Journal of Education and e-Learning Research
Vol. 11, No. 3, 499-509, 2024
ISSN(E) 2410-9991 / ISSN(P) 2518-0169
DOI: 10.20448/jeelr.v11i3.5861

Reliability and separation index analysis of mathematics questions integrated with the cultural architecture framework using the Rasch model

Muh. Fitrah[1]
Anastasia Sofroniou[2]
Ofianto[3] ✉
Loso Judijanto[4]
Widihastuti[5]

(✉ *Corresponding Author*)

[1,5]*Department of Education Research and Evaluation, Graduate School, Yogyakarta State University, Yogyakarta, Indonesia.*
[1]*Email: muhfitrah.2023@student.uny.ac.id*
[5]*Email: widihastuti@uny.ac.id*
[2]*School of Computing and Engineering, University of West London, London, UK.*
[2]*Email: anastasia.sofroniou@uwl.ac.uk*
[3]*Padang State University, Padang, Indonesia.*
[3]*Email: ofianto@fis.unp.ac.id*
[4]*IPOSS Jakarta, Jakarta, Indonesia.*
[4]*Email: losojudijantobumn@gmail.com*

Abstract

This research uses Rasch model analysis to identify the reliability and separation index of an integrated mathematics test instrument with a cultural architecture structure in measuring students' mathematical thinking abilities. The study involved 357 students from six eighth-grade public junior high schools in Bima. The selection of schools was based on average school exam scores and considered the effectiveness of the learning process that used cultural settings to explore mathematical content. Data analysis was conducted using Microsoft Excel to calculate the content validity of Aiken's index with four experts and the jMetrik software to measure reliability and the separation index. The research results indicate that the mathematics test instrument passed validation by mathematics experts and measurements with a valid content validity level. Rasch model calibration shows a very high level of instrument reliability. Separation analysis on the logit scale indicates the instrument's ability to differentiate students with different ability levels with good homogeneity in the distribution of test items and individual abilities. Scale quality statistics show good item response variability, low error rates and a high separation index. This study has limitations because it focuses solely on multiple-choice questions. Similar research must be conducted using other types of questions (such as those used in PISA, namely open-constructed and closed-constructed questions) and integrating other mathematical materials within relevant cultural architectural structures.

Keywords: Culture, Mathematics test instrument, Rasch model, Reliability, Separation index.

Citation | Fitrah, M., Sofroniou, A., Ofianto, Judijanto, L., & Widihastuti. (2024). Reliability and separation index analysis of mathematics questions integrated with the cultural architecture framework using the Rasch model. *Journal of Education and E-Learning Research, 11*(3), 409–519. 10.20448/jeelr.v11i3.5861

History:
Received: 23 February 2024
Revised: 29 May 2024
Accepted: 12 June 2024
Published: 31 July 2024

Publisher: Asian Online Journal Publishing Group

Funding: This research is supported by Indonesian Higher Education Financing Agency of the Ministry of Education, Culture, Research and Technology of the Republic of Indonesia, Education Fund Management Institution and Indonesian Education Scholarship (Grant number: 01321/BPPT/BPI.06/9/2023).

Institutional Review Board Statement: The Ethical Committee of the Yogyakarta State University's Ethics Committee, Indonesia has granted approval for this study on 7 October 2023 (Ref. No. B/2981/UN34.17/KM/2023).

Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Competing Interests: The authors declare that they have no competing interests.

Authors' Contributions: Contributed to conceptualization, data curation, formal analysis, software development, and drafting the original manuscript, M.F.; involved in reviewing and editing the manuscript, as well as contributing to methodology and validation, A.S. and W.I.D.; played a role in investigation, resourcing, data curation, and formal analysis, O.F.I.; contributed to formal analysis, visualization, and project administration, L.J. All authors have read and agreed to the published version of the manuscript.

Contents

Contribution of this paper to the literature

Integration of culture in mathematics learning specifically the architectural structure of culture is aimed at developing and maximizing mathematics test instruments using Rasch model calibration. This approach measures the reliability and separation index of valid instruments which can differentiate the levels of student ability ultimately impacting student engagement and academic achievement.

1. Introduction

Improving the quality of education is key to create high-quality human resources. One strategy to achieve this is to develop students' critical thinking abilities. This skill is crucial for forming sound arguments and making informed decisions (Hernawati & Nurbayani, 2018; Kivunja, 2014; Mahdi, Nassar, & Almuslamani, 2020). Mathematics is considered one of the most effective subjects for enhancing students' critical thinking abilities (Aizikovitsh-Udi & Cheng, 2015; Jones, 2015).

The development in the fields of science and technology in the 21st century has resulted in significant challenges. The characteristics of the 21st century are marked by increasingly interconnected scientific disciplines leading to rapid synergy between them. Such rapid changes affect various aspects, particularly education and specifically mathematics education. A robust assessment process is required to achieve these developments in teaching mathematics.

Assessing student learning outcomes in accordance with Ministry of Education and Culture Regulation No. 66 of 2013 encompasses attitudes, knowledge and skills competencies. One of the essential knowledge domains in the 21st century is critical thinking skills which are cultivated through mathematics education. Measuring students' critical thinking abilities in mathematics through questions is crucial in educational evaluation (Harjo, Kartowagiran, & Mahmudi, 2019; Monrat, Phaksunchai, & Chonchaiya, 2022). Critical thinking skills necessitate students to analyze, evaluate and draw logical and rational conclusions from information (Raj, Chauhan, Mehrotra, & Sharma, 2022).

Several studies conducted by previous researchers have highlighted the issue that the questions used by teachers have not effectively measured students' critical thinking abilities (Adams & Wieman, 2011; Bray, Girvan, & Chorcora, 2023; Priatna, Lorenzia, & Widodo, 2020; Widana, Parwata, & Sukendra, 2018). Therefore, valid mathematics questions are needed with the goal of impacting student achievement particularly by enhancing and developing students' critical thinking abilities.

The Rasch model is one of the tools used in measurement to answer an item correctly and solely depends on the ability of the student and the difficulty level of the item (Andrich & Marais, 2019). The Rasch model provides a statistical interpretation of the difficulty level of items based on student responses (Clements, Sarama, & Liu, 2008; Karlimah, Andriani, & Suryana, 2020). Analysis using the Rasch model can provide information about the quality of the instrument used and the overall quality of student responses as well as the interaction between respondents and test items (Chan, Ismail, & Sumintono, 2014).

The Rasch model is used to measure students' abilities based on their responses to test items that have been developed (Doyle, Hula, McNeil, Mikolic, & Matthews, 2005; Gorin, Embretson, & McKay, 2008; Planinic, Boone, Susac, & Ivanjek, 2019) and can be used to test statistical assumptions such as item invariance (Engelhard Jr, 2013; Kubinger, Rasch, & Yanagida, 2011; Makransky, Rogers, & Creed, 2015; Schneider, Strobl, Zeileis, & Debelak, 2022). One can test item and person invariance but it is necessarily a consequence of the Rasch model (Holland, 1990). The Rasch model assists improve the validity and reliability of assessment tools but users need to critically comprehend the statistical concepts and underlying assumptions to interpret the results validly and reliably.

Item fit reflects how well the items operate according to the Rasch model. However, it should be noted that although an item may function normally, it does not always indicate conformity to the Rasch model. This analysis is useful for teachers in their efforts to improve the quality of their teaching (Sumintono & Widhiarso, 2015). Studies indicate that the Rasch model provides accurate feedback for improvement (Noben, Maulana, Deinum, & Hofman, 2021; Van der Lans, Van de Grift, & van Veen, 2018) reflects teaching practices (Kaspersen, Pepin, & Sikko, 2017; Zile-Tamsen, 2017) and helps teachers understand students' needs (Razak, bin Khairani, & Thien, 2012).

A problem identified in the school is that teachers have not been able to develop assessment instruments to be implemented with contextual, complex, non-routine written test techniques that require reasoning, argumentation, and creativity to solve. This is based on the analysis of documents held by middle school mathematics teachers in Bima, particularly evaluation tools used to measure mathematics learning achievement such as assessment instruments. Currently, there is a lack of authoritative literature or guidelines for the use of mathematical ability measuring tools particularly in middle school mathematics in addition to supporting data. Furthermore, this includes the concept of mathematics test instruments integrated with local culture.

Additionally, based on initial interviews with mathematics teachers at the school, it shows a lack of knowledge among the teachers in ensuring the validity of questions for each mathematical content in class. Teachers tend to rely on questions from the internet without verifying their quality and validity. Quality improvement activities for teacher professionalism such as training and workshops are also not implemented constructively and continuously.

This indicates the need for further efforts to strengthen teachers' capacity to design and use quality assessment instruments.

The integration of culture into mathematics test instruments is significantly relevant to recognizing cultural diversity in the students' learning environment. In this concept, mathematics test instruments are designed by considering the cultural architecture structure as a strategy to understand student diversity through item questions. This approach reflects a commitment to creating a constructive learning environment. The learning evaluation process becomes not only a constructive measure of academic performance but also a means to understand the cultural context of students in understanding and applying mathematical concepts by incorporating cultural aspects into mathematics test instruments.

The importance of local wisdom values in mathematics learning as an effort to address moral degradation and shape character is emphasized. Cultural values are integrated into the 2013 curriculum student books (Nuraini, 2022). Examples of this integration include cultural aspects of mathematics such as calculation, measurement, building design, location determination, playing activities, thinking activities and problem-solving activities. Various studies conducted by researchers in Indonesia to describe culture through mathematics include ethno mathematics exploration in local culture (Lidinillah, Rahman, Wahyudin, & Aryanto, 2022). High-quality tests should not only provide challenges appropriate to the expected difficulty level but also have strong reliability. The reliability of a test measures the extent to which it consistently yields uniform results. Reliability refers to the degree of consistency or stability in test results (Reynolds, Livingston, Willson, & Willson, 2010) reflecting the consistency of test scores when measured through the same process (Jonsson & Svingby, 2007).

High reliability levels explain the consistency of measurement results shown at different times on the same subject. A test is considered reliable if the scores obtained have a high correlation with the total scores. The reliability value of an instrument is influenced by the subjects being measured, the instrument's user and the instrument itself. Meanwhile, the separation index explains how well questions can differentiate students' abilities among individuals for a specific criterion. The separation index has a range of values varying from zero to infinity (Leeming & Wong, 2016) and when the separation index value is high, it indicates that test items are well distributed across difficulty levels with values above 2.0 considered acceptable (Bond & Fox, 2013). A low separation index value indicates that a developed instrument is not effective particularly in identifying students' differences (Leeming & Wong, 2016). Research findings by Thomas, Anderson, and Nashon (2008) show a separation index of 5.93. Resnick (2005) describes a separation index of 4.70 and 9.88 in measurement tool development. Additionally, Sari and Abdurrahman (2019) achieved a separation index of 3.19 in test product development.

Various studies using the Rasch model to analyze measurement instrument quality have produced significant findings. Erfan, Maulyda, Ermiana, Hidayati, and Widodo (2020) found significant differences in validity and reliability between classical test theory and the Rasch model approaches in measuring the ability to differentiate between series and parallel circuits. Schulz (2023) assessed students' problem-solving abilities in permutation and combination using the Rasch model. The arithmetic operation abilities of elementary school students were measured using various Rasch analyses with the Rasch model specifically used to analyze students' difficulties with decimal numbers (Bolondi, Cascella, & Giberti, 2017).

Furthermore, other studies have demonstrated the validity, reliability, difficulty levels and discriminative abilities of test instruments using the Rasch model (Mui Lim, Rodger, & Brown, 2009; Neumann, Neumann, & Nehm, 2011). Ridzuan, Lian, Fozee, and Nasser (2020) explored reliability and validity using a superitem test while Saidi and Siew (2019) focused on measuring the reliability and validity of statistical thinking test instruments.

There are some relevant studies that explain the impact of using the Rasch model on critical thinking skills instruments in mathematics especially reliability and separation index. Research that specifically discusses the reliability and separation index of mathematical test instruments integrated with the cultural architecture framework is still very limited. The average of these relevant studies shows the level of validity, level of difficulty and other psychometric characteristics. Therefore, the formulation of the research problem is how the level of reliability and separation index of mathematical test instruments integrated with the cultural architecture framework in measuring students' mathematical abilities based on the Rasch Model?

2. Literature Review

2.1. Mathematics with Culture

The process of integrating cultural architecture into mathematics education is an innovative approach aimed at enhancing students' achievement, motivation and knowledge of the subject matter (Fauzi, Hanum, Jailani, & Jatmiko, 2022; Kurniawan, Purwoko, & Setiana, 2023; Prasad Pant & Chandra Luitel, 2020). It can have a profound impact on students' enjoyment, understanding and learning of mathematics within the context of everyday life by transforming certain elements of cultural architecture into instructional materials and problems. The advantages of integrating cultural elements into mathematics include fostering learning motivation, improving students' critical thinking skills and enabling effective problem-solving (Fouze & Amit, 2017; Simamora & Saragih, 2019).

Examples of implementing mathematics problems integrated with cultural architecture such as calculating the surface area of a temple (Munthahana & Budiarto, 2020) allow students not only to learn mathematical formulas but also to understand how these concepts are reflected in the cultural architecture itself. Furthermore, integrating cultural architecture into mathematics education can boost students' pride in their own culture (Meaney, Trinick, & Allen, 2021; Zubaidah & Arsih, 2021). Students become more connected to their cultural heritage and feel valued in the learning process by incorporating examples of local architecture into mathematics education. This can lead to higher motivation to learn and increased participation in mathematics classes (Asfar, Asfar, & Nurannisa, 2021; Garcia & Pacheco, 2013).

Numerous studies have emphasized and demonstrated the importance of integrating culture into mathematics education. Research by Parker, Bartell and Novak (2017) show that students are more enthusiastic about understanding mathematical concepts when presented within a cultural context. Integrating culture into mathematics instruction not only boosts student motivation but also strengthens their understanding of mathematical concepts (Wong & Wong, 2021).

2.2. Rasch Model in Item Measurement

The use of the Rasch model and its paradigm refers to employ the Rasch model as a reference framework in developing measurement tools. The Rasch model is based on principles of unidimensionality and local independence and includes principles such as monotonicity and invariance (Andrich & Marais, 2019; Baghaei, 2012). The Rasch model has various applications including measuring individual abilities and item difficulties in test development and analyzing test data to identify poorly functioning items (Edelsbrunner & Dablander, 2018; Petra & Aziz, 2020; Sinnema, Ludlow, & Robinson, 2016; Takács, Kárász, Takács, Horváth, & Oláh, 2021). This implies that criteria for evaluating test results are determined by the properties of the Rasch model. If test results or data do not meet the criteria, the necessary action is to inspect or check the data rather than seeking another model to explain the data. This aligns with the Rasch paradigm. The model is used to assess the degree of conformity of the generated data with model criteria. Non-conformity with model criteria guides necessary improvements. Analysis to assess the fit between the model and test results is commonly performed.

Statistics from fit analysis are used as the basis for determining whether an item fits. Analysis using the Rasch model solely to obtain fit statistics is suboptimal as it underutilizes the Rasch model as a reference framework and diagnostic tool in developing measurement instruments. Embretson and Reise (2000) acknowledge the strengths of the Rasch model but do not recommend its use in all situations to prevent the removal of important items that may alter the measurement construct. According to the Rasch model, the removal of items should not be solely based on statistical criteria. The optimal utilization of the Rasch model includes more in-depth fit analysis, unidimensionality and independent response analysis and Differential Item Functioning (DIF) analysis.

Some characteristics of the Rasch model include: 1) the Rasch logit scale. 2) Item characteristic curve (ICC). 3) Item difficulty in the Rasch model. 4) Objective comparison and 5) guessing in the Rasch model (Andrich, Marais, & Humphry, 2012; Bansilal, 2015; Kaspersen et al., 2017; Long, Bansilal, & Debba, 2014).

3. Methods

3.1. Research Design and Participants

The quantitative method employed in this research is the Rasch model analysis which focuses on evaluating the reliability and separation index of mathematics questions integrated with cultural architecture. Eighth-grade students from a junior high school in Bima City, West Nusa Tenggara, Indonesia participated in this study. The research sample was selected based on the classification of average school exam scores in mathematics categorized into high, medium and low criteria. Additionally, schools actively implementing outdoor learning approaches involving visits to cultural locations in the surrounding areas, including the Asi Mbojo Museum (Bima), Heroes' Cemetery and traditional village were selected. Thus, the selected participants in this research context representatively reflect the cultural influence on the reliability and separation index analysis of mathematics questions using the Rasch model. Therefore, 6 public junior high schools in Bima became the subjects of the study (see Table 1).

Table 1. Research sample.

School	Average school exam score in mathematics	Sample	Category
Public Junior High School 7 Bima City	36.58	50	High
Public Junior High School 1 Bima City	36.51	68	High
Public Junior High School 6 Bima City	35.62	54	Medium
Public Junior High School 2 Bima City	35.28	60	Medium
Public Junior High School 4 Bima City	34.59	65	Low
Public Junior High School 14 Bima City	34.51	60	Low

Note: Categories refer to the classification of schools based on students' final examination scores.

The selection of subjects is based on the need for developing questions to be used in the learning process. The subjects of this research are eighth-grade students from public junior high schools who have completed all the subject matter in the basic competency of solving problems related to the surface area and volume of flat-sided space objects (cubes) as well as solving problems related to the surface area and volume of flat-sided space objects (rectangular prisms). This is done considering that the test instrument developed refers to the mathematics graduate competency standards for junior high school. Therefore, out of the 6 identified schools, the subjects in this study are 357 students.

3.2. Instruments and Data Collection

The instrument used is a set of mathematics questions developed by considering relevant cultural aspects in the form of 25 multiple-choice items. Data collection was conducted through a written exam given to the participants who then answered within the specified time limit. This multiple-choice item is arranged based on the curriculum framework of junior high school mathematics set by the standard, curriculum and educational assessment body in a standardized manner (see Table 2).

Table 2. Matrix of mathematics questions integrated with cultural architectural structures.

Competency basic	Subject matter	Question indicators	Cognitive aspects	Question forms	Number of questions
Solving problems related to the surface area and volume of flat-sided spatial structures (Cubes).	Cube, rectangular prism and pyramid.	Students can calculate the volume of a cube given the surface area of the cube.	C3, C4, C5, and C6	Multiple choice	25
Solving problems related to the surface area and volume of flat-sided spatial structures (Rectangular prisms).		Students can calculate the volume of a rectangular prism given the length, width and surface area of the prism. Students can calculate the surface area of a rectangular prism given the length, width and height of the prism. Students can calculate the surface area of a pyramid given the base and height of the pyramid. Determining the volume of a prism given the base and surface area of the prism.	-	-	-

3.3. Data Analysis and Interpretation

The collected data was analyzed using the Rasch model method with the jMetrik software. jMetrik software is one of the programs that can be used for item response theory (Aksu, Guzeller, & Eser, 2019). jMetrik can be used to analyze the Rasch model, one- parameter logistic model , two- parameter logistic model , three-parameter logistic model, partial credit model and others (Avetisyan, 2015). The instrument was validated through a content validation process by experts (expert judgment) before analyzing the data using the jMetrik software. Four experts were involved in assessing the content validity of the instrument and expert consensus was used to determine the level of content validity.

The instrument's validity in this study was measured through validation processes, particularly content validity aimed at assessing instruments that can measure students' thinking abilities developed based on learning indicators and materials in middle school mathematics through assessment by experts. In this study, four experts were involved in the validation assessment itself. The instrument's validity process was conducted using a validation sheet consisting of four rating scales: 4 (very appropriate), 3 (appropriate), 2 (inappropriate) and 1 (very inappropriate). This scale reflects the level of conformity between the item questions and the specified indicators. Expert agreement is measured with the Aiken index as described by Retnawati (2016).

$$\text{Formula: } V = \frac{\sum_{i=1}^{n} s_i}{R(c-1)}$$

Explanation

V : Index of expert agreement on item validity.

$\sum_{i=1}^{n} s_i$: The sum of scores given by each expert is subtracted from the lowest score within the used category.

R : Number of experts.

c : Number of categories that can be chosen by experts.

The interpretation of the V index calculation results can be categorized as follows: if the index is less than or equal to 0.4, the validity is low. If the index is between 0.4-0.8, the validity is moderate and if the index is greater than 0.8, the validity is high (Retnawati, 2016).

Content validity analysis is performed using Microsoft Excel.

After content validity, this research focuses on the main objective which is the analysis of the level of reliability of the mathematics test instrument and the effectiveness of the separation index integrated with cultural architecture using the Rasch model with the jMetrik software. The outcome of jMetrik manifests as the table of Joint Maximum Likelihood Estimation (JMLE) item statistics referred to as the final JMLE item statistics serving as the pivotal element in addressing the research questions. This table focuses on item, difficulty, WMS (weighted mean-square (infit)) and UMS (unweighted mean-square (outfit)) in relation to 25 mathematics questions answered by 357 students.

"Item" denotes the unique identification of each question, "difficulty" reflects the level of complexity while WMS and UMS provide metrics for evaluating student performance. This comprehensive analysis of the final JMLE item statistics table offers insights into both the difficulty and comprehension levels of students regarding the mathematics questions serving as a guiding tool for enhancing mathematics education in the educational setting. The interpretation of the Rasch model analysis results uses criteria from Fisher (2007) to conclude the reliability level of the instrument. For further clarity, please refer to Table 3.

Table 3. Rating scale instrument quality criteria.

Criterion	Poor	Fair	Good	Very good	Excellent
Item model fit mean-square range extremes	< 0.33 or >3.0	0.34 - 2.9	0.5 - 2.0	0.71 - 1.4	0.77 - 1.3
Person and item measurement reliability	<0.67	0.67-.80	0.81-0.90	0.91-0.94	>0.94
Person and item strata separated	2 or less	2-3	3-4	4-5	>5
Ceiling effect: % maximum extreme scores	>5%	2-5%	1-2%	0.5-1%	<0.5%
Floor effect: % minimum extreme scores	>5%	2-5%	1-2%	0.5-1%	<0.5%

4. Results

The research instrument has undergone a validation process by four experts, namely, a mathematics learning expert and a measurement expert. The results of the content validity analysis indicate that each item of the instrument has been declared valid for use. This validity is obtained because the material to be tested has been adjusted to the 2013 curriculum and the implementation pattern of the independent curriculum in junior high schools. Please refer to the Aiken validity index table below.

Table 4. Content validity with the Aiken validity index.

Items	Rater 1	Rater 2	Rater 3	Rater 4	s1	s2	s3	s4	Σs	V
1	3	3	4	3	2	2	3	2	9	0.75
2	3	3	4	3	2	2	3	2	9	0.75
3	4	3	3	3	3	2	2	2	9	0.75
4	3	4	3	4	2	3	2	3	10	0.83
5	3	4	3	3	2	3	2	2	9	0.75
6	2	2	3	3	1	1	2	2	6	0.50
7	3	3	4	4	2	2	3	3	10	0.83
8	3	4	3	4	2	3	2	3	10	0.83
9	4	3	3	4	3	2	2	3	10	0.83
10	4	4	4	3	3	3	3	2	11	0.92
11	3	4	3	4	2	3	2	3	10	0.83
12	4	4	3	4	3	3	2	3	11	0.92
13	3	3	4	4	2	2	3	3	10	0.83
14	2	2	3	3	1	1	2	2	6	0.50
15	3	3	3	2	2	2	2	1	7	0.58
16	3	3	3	3	2	2	2	2	8	0.67
17	3	4	4	4	2	3	3	3	11	0.92
18	4	3	3	4	3	2	2	3	10	0.83
19	3	3	4	3	2	2	3	2	9	0.75
20	2	3	2	3	1	2	1	2	6	0.50
21	2	3	2	2	1	2	1	1	5	0.42
22	4	4	4	4	3	3	3	3	12	1.00
23	4	3	4	4	3	2	3	3	11	0.92
24	3	4	3	4	2	3	2	3	10	0.83
25	3	3	4	3	2	2	3	2	9	0.75
Average										0.76

Note: The column $R(c-1)$ is not included in the table as its value is constant (always 12). This decision was made to tidy up the data presentation and enhance the clarity of the analysis without compromising crucial information.

Based on Table 4, it can be explained that the average Aiken validity index is 0.76 which falls into the moderate category. This value indicates an acceptable level of validity for the research instrument. It means that the overall consistency of the raters' assessment of the content of each item is quite good, and the results are valid. This conclusion is based on the interpretation of the Aiken validity index value which generally indicates the level of agreement among raters regarding the content of each research item. Therefore, this research instrument can be relied upon to measure the critical thinking abilities of eighth-grade junior high school students.

The results of the Rasch model calibration to measure the quality of mathematics questions using the jMetrik software are shown in Table 5. In the context of evaluating the quality of questions, this calibration provides information about the difficulty level, weighted mean square (WMS) for infit and unweighted mean square (UMS) for outfit. The difficulty level of items in the context of the Rasch model analyzed with jMetrik is akin to Z scores. The comparison with Z scores is employed because it aids in understanding that lower values indicate lower or easier difficulty levels while higher values indicate higher or more challenging difficulty levels. The range of values used in jMetrik analysis is situated between -3 and +3. The lower the difficulty level value, the easier the item; conversely, the higher the value, the more difficult the item.

Table 5. Joint maximum likelihood estimation (JMLE) for mathematics items: Cultural architecture integration and reliability analysis.

Items	Difficulty	Std. error	WMS	UMS
item1	-0.73	0.14	0.88	0.68
item2	-0.87	0.14	0.73	0.55
item3	-0.10	0.13	1.21	1.20
item4	0.50	0.13	0.75	0.65
item5	-0.51	0.13	1.42	1.39
item6	0.95	0.13	0.75	0.83
item7	-0.31	0.13	0.86	0.75
item8	-0.32	0.13	1.29	1.29
item9	1.31	0.13	1.21	1.42
item10	0.54	0.13	0.72	0.65
item11	-1.09	0.14	0.66	0.44
item12	0.14	0.13	0.78	0.65
item13	-1.13	0.15	0.88	0.95
item14	-0.48	0.13	0.78	0.63
item15	-0.31	0.13	0.86	0.68
item16	0.42	0.13	1.39	1.34
item17	-0.48	0.13	0.79	0.65

Items	Difficulty	Std. error	WMS	UMS
item18	0.44	0.13	0.88	0.82
item19	-0.51	0.13	0.73	0.56
item20	-0.76	0.14	0.76	0.56
item21	2.30	0.14	1.74	3.31
item22	-0.41	0.13	0.82	0.62
item23	0.06	0.13	1.12	0.98
item24	1.97	0.14	2.03	3.82
item25	-0.62	0.14	0.73	0.57

According to Table 5, it can be interpreted that questions with very low difficulty levels (easy) include items 11, 13, 1, 2, 20, 25, 5, 7, 8, 15, 17, 14, 17, 19, and 3. Meanwhile, questions considered easy items include 4, 10, 12, 16, 18, and 23. On the other hand, questions with high difficulty levels include items 6, 9, 21, and 24. The figures below show the representation of the difficulty levels of questions by eighth-grade students at public junior high schools in Bima City.

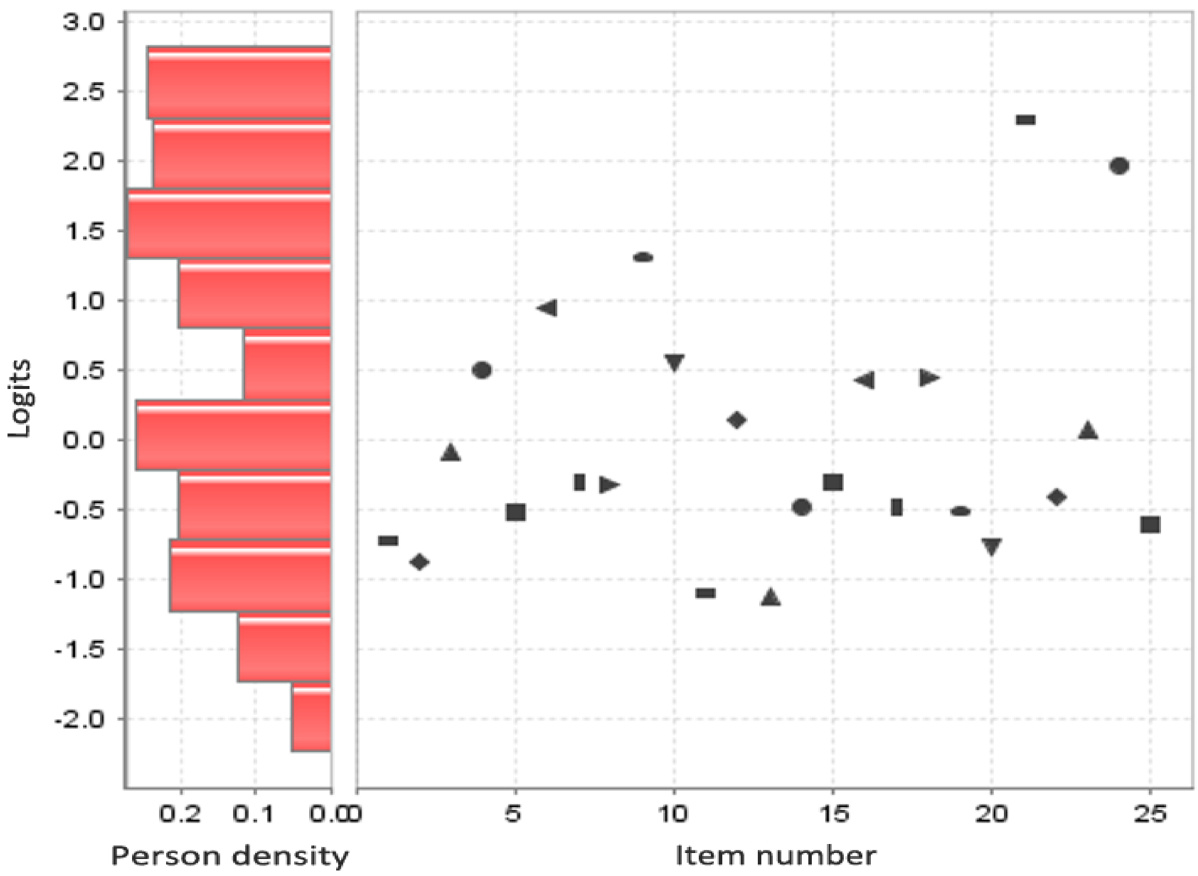

Figure 1. Wright's map of items and individual abilities in math problems integrated with cultural architectural structures.

Figure 1 presents a logit scale depicting the difficulty level of questions and individual abilities in the context of this research. The logit scale is used to provide information about how difficult or easy a question item is and the extent of an individual student's ability to answer the questions. First and foremost, it is important to note that the separation for question items on this logit scale appears quite good. Good separation indicates that this research instrument is capable of distinguishing between students with different levels of ability.

Figure 1 shows a high level of homogeneity in the distribution of mathematics question items and the individual abilities of middle school students. This indicates that the question items in mathematics integrated with cultural architectural structures are evenly distributed in terms of difficulty levels providing accurate data related to students' abilities. The analysis of students' critical thinking abilities through mathematics questions integrated with cultural architectural structures has proven to have a significant impact using the instrument in this research. In other words, there are no question items that are too difficult or too easy. The presence of extreme items can affect the validity of the measurement but in this visualization, the distribution of question items appears balanced and even. Thus, Figure 1 provides a positive visual representation regarding the validity and reliability of this research instrument in measuring students' critical thinking abilities using mathematics questions integrated with cultural architectural structures.

The Rasch model analysis in Table 5 regarding weighted mean square (WMS) or infit and unweighted mean square (UMS) or outfit provides a comprehensive overview of the quality of question items and their fit with the Rasch model. In this context, 16 out of 25 items show WMS values close to 1 indicating excellent item quality, especially in accurately measuring students' abilities. Meanwhile, UMS in jMetrik in this research reflects the accuracy of question items in the Rasch model. Outfit values should range from 0.5 to 1.5. If the outfit value is less than 0.5 or greater than 1.5, it indicates a lack of fit with the Rasch model. From the jMetrik output in Table 5, less satisfactory UMS values are found for item 24, item 21 and item 11.

Furthermore, Table 6 also shows the results of Rasch Person Statistics for 7 selected students based on low and high abilities. From this analysis, focusing particularly on the UMS (outfit) values, it can be identified that some students exhibit significant misfit.

Table 6. Rasch person statistics.

Student_No:	Sum	Vsum	Theta	Stderr	Extreme	WMS	Stdwms	UMS	Stdums
Student No. 28	17.0	17.0	0.83	0.46	No	0.82	-0.78	0.72	-10.73
Student No. 35	23.0	23.0	2.81	0.78	No	0.60	-0.59	0.21	-10.28
Student No. 82	6.0	6.0	-1.34	0.48	No	10.82	0.42	18.06	16.67
Student No.115	8.0	8.0	-0.90	0.45	No	10.73	0.48	11.00	0.41
Student No.130	16.0	16.0	0.62	0.45	No	12.11	10.76	12.13	0.94
Student No.255	9.0	9.0	-0.70	0.44	No	12.29	15.26	12.99	1.07
Student No.259	15.0	15.0	0.42	0.44	No	0.80	-11.52	0.74	-12.89

Rasch model analysis in Table 6 indicates the Weighted Mean Square (WMS or infit) and Unweighted Mean-Square (UMS or outfit) statistics for individuals and questions. The focus is on eight students who show significant misfit. These students can be grouped into two categories based on their UMS (outfit) scores. First, students with high UMS (outfit) scores such as student no. 82 with a total score of 6.0, logit ability of -1.34 and UMS (outfit) of 18.06. Similarly, student no. 130 exhibits significant misfit despite having different total scores and logit abilities.

On the other hand, students with low UMS (outfit) scores like student no. 85 with a total score of 23.0, logit ability of 2.81 and UMS (outfit) of 0.21 show a lower level of misfit. Other students, namely student no. 28, student no: 115 and student no. 259 also display relatively low misfit even though they have variations in total scores.

In a nutshell, Rasch analysis provides a comprehensive view of students who exhibit misfit in this exam. It can be identified that students no. 82, 130 and 255 require special attention while other students show a lower level of misfit by focusing on UMS (outfit) values. Further understanding of the exam and potential difficulties in specific questions can improve the quality of the assessment and enhance students' understanding of the tested material. Next, the output from the jMetrik software in this research is scale quality statistics. This table is crucial in addressing the first and second research questions.

Table 7. Scale quality statistics.

Statistic	Items	Persons
Observed variance	0.75	19.16
Observed std. dev.	0.86	13.84
Mean square error	0.01	0.29
Root MSE	0.13	0.54
Adjusted variance	0.73	16.21
Adjusted std. dev.	0.85	12.73
Separation index	6.43	2.34
Number of strata	8.91	3.46
Reliability	0.97	0.84

According to Table 7, it is illustrated that the variability in item responses is expressed through an observed variance of 0.75 while the observed standard deviation reaches 0.86. This indicates a significant variation in response patterns to mathematics questions integrated with the cultural architectural structure and the high standard deviation suggests an even distribution of item responses. The mean square error for items is 0.01 indicating a level of inaccuracy in measurement. However, with a root mean square error (RMSE) of 0.13, we can observe that the error rate is relatively low indicating a reasonably accurate estimation of item responses.

In the context of adjusted variability, the adjusted variance is 0.73 with a standard deviation of 0.85. This suggests that after adjustment, item responses still have a significant level of variability but the lower standard deviation indicates some control or adjustment to this variability.

The separation index of 6.43 indicates a level of difference among individual abilities. Meanwhile, the number of strata at 8.91 reflects complexity on the scale indicating a variety of discernible ability levels. The item reliability is 0.97 and the person reliability is 0.84 indicating a very high level of reliability for measuring the critical thinking abilities of junior high school students with mathematics questions integrated with the cultural architectural structure. This is evident from the rating scale instrument quality criteria according to Fisher which are in the excellent category, i.e., above 0.94 indicating that this instrument is reliable in measuring the critical thinking abilities of junior high school students with mathematics questions integrated with the cultural architectural structure. This can also be seen in the correlation between theta and the sum of 25 items.

5. Discussion

5.1. Reliability of Mathematics Test Instruments Integrated with Cultural Architectural Structure Using the Rasch Model

The research instrument underwent a validation process by two experts, a mathematics learning expert and a measurement expert. The content validity of the instrument was measured using the Aiken validity index with an overall result at a moderate level (average Aiken validity of 0.76). This value indicates that the overall consistency in assessing the content of each item by raters is good and the instrument is considered valid for measuring the critical thinking abilities of 8th-grade junior high school students.

Furthermore, the calibration results using the Rasch model indicate that the quality of the mathematics test instrument has been effectively measured using the jMetrik software. The analysis results from jMetrik include the difficulty level (item difficulty), WMS (weighted mean square for infit), and UMS (unweighted mean-square for outfit). The Rasch model also focused on the item characteristics located in the meaning of the difficulty level. The instrument's reliability measured with jMetrik reached a high level of 0.97. This indicates that the instrument is consistent and stable in measuring the desired construct. High instrument reliability is a crucial foundation with decisions related to test results. In interpreting this value, we understand that this instrument is reliable and provides consistent information about students' abilities in the context of mathematics. This is consistent with the criteria in Fisher's scale quality instrument criteria stating that an item's measurement reliability (Tarigan, Nilmarito, Islamiyah, Darmana, & Suyanti, 2022).

It is concluded that some questions are valid but have low reliability based on the research findings that focused on the reliability of mathematics questions from previous studies (Hamimi, Zamharirah, & Rusydy, 2020). Similar research shows a disproportion in the difficulty level of questions though their reliability is quite good (Susanto, Rinaldi, & Novalia, 2015). In the context of creating questions by teachers, the quality of multiple-choice tests indicates that all test items are valid with a high level of reliability (Brown & Abdulnabi, 2017; DiBattista & Kurzawa, 2011; Gierl, Bulut, Guo, & Zhang, 2017). Meanwhile, Supandi and Farikhah (2016) found that most items are valid with high reliability and varying difficulty levels. The analysis of test items in mathematics competitions can help identify invalid questions (Karim, 2018). On the contrary, Tezer and Özcan (2015) found that the reliability of the scale is a significant factor in factor analysis.

5.2. Separation Index of Mathematics Test Instruments Integrated with Cultural Architectural Structure Using the Rasch Model

The analysis of separation for test items on the logit scale showed quite good results. A good separation indicates that this research instrument can differentiate between students with different ability levels. The visualization of the logit scale shows a good level of homogeneity in the distribution of test items and individual abilities. This indicates that the test items cover various levels of difficulty uniformly. There are no extreme items (too difficult or too easy). The results of the Rasch Person statistics analysis for 7 students indicated that there were students showing significant discrepancies and a deeper understanding of difficulties in specific questions could help improve the quality of the evaluation and students' understanding of the tested material. A large item separation index indicates that respondents have diverse abilities (Fisher, 2007).

The research results are consistent with previous researchers' findings indicating that an item validity of 0.93 shows that the questions can measure critical thinking abilities as supported by a separation index of 4.34 (Nuryanti, Masykuri, & Susilowati, 2018). The separation value indicates that these test items have a good response distribution. Consistency within groups of individuals in providing information about the difficulty of items in forming a scale is reflected in the item separation index (Curtis & Boman, 2007). The higher the separation index estimate, the more accurately the analysis of overall item separation aligns with the model used (Hamimi et al., 2020).

In practical terms, the separation index has significant implications particularly in decision-making such as classification or selection. Information from the separation index can be applied directly in the field with the instrument's ability to separate individuals with a high level of clarity. A deep understanding of the variation in respondents' abilities can help decision-makers identify and respond to specific educational or training needs.

5.3. Scale Quality Statistics

The output from the jMetrik software shows scale quality statistics in Table 7. Variability in item responses is expressed through an observed variance of 0.75 with a high standard deviation indicating an even distribution of item responses. Although the level of measurement inaccuracy (Mean Square Error) is 0.01, the relatively low Root MSE value (0.13) indicates good accuracy in estimating item responses. The adjusted variability has a variance of 0.73 with a standard deviation of 0.85 showing that item responses still have a significant level of variation. A high separation index (6.43) indicates a good level of difference between individual abilities. The very high reliability level (0.97) indicates that this instrument is reliable in measuring the critical thinking abilities of 8th-grade junior high school students with mathematics questions integrated with a cultural architectural structure.

6. Conclusion

This study concludes that the mathematics test instrument integrated with a cultural architectural structure using the Rasch model has undergone adequate validation by mathematics and measurement experts. The content validity results with the Aiken validity index reached a moderate level indicating good consistency in the content assessment of each item by raters. The Rasch model calibration results at a high instrument reliability level were 0.97. Furthermore, the separation index analysis indicates that the mathematics test instrument integrated with the cultural architecture structure has been able to distinguish students' abilities with different levels and individual abilities. Additionally, the scale quality statistics also indicate good variation in item responses, low error rates and a high separation index.

Based on the results of the study on mathematics test instruments integrated with cultural architecture structure, recommendations for future researchers prioritize enhancing validity and reliability and focus on Differential Item Functioning (DIF) analysis. In terms of practice, attention should be paid to teachers' involvement in developing the instrument as its technical implementers in schools so that its relevance to the curriculum can trigger quality learning achievement. Another aspect is the need for further researcher support regarding the tested and developed instruments especially in the implementation process.

References

Adams, W. K., & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education, 33*(9), 1289–1312. https://doi.org/https://doi.org/10.1080/09500693.2010.512369

Aizikovitsh-Udi, E., & Cheng, D. (2015). Developing critical thinking skills from dispositions to abilities: Mathematics education from early childhood to high school. *Creative Education, 6*(4), 455-462. https://doi.org/https://doi.org/10.4236/ce.2015.64045

Aksu, G., Guzeller, C. E. M., & Eser, M. (2019). Jmetrik: Classical test theory and item response theory data analysis software. *Journal of Measurement and Evaluation in Education and Psychology, 10*(2), 165-178.

Andrich, D., & Marais, I. (2019). *A course in rasch measurement theory in springer texts in education*. Singapore: Springer Nature Singapore.

Andrich, D., Marais, I., & Humphry, S. (2012). Using a theorem by Andersen and the dichotomous rasch model to assess the presence of random guessing in multiple choice items. *Journal of Educational and Behavioral Statistics, 37*(3), 417–442. https://doi.org/https://doi.org/10.3102/1076998611411914

Asfar, A. M. I. T., Asfar, A. M. I. A., & Nurannisa, A. (2021). Integration of local traditions bugis-makassarese: Learning strategies to improve mathematical communication skills. *Journal of Physics: Conference Series, 1808*(1), 12064. https://doi.org/10.1088/1742-6596/1808/1/012064

Avetisyan, V. (2015). Survey of software for the test quality analysis. *International Journal" Information Content and Processing, 2*(1), 82–92.

Baghaei, P. (2012). The application of multidimensional rasch models in large scale assessment and validation: An empirical example. *Electronic Journal of Research in Educational Psychology, 10*(1), 233-252.

Bansilal, S. (2015). A rasch analysis of a grade 12 test written by mathematics teachers. *South African Journal of Science, 11*(5-6), 1–9. https://doi.org/https://doi.org/10.17159/sajs.2015/20140098

Bolondi, G., Cascella, C., & Giberti, C. (2017). Highlights on gender gap from Italian standardized assessment in mathematics diversity in mathematics education. *Proceedings of the International Symposium Elementary Maths Teaching, 17*, 81–90. https://doi.org/10.1016/j.econedurev.2017.03.001

Bond, T. G., & Fox, C. M. (2013). *Applying the rasch model: Fundamental measurement in the human sciences.* New York: Psychology Press.

Bray, A., Girvan, C., & Chorcora, E. N. (2023). Students' perceptions of pedagogy for 21st century learning instrument (S-POP-21): Concept, validation, and initial results. *Thinking Skills and Creativity, 49*, 101319. https://doi.org/https://doi.org/10.1016/j.tsc.2023.101319

Brown, G. T., & Abdulnabi, H. H. (2017). Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades. *Frontiers, 2*(24), 1-12. https://doi.org/https://doi.org/10.3389/feduc.2017.00024

Chan, S. W., Ismail, Z., & Sumintono, B. (2014). A rasch model analysis on secondary students' statistical reasoning ability in descriptive statistics. *Procedia-Social and Behavioral Sciences, 129*, 133–139. https://doi.org/https://doi.org/10.1016/j.sbspro.2014.03.658

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the rasch model: The research-based early maths assessment. *Educational Psychology, 28*(4), 457–482. https://doi.org/10.1080/01443410701777272

Curtis, D. D., & Boman, P. (2007). X-ray your data with rasch. *International Education Journal, 8*(2), 249–259.

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning, 2*(2), 4. https://doi.org/10.5206/cjsotl-rcacea.2011.2.4

Doyle, P. J., Hula, W. D., McNeil, M. R., Mikolic, J. M., & Matthews, C. (2005). An application of rasch analysis to the measurement of communicative functioning. https://doi.org/10.1044/1092-4388(2005/098)

Edelsbrunner, P. A., & Dablander, F. (2018). The psychometric modeling of scientific reasoning: A review and recommendations for future avenues. *Educational Psychology Review, 31*(1), 1–34. https://doi.org/10.1007/s10648-018-9455-5

Embretson, S. E., & Reise, P. E. (2000). *Item response theory for psychologists.* Mahwah: Lawrence Earlbaum Associates.

Engelhard Jr, G. (2013). *Invariant measurement: Using rasch models in the social, behavioral, and health sciences.* New York: Routledge.

Erfan, M., Maulyda, M. A., Ermiana, I., Hidayati, V. R., & Widodo, A. (2020). Validity and reliability of cognitive tests study and development of elementary curriculum using rasch model. *Psychology, Evaluation, and Technology in Educational Research, 3*(1), 26–33. https://doi.org/https://doi.org/10.33292/petier.v3i1.51

Fauzi, L. M., Hanum, F., Jailani, J., & Jatmiko, J. (2022). Ethnomathematics: Mathematical ideas and educational values on the architecture of Sasak traditional residence. *International Journal of Evaluation and Research in Education, 11*(1), 250-259. https://doi.org/10.11591/ijere.v11i1.21775

Fisher, W. P. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions, 21*(1), 1095.

Fouze, A. Q., & Amit, M. (2017). Development of mathematical thinking through integration of ethnomathematic folklore game in math instruction. *Eurasia Journal of Mathematics, Science and Technology Education, 14*(2), 617-630. https://doi.org/10.12973/ejmste/80626

Garcia, I., & Pacheco, C. (2013). A constructivist computational platform to support mathematics education in elementary school. *Computers & Education, 66*, 25-39. https://doi.org/10.1016/j.compedu.2013.02.004

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*(6), 1082-1116. https://doi.org/https://doi.org/10.3102/0034654317726529

Gorin, J. S., Embretson, S. E., & McKay, D. (2008). Item response theory and Rasch models. *Handbook of Research Methods in Abnormal and Clinical Psychology*, 271-292.

Hamimi, L., Zamharirah, R., & Rusydy, R. (2020). Analysis of class VII mathematics exam questions for the odd semester of the 2017/2018 academic year. *Mathema: Journal of Mathematics Education, 2*(1), 57–66.

Harjo, B., Kartowagiran, B., & Mahmudi, A. (2019). Development of critical thinking skill instruments on mathematical learning high school. *International Journal of Instruction, 12*(4), 149-166. https://doi.org/10.29333/iji.2019.12410a

Hernawati, A., & Nurbayani, S. (2018). *The importance of critical thinking to face global challenges in the era of industry 4.0 through social studies.* Paper presented at the 3rd International Seminar on Social Studies and History Education (ISSSHE Bandung, Indonesia).

Holland, P. W. (1990). On the sampling theory roundations of item response theory models. *Psychometrika, 55*, 577-601.

Jones, A. (2015). *A disciplined approach to critical thinking. In The Palgrave handbook of critical thinking in higher education.* New York: Palgrave Macmillan. https://doi.org/10.1057/9781137378057_11.

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

Karim, A. (2018). Analysis of the quality of mathematics competition questions at the high school level. *Titian Ilmu: Jurnal Ilmiah Multi Sciences, 10*(1), 1–8. https://doi.org/https://doi.org/10.30599/jti.v10i1.126

Karlimah, K., Andriani, D., & Suryana, D. (2020). Development of mathematical anxiety instruments with a Rasch model analysis. *The Open Psychology Journal, 13*(1), 181-192. https://doi.org/10.2174/1874350102013010181

Kaspersen, E., Pepin, B., & Sikko, S. A. (2017). Measuring student teachers' practices and beliefs about teaching mathematics using the Rasch model. *International Journal of Research & Method in Education, 40*(4), 421-442. https://doi.org/https://doi.org/10.1080/1743727x.2016.1152468

Kivunja, C. (2014). Do you want your students to be job-ready with 21st century skills? Change pedagogies: A pedagogical paradigm shift from vygotskyian social constructivism to critical thinking, problem solving and siemens' digital connectivism. *International Journal of Higher Education, 3*(3), 81–91. https://doi.org/https://doi.org/10.5430/ijhe.v3n3p81

Kubinger, K. D., Rasch, D., & Yanagida, T. (2011). A new approach for testing the Rasch model. *Educational Research and Evaluation, 17*(5), 321-333. https://doi.org/10.1037/e578442014-092

Kurniawan, H., Purwoko, R. Y., & Setiana, D. S. (2023). Integrating cultural artifacts and tradition from remote regions in developing mathematics lesson plans to enhance mathematical literacy. *Journal of Pedagogical Research, 8*(1), 61-74.

Leeming, P., & Wong, A. (2016). Using dictation to measure language proficiency: A Rasch analysis. *Language Testing and Assessment, 5*(2), 1-25. https://doi.org/10.58379/mbsw8958

Lidinillah, D. A. M., Rahman, R., Wahyudin, W., & Aryanto, S. (2022). Integrating sundanese ethnomathematics into mathematics curriculum and teaching: A systematic review from 2013 to 2020. *Infinity Journal, 11*(1), 33-54. https://doi.org/10.22460/infinity.v11i1.p33-54

Long, C., Bansilal, S., & Debba, R. (2014). An investigation of mathematical literacy assessment supported by an application of Rasch measurement. *Pythagoras, 35*(1), 1-17. https://doi.org/https://doi.org/10.4102/pythagoras.v35i1.235

Mahdi, O. R., Nassar, I. A., & Almuslamani, H. A. I. (2020). The role of using case studies method in improving students' critical thinking skills in higher education. *International Journal of Higher Education, 9*(2), 297-308. https://doi.org/https://doi.org/10.5430/ijhe.v9n2p297

Makransky, G., Rogers, M. E., & Creed, P. A. (2015). Analysis of the construct validity and measurement invariance of the career decision self-efficacy scale: A Rasch model approach. *Journal of Career Assessment, 23*(4), 645-660. https://doi.org/10.1177/1069072714553555

Meaney, T., Trinick, T., & Allen, P. (2021). Ethnomathematics in education: The need for cultural symmetry. In Handbook of Cognitive Mathematics. In (pp. 1–29): Springer International Publishing. https://doi.org/10.1007/978-3-030-44982-7_4-1.

Monrat, N., Phaksunchai, M., & Chonchaiya, R. (2022). Developing students' mathematical critical thinking skills using open-ended questions and activities based on student learning preferences. *Education Research International*, *2022*, 1–11. https://doi.org/10.1155/2022/3300363

Mui Lim, S., Rodger, S., & Brown, T. (2009). Using Rasch analysis to establish the construct validity of rehabilitation assessment tools. *International Journal of Therapy and Rehabilitation*, *16*(5), 251-260. https://doi.org/10.12968/ijtr.2009.16.5.42102

Munthahana, J., & Budiarto, M. T. (2020). Ethnomathematics exploration in panataran temple and its implementation in learning. *Indonesian Journal of Science and Mathematics Education*, *3*(2), 196-209. https://doi.org/10.24042/ijsme.v3i2.6718

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

Noben, I., Maulana, R., Deinum, J. F., & Hofman, W. A. (2021). Measuring university teachers' teaching quality: A Rasch modelling approach. *Learning Environments Research*, *24*, 87-107. https://doi.org/10.1007/s10984-020-09319-w

Nuraini, L. (2022). Integration of local wisdom values in mathematics learning for elementary schools (SD/MI) based on the 2013 curriculum. *Journal of Mathematics Education (Holy)*, *1*(2). http://dx.doi.org/10.21043/jpm.v1i2.4873

Nuryanti, S., Masykuri, M., & Susilowati, E. (2018). Item analysis and Rasch model in developing critical thinking ability instruments for vocational high school students. *Journal of Science Education Innovation*, *4*(2), 224–233.

Parker, F., Bartell, T. G., & Novak, J. D. (2017). Developing culturally responsive mathematics teachers: Secondary teachers' evolving conceptions of knowing students. *Journal of Mathematics Teacher Education*, *20*(4), 385-407. https://doi.org/10.1007/s10857-015-9328-5

Petra, T. Z. H. T., & Aziz, M. J. A. (2020). Investigating reliability and validity of student performance assessment in higher education using Rasch model. *Journal of Physics: Conference Series*, *1529*(4), 42088. https://doi.org/10.1088/1742-6596/1529/4/042088

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), 020111. https://doi.org/10.1103/physrevphyseducres.15.020111

Prasad Pant, B., & Chandra Luitel, B. (2020). Incorporating culturally relevant pedagogy in teaching mathematics in a blended learning environment. In Innovative Technologies and Pedagogical Shifts in Nepalese Higher Education. In (pp. 167–181): BRILL. https://doi.org/10.1163/9789004448865_010.

Priatna, N., Lorenzia, S., & Widodo, S. A. (2020). STEM education at junior high school mathematics course for improving the mathematical critical thinking skills. *Journal for the Education of Gifted Young Scientists*, *8*(3), 1173-1184. https://doi.org/10.17478/JEGYS.728209

Raj, T., Chauhan, P., Mehrotra, R., & Sharma, M. (2022). Importance of critical thinking in the education. *World Journal of English Language*, *12*(3), 126-133. https://doi.org/https://doi.org/10.5430/wjel.v12n3p126

Razak, N. b. A., bin Khairani, A. Z., & Thien, L. M. (2012). Examining quality of mathemtics test items using rasch model: Preminarily analysis. *Procedia-Social and Behavioral Sciences*, *69*, 2205-2214. https://doi.org/10.1016/j.sbspro.2012.12.187

Resnick, B. (2005). Reliability and validity of the outcome expectations for exercise scale-2. *Journal of Aging and Physical Activity*, *13*(4), 382-394. https://doi.org/10.1123/japa.13.4.382

Retnawati, H. (2016). *Reliability validity and item characteristics*. Yogyakarta: Parama Publishing.

Reynolds, C. R., Livingston, R. B., Willson, V. L., & Willson, V. (2010). *Measurement and assessment in education*. Upper Saddle River: Pearson Education International.

Ridzuan, M. F., Lian, L. H., Fozee, F. A. A., & Nasser, S. (2020). Rasch analysis model: Reliability and validity of superitem test instrument. *International Journal of Academic Research in Progressive Education and Development*, *9*(4), 1-11. https://doi.org/10.6007/ijarped/v9-i4/8166

Saidi, S. S., & Siew, N. M. (2019). Reliability and validity analysis of statistical reasoning test survey instrument using the Rasch measurement model. *International Electronic Journal of Mathematics Education*, *14*(3), 535-546.

Sari, N., & Abdurrahman, S. (2019). Developing and validating of the three tier diagnostic test based'higher order thinking skills' instrument. *Dynamics of the Scientific Journal of Basic Education*, *11*(2), 87-93.

Schneider, L., Strobl, C., Zeileis, A., & Debelak, R. (2022). An R toolbox for score-based measurement invariance tests in IRT models. *Behavior Research Methods*, *54*(5), 2101-2113. https://doi.org/10.31234/osf.io/r9w34

Schulz, A. (2023). Assessing student teachers' procedural fluency and strategic competence in operating and mathematizing with natural and rational numbers. *Journal of Mathematics Teacher Education*, 1-28. https://doi.org/https://doi.org/10.1007/s10857-023-09590-7

Simamora, R. E., & Saragih, S. (2019). Improving students' mathematical problem solving ability and self-efficacy through guided discovery learning in local culture context. *International Electronic Journal of Mathematics Education*, *14*(1), 61-72. https://doi.org/https://doi.org/10.12973/iejme/3966

Sinnema, C., Ludlow, L., & Robinson, V. (2016). Educational leadership effectiveness: A Rasch analysis. *Journal of Educational Administration*, *54*(3), 305-339.

Sumintono, B., & Widhiarso, W. (2015). *Application of Rasch modeling in educational assessment*. Cimahi: Trim Komunikata Publishing House.

Supandi, S., & Farikhah, L. (2016). Analisis butir soal matematika pada instrumen uji coba materi segitiga. *JIPMat*, *1*(1), 90880. https://doi.org/https://doi.org/10.26877/jipmat.v1i1.1085

Susanto, H., Rinaldi, A., & Novalia, N. (2015). Analysis of the validity and reliability of the level of difficulty and different power in the odd semester final exam questions in mathematics class XII Ips at SMA Negeri 12 Bandar Lampung in the 2014/2015 academic year. *Al-Jabar: Journal of Mathematics Education*, *6*(2), 203–218. https://doi.org/https://doi.org/10.24042/ajpm.v6i2.50

Takács, R., Kárász, J. T., Takács, S., Horváth, Z., & Oláh, A. (2021). Applying the Rasch model to analyze the effectiveness of education reform in order to decrease computer science students' dropout. *Humanities and Social Sciences Communications*, *8*(1), 1-8. https://doi.org/10.1057/s41599-021-00725-w

Tarigan, E. F., Nilmarito, S., Islamiyah, K., Darmana, A., & Suyanti, R. D. (2022). Analysis of test instruments using Rasch model and SPSS 22.0 software (analysis of test instruments using Rasch model and SPSS 22.0 software). *Journal of Chemical Education Innovation*, *16*(2), 92–96. https://doi.org/https://doi.org/10.15294/jipk.v16i2.30530

Tezer, M., & Özcan, D. (2015). A study of the validity and reliability of a mathematics lesson attitude scale and student attitudes. *Eurasia Journal of Mathematics, Science and Technology Education*, *11*(2), 371-379.

Thomas, G., Anderson, D., & Nashon, S. (2008). Development of an instrument designed to investigate elements of science students' metacognition, self-efficacy and learning processes: The SEMLI-S. *International Journal of Science Education*, *30*(13), 1701-1724.

Van der Lans, R. M., Van de Grift, W. J., & van Veen, K. (2018). Developing an instrument for teacher feedback: Using the rasch model to explore teachers' development of effective teaching strategies and behaviors. *The Journal of Experimental Education*, *86*(2), 247-264. https://doi.org/10.1080/00220973.2016.1268086

Widana, I. W., Parwata, I., & Sukendra, I. K. (2018). Higher order thinking skills assessment towards critical thinking on mathematics lesson. *International Journal of Social Sciences and Humanities*, *2*(1), 24-32. https://doi.org/https://doi.org/10.29332/ijssh.v2n1.74

Wong, S. L., & Wong, S. L. (2021). Effects of motivational adaptive instruction on student motivation towards mathematics in a technology-enhanced learning classroom. *Contemporary Educational Technology*, *13*(4), ep326. https://doi.org/10.30935/cedtech/11199

Zile-Tamsen, V. C. (2017). Using Rasch analysis to inform rating scale development. *Research in Higher Education*, *58*(8), 922-933.

Zubaidah, S., & Arsih, F. (2021). *Indonesian culture as a means to study science*. Paper presented at the The 4th International Conference on Mathematics and Science Education (ICoMSE) 2020: Innovative Research in Science and Mathematics Education in The Disruptive Era AIP Publishing.