5 to 1.5. These results agree with previous studies on characteristics based on JMLE measure values (e.g., Darman et al., 2024; Kreijns et al., 2020; Soeharto & Csapó, 2022). The evaluation of instruments and student performance can continue to be developed to achieve the ideal model between the two.

Apart from the MNSQ score, you can also look at the ZSTD score. Fit category ZSTD values fall between − 2.0 and + 2.0 (). The person or item can indicate a misfit when it exceeds the accepted range of ZSTD values. For instance, a person with code 17PD (see. Figure 5) exceeds the limit mark 2.0, which can indicate a misfit because it exceeds the range of existing values. Besides that, the item about Q7 also got an underfit limit with a value of 1.97. Previous studies have also used confederate constructs to confirm consistency with the intended measurement constructs (Massof, 2005, 2011; Massof & Rubin, 2001). This analysis highlights areas where both person and item measures could be refined for improved precision and better fit in the assessment tool.

Continue from previous results with the presentation Wright map to see the balance between difficulty questions (items) and the student's level of ability (person). It can be seen that the most challenging items are Q15 (1.26) and the easiest Q8 (-1.21). Although item Q15 is the most challenging item, 14 students were able to answer Q8, which is why there is still one student who needs help getting the most accessible item answered correctly, namely student 34PD. Several studies are consistent in measuring levels of difficulty on question items (e.g., Karlin & Karlin, 2018; Planinic et al., 2019; Tennant & Conaghan, 2007) and student abilities (e.g., Arnold et al., 2018; Chan et al., 2021; Talib et al., 2018), following the results of this research. Through this Wight map, we can see directly between the quality of the questions and student performance in a lesson, which is measured in this case of learning physics. That way, it can become a reference for teachers in evaluating instruments for students' quality and performance that align with previous studies (e.g., Brochado, 2009; Spooren et al., 2007; Zeng et al., 2023).

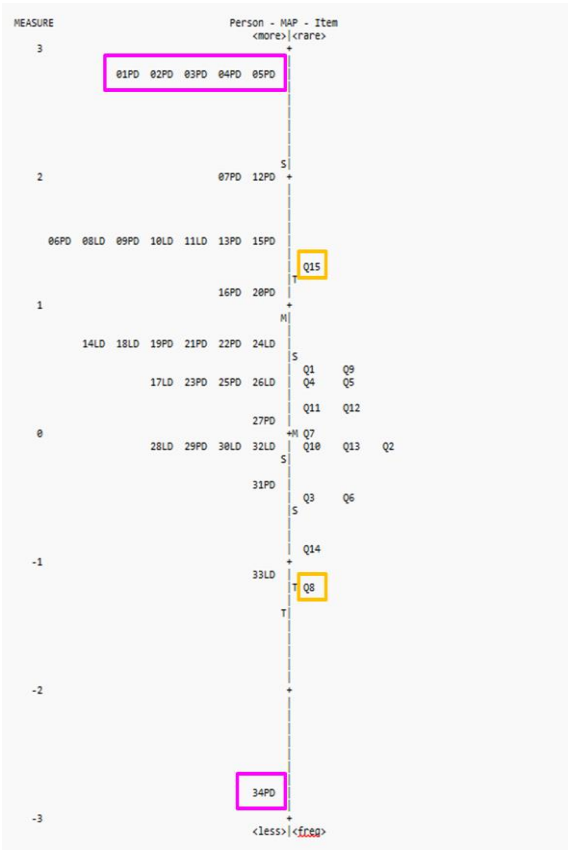

Figure 6 Wright mMap

Figure 6 depicts two categories: person (left) and item (right). The person section states the student's ability level from the lowest to the highest, while the items state the difficulty level of the questions from the easiest to the most difficult (Boone & Noltemeyer, 2017). This Wright map is presented based on the visible JMSE measure (logit) value (i.e., -3 to 3). Students with high abilities, namely 01PD, 02PD, 03PD, 04PD, and 05PD, have a logit value of 2.78. Meanwhile, the most difficult question item is Q15, with a logit value of 1.26. Then, the person with the lowest ability, namely 34PD, has a logit value of -2.97, while the easiest question is Q8, with a logit value of -1.21. Through the Wright Map, which measures the ability and difficulty level of questions based on logit values, people with higher logit values can work on questions with lower logit values. Instead, people with a logit value lower than the logit value of the question item are unable to answer the question.

CONCLUSION

According to Rasch's analysis, the measurement instrument used demonstrates moderate reliability and validity. The person reliability score of 0.63 suggests moderate consistency in measuring individual abilities, while the item reliability score of 0.46 highlights the need for item refinement. The person separation index of 1.31 and the item separation index of 0.93 reflect a moderate ability to distinguish between different levels of abilities and item difficulties, respectively. The Cronbach alpha of 0.70 indicates acceptable test score reliability, and the raw variance explained by the measures (41.6%) exceeds the minimum threshold, further supporting the instrument's validity. The unidimensionality assumption was met, suggesting that the data primarily measures a single construct despite multiple dimensions or factors. The characteristics of the items and persons reveal that most items align well with the model expectations, though some items and persons exhibit variability or inconsistencies that need addressing.

This study was limited to a small sample in one school in Indonesia. In addition, this study cannot be generalized to different contexts. Item refinement is necessary to improve the relatively low item reliability score, focusing on enhancing item quality to differentiate item difficulties better. Increasing the sample size could provide a more comprehensive understanding of the instrument's reliability and validity, and exploring additional dimensions through factor analysis might reveal more about the measured construct. Incorporating qualitative feedback from respondents can inform item revisions, while longitudinal studies could track changes over time, aiding in understanding the instrument's stability across different contexts. Comparative studies with other established tools can highlight relative strengths and weaknesses, providing a basis for further refinement and validation. Addressing these suggestions will enhance the reliability and validity of measurement instruments in educational assessments and beyond.

ACKNOWLEDGEMENT

The authors would like to thank the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia with DRTPM and “Program Pendidikan Magister Menuju Doktor untuk Sarjana Unggul (PMDSU) Batch VII and Universitas Pendidikan Indonesia

FUNDING

The authors declare that no financial support was received for the research, authorship, and publication of this article.

DATA AVAILABILITY STATEMENT

Data will be made available on request

CONFLICT OF INTEREST

The author declare no conflicts of interest" should be included if there is no conflict of interest.

REFERENCES

Abdellatif, H. (2023). Test results with and without blueprinting:Psychometric analysis using the Rasch model. *Educación Médica*, *24*(3), 100802. https://doi.org/10.1016/j.edumed.2023.100802

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459. https://doi.org/10.1002/wics.101

Ahmed, V., Olanipekun, A., Opoku, A., & Sutrisna, M. (2022). Understanding reliability in research. In *Validity and Reliability in Built Environment Research* (pp. 3–15). Routledge. https://doi.org/10.1201/9780429243226-2

Akbulut, Y., & Cardak, C. S. (2012). Adaptive educational hypermedia accommodating learning styles: A content analysis of publications from 2000 to 2011. *Computers & Education*, *58*(2), 835–842. https://doi.org/10.1016/j.compedu.2011.10.008

Akkoyunlu, B., & Soylu, M. Y. (2008). A study of student's perceptions in a blended learning environment based on different learning styles. *Journal of Educational Technology & Society*, *11*(1), 183–193. https://doi.org/10.1016/j.iheduc.2007.12.006

Al-Hawamdeh, B. O. S., Hussen, N., & Abdelrasheed, N. S. G. (2023). Portfolio vs. summative assessment: impacts on EFL learners' writing complexity, accuracy, and fluency (CAF); self-efficacy; learning anxiety; and autonomy. *Language Testing in Asia*, *13*(1), 12. https://doi.org/10.1186/s40468-023-00225-5

Al-Sagarat, A. Y., Yaghmour, G., & Moxham, L. (2017). Intentions and barriers toward breastfeeding among Jordanian mothers—A cross sectional descriptive study using quantitative method. *Women and Birth*, *30*(4), e152–e157. https://doi.org/10.1016/j.wombi.2016.11.001

Amiruddin, M. Z. Bin, Samsudin, A., Suhandi, A., Kaniawati, I., COŞTU, B., Aminuddin, A. H., & Kuniawan, F. (2023). Validity and Reliability of the Global Warming Instrument: A Pilot Study Using Rasch Model Analysis. *Jurnal Pendidikan MIPA*, *24*(4), 912–922. https://doi.org/10.23960/jpmipa/v24i4.pp912-922

Arifin, Z., & Setiawan, B. (2022). Utilising Gamification for Online Evaluation through Quizizz: Teachers' Perspectives and Experiences. *World Journal on Educational Technology: Current Issues*, *14*(3), 781–796. https://doi.org/10.18844/wjet.v14i3.7278

Arnold, J. C., Boone, W. J., Kremer, K., & Mayer, J. (2018). Assessment of competencies in scientific inquiry through the application of Rasch measurement techniques. *Education Sciences*, *8*(4), 184. https://doi.org/10.3390/educsci8040184

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons. https://doi.org/10.1002/9781119970583

Behi, R., & Nolan, M. (1995). Reliability: consistency and accuracy in measurement. *British Journal of Nursing*, *4*(8), 472–475. https://doi.org/10.12968/bjon.1995.4.8.472

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief. *Office of Educational Technology, US Department of Education*.

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education*, *4*(1), 1416898. https://doi.org/10.1080/2331186X.2017.1416898

Boone, W. J., Staver, J. R., Yale, M. S., Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Wright maps: First steps. *Rasch Analysis in the Human Sciences*, 111–136. https://doi.org/10.1007/978-94-007-6857-4_6

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061. https://doi.org/10.1037/0033-295X.111.4.1061

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: the definitive guide to questionnaire design--for market research, political polls, and social and health questionnaires*. John Wiley & Sons.

Brochado, A. (2009). Comparing alternative instruments to measure service quality in higher education. *Quality Assurance in Education*, *17*(2), 174–190. https://doi.org/10.1108/09684880910951381

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, *30*(1), 3–12. https://doi.org/10.1111/j.1745-3992.2010.00195.x

Care, E., Kim, H., Vista, A., & Anderson, K. (2018). Education System Alignment for 21st Century Skills: Focus on Assessment. *Center for Universal Education at The Brookings Institution*.

Chalmers, D. (2007). A review of Australian and international quality systems and indicators of learning and

teaching. *Carrick Institute for Learning and Teaching in Higher Education*, *1*(2), 1–122.
Chan, S.-W., Looi, C.-K., & Sumintono, B. (2021). Assessing computational thinking abilities among Singapore secondary students: a Rasch model measurement analysis. *Journal of Computers in Education*, *8*, 213–236. https://doi.org/10.1007/s40692-020-00177-2
Cohen, D., & Sasson, I. (2016). Online quizzes in a virtual learning environment as a tool for formative assessment. *Journal of Technology and Science Education (JOTSE)*, *6*(3), 188–208.
Cohen, L., Manion, L., & Morrison, K. (2017). Validity and reliability. In *Research methods in education* (pp. 245–284). Routledge. https://doi.org/10.4324/9781315456539-14
Colville, G., Darkins, J., Hesketh, J., Bennett, V., Alcock, J., & Noyes, J. (2009). The impact on parents of a child's admission to intensive care: Integration of qualitative findings from a cross-sectional study. *Intensive and Critical Care Nursing*, *25*(2), 72–79. https://doi.org/10.1016/j.iccn.2008.10.002
Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *The American Journal of Medicine*, *119*(2), 166-e7. https://doi.org/10.1016/j.amjmed.2005.10.036
Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.
Darman, D. R., Suhandi, A., Kaniawati, I., Samsudin, A., & Wibowo, F. C. (2024). Development and Validation of Scientific Inquiry Literacy Instrument (SILI) Using Rasch Measurement Model. *Education Sciences*, *14*(3), 322. https://doi.org/10.3390/educsci14030322
De Castella, K., & Byrne, D. (2015). My intelligence may be more malleable than yours: The revised implicit theories of intelligence (self-theory) scale is a better predictor of achievement, motivation, and student disengagement. *European Journal of Psychology of Education*, *30*, 245–267. https://doi.org/10.1007/s10212-015-0244-y
Dunn, K. E., Airola, D. T., Lo, W.-J., & Garrison, M. (2013). What teachers think about what they can do with data: Development and validation of the data driven decision-making efficacy and anxiety inventory. *Contemporary Educational Psychology*, *38*(1), 87–98. https://doi.org/10.1016/j.cedpsych.2012.11.002
Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measuremen
https://doi.org/10.1037/0021-9010.73.3.421on belief, attitude, intention, and behavior. *Journal of Applied Psychology*, *73*(3), 421.
Fink, A., & Litwin, M. S. (1995). *How to measure survey reliability and validity* (Vol. 7). Sage. https://doi.org/10.4135/9781483348957
Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 286. https://doi.org/10.1037/1040-3590.7.3.286
Galy, E., Downey, C., & Johnson, J. (2011). The effect of using e-learning tools in online and campus-based classrooms on student performance. *Journal of Information Technology Education: Research*, *10*(1), 209–230. https://doi.org/10.28945/1503
Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods*, *18*(4), 454. https://doi.org/10.1037/a0030005
Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education*, *57*(4), 2333–2351. https://doi.org/10.1016/j.compedu.2011.06.004
Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The Qualitative Report*, *8*(4), 597–607.
Green, S. B., Levy, R., Thompson, M. S., Lu, M., & Lo, W.-J. (2012). A proposed solution to the problem with using completely random data to assess the number of factors with parallel analysis. *Educational and Psychological Measurement*, *72*(3), 357–374. https://doi.org/10.1177/0013164411422252
Groves, R. M. (1987). Research on survey data quality. *The Public Opinion Quarterly*, *51*, S156–S172. https://doi.org/10.1086/269077
Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, *43*, 115–135. https://doi.org/10.1007/s11747-014-0403-8
Hitt, D. H., & Tucker, P. D. (2016). Systematic review of key leader practices found to influence student achievement: A unified framework. *Review of Educational Research*, *86*(2), 531–569. https://doi.org/10.3102/0034654315614911
Hora, M. T., Bouwma-Gearhart, J., & Park, H. J. (2017). Data driven decision-making in the era of accountability: Fostering faculty data cultures for learning. *The Review of Higher Education*, *40*(3), 391–426. https://doi.org/10.1353/rhe.2017.0013
Jackson, D. A. (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, *74*(8), 2204–2214. https://doi.org/10.2307/1939574
Johnson, R. B., & Christensen, L. (2019). *Educational research: Quantitative, qualitative, and mixeapproaches*.

Sage publications.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Karlin, O., & Karlin, S. (2018). Making Better Tests with the Rasch Measurement Model. *InSight: A Journal of Scholarly Teaching*, *13*, 76–100. https://doi.org/10.46504/14201805ka

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, *65*(23), 2276–2284. https://doi.org/10.2146/ajhp070364

Kreijns, K., Bijker, M., & Weidlich, J. (2020). A Rasch analysis approach to the development and validation of a social presence measure. *Rasch Measurement: Applications in Quantitative Educational Research*, 197–221. https://doi.org/10.1007/978-981-15-1800-3_11

Kulasegaram, K., & Rangachari, P. K. (2018). Beyond "formative": assessments to enrich student learning. *Advances in Physiology Education*, *42*(1), 5–14. https://doi.org/10.1152/advan.00122.2017

Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education*, *36*, 143–152. https://doi.org/10.1016/j.tate.2013.07.010

Lewis, B. R., Templeton, G. F., & Byrd, T. A. (2005). A methodology for construct development in MIS research. *European Journal of Information Systems*, *14*(4), 388–400. https://doi.org/10.1057/palgrave.ejis.3000552

Lim, T. M., & Yunus, M. M. (2021). Teachers' perception towards the use of Quizizz in the teaching and learning of English: A systematic review. *Sustainability*, *13*(11), 6436. https://doi.org/10.3390/su13116436

Martínez-Caro, E., Cegarra-Navarro, J. G., & Cepeda-Carrión, G. (2015). An application of the performance-evaluation model for e-learning quality in higher education. *Total Quality Management & Business Excellence*, *26*(5–6), 632–647. https://doi.org/10.1080/14783363.2013.867607

Massof, R. W. (2005). Application of stochastic measurement models to visual function rating scale questionnaires. *Ophthalmic Epidemiology*, *12*(2), 103–124. https://doi.org/10.1080/09286580590932789

Massof, R. W. (2011). Understanding Rasch and item response theory models: applications to the estimation and validation of interval latent trait measures from responses to rating scale questionnaires. *Ophthalmic Epidemiology*, *18*(1), 1–19. https://doi.org/10.3109/09286586.2010.545501

Massof, R. W., & Rubin, G. S. (2001). Visual function assessment questionnaires. *Survey of Ophthalmology*, *45*(6), 531–548. https://doi.org/10.1016/S0039-6257(01)00194-1

Maydeu-Olivares, A. (2017). Assessing the size of model misfit in structural equation models. *Psychometrika*, *82*(3), 533–558. https://doi.org/10.1007/s11336-016-9552-7

Neumann, I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-baseanalyses of a nature of science test. *International Journal of Science Education*, *33*(10), 1373–1405. https://doi.org/10.1080/09500693.2010.511297

Park, M. S., Kang, K. J., Jang, S. J., Lee, J. Y., & Chang, S. J. (2018). Evaluating test-retest reliability in patient-reported outcome measures for older people: A systematic review. *International Journal of Nursing Studies*, *79*, 58–69. https://doi.org/10.1016/j.ijnurstu.2017.11.003

Pellegrino, J. W., & Quellmalz, E. S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education*, *43*(2), 119–134. https://doi.org/10.1080/15391523.2010.10782565

Pianta, R. C., Hamre, B. K., & Allen, J. P. (2012). Teacher-student relationships and engagement: Conceptualizing, measuring, and improving the capacity of classroom interactions. In *Handbook of research on student engagement* (pp. 365–386). Springer. https://doi.org/10.1007/978-1-4614-2018-7_17

Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research*, *15*(2), 20111. https://doi.org/10.1103/PhysRevPhysEducRes.15.020111

Radhakrishna, R. B. (2007). Tips for developing and testing questionnaires/instruments. *The Journal of Extension*, *45*(1), 25.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*(6), 544–559. https://doi.org/10.1080/00223891.2010.496477

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment*, *12*(3), 287. https://doi.org/10.1037/1040-3590.12.3.287

Shen, J., Cooley, V. E., Ma, X., Reeves, P. L., Burt, W. L., Rainey, J. M., & Yuan, W. (2012). Data-informed decision making on high-impact strategies: Developing and validating an instrument for principals. *The Journal of Experimental Education*, *80*(1), 1–25. https://doi.org/10.1080/00220973.2010.550338

Soeharto, S., & Csapó, B. (2022). Assessing Indonesian student inductive reasoning: Rasch analysis. *Thinking Skills and Creativity*, *46*, 101132. https://doi.org/10.1016/j.tsc.2022.101132

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of

the art. *Review of Educational Research*, *83*(4), 598–642. https://doi.org/10.3102/0034654313496870

Spooren, P., Mortelmans, D., & Denekens, J. (2007). Student evaluation of teaching quality in higher education: development of an instrument based on 10 Likert-scales. *Assessment & Evaluation in Higher Education*, *32*(6), 667–679. https://doi.org/10.1080/02602930601117191

Stiggins, R. (2010). Essential formative assessment competencies for teachers and school leaders. In *Handbook of formative assessment* (pp. 233–250). Routledge.

Stigler, J. W., & Hiebert, J. (2009). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. Simon and Schuster.

Sürücü, L., & Maslakci, A. (2020). Validity and reliability in quantitative research. *Business & Management Studies: An International Journal*, *8*(3), 2694–2726. https://doi.org/10.15295/bmij.v8i3.1540

Suskie, L. (2018). *Assessing student learning: A common sense guide*. John Wiley & Sons.

Talib, A. M., Alomary, F. O., & Alwadi, H. F. (2018). Assessment of student performance for course examination using Rasch measurement model: A case study of information technology fundamentals course. *Education Research International*, *2018*(1), 8719012. https://doi.org/10.1155/2018/8719012

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

Törmäkangas, K. (2011). Advantages of the Rasch measurement model in analysing educational tests: an applicator's reflection. *Educational Research and Evaluation*, *17*(5), 307–320. https://doi.org/10.1080/13803611.2011.630562

Van Iddekinge, C. H., Roth, P. L., Raymark, P. H., & Odle-Dusseau, H. N. (2012). The criterion-related validity of integrity tests: An updated meta-analysis. *Journal of Applied Psychology*, *97*(3), 499. https://doi.org/10.1037/a0021196

Vaske, J. J., Beaman, J., & Sponarski, C. C. (2017). Rethinking internal consistency in Cronbach's alpha. *Leisure Sciences*, *39*(2), 163–173. https://doi.org/10.1080/01490400.2015.1127189

Viladrich, C., Angulo-Brunet, A., & Doval, E. (2017). A journey around alpha and omega to estimate internal consistency reliability. *Anales de Psicología*, *33*(3), 755–782. https://doi.org/10.6018/analesps.33.3.268401

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles, Policy & Practice*, *11*(1), 49–65. https://doi.org/10.1080/0969594042000208994

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 97–116. https://doi.org/10.1111/j.1745-3984.1977.tb00031.x

Zeng, L. M., Fryer, L. K., & Zhao, Y. (2023). A comparison of three major instruments used for the assessment of university student experience: Toward a comprehensive and distributed approach. *Higher Education Quarterly*, *77*(1), 27–44. https://doi.org/10.1111/hequ.12363

Zumbo, B. D. (2006). 3 validity: foundational issues and statistical methodology. *Handbook of Statistics*, *26*, 45–79. https://doi.org/10.1016/S0169-7161(06)26003-6