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

Muh. Fitrah1, Anastasia Sofroniou2, Ofianto3, Lusdi Judijanto4, and Widihastuti1

1 Department of Education Research and Evaluation, Graduate School, Yogyakarta State University, Yogyakarta, Indonesia. Email: mfitrah@2023@student.eway.ac.id
2 School of Computing and Engineering, University of West London, London, UK. Email: anastasia.sofroniou@uel.ac.uk
3 Padang State University, Padang, Indonesia. Email: ofianto@fs.unp.ac.id
4 POSIS-Jakarta, Jakarta, Indonesia. Email: lusdi.judijanto@uny.ac.id

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.

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: 01821/BPPT/BPL/06/2023)

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. R/2023/1/UN34.17/RM/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.


History:
Received: 23 February 2023
Revised: 29 May 2023
Accepted: 12 June 2024
Published: 31 July 2024

Licensed: This work is licensed under a Creative Commons Attribution 4.0 License [CCBY 4.0].

Publisher: Asian Online Journal Publishing Group

© 2024 by the authors; licensee Asian Online Journal Publishing Group
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, 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, 2025; 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; Karlmaia, 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, Mikolje, & 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; Kuhinger, 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 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 & Widhiarsi, 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, & Sikk, 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, & Aryanoto, 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 5.29. Besnick (2005) describes a separation index of 4.70 and 9.88 in measurement tool development. Additionally, Sari and Abdurrahman (2010) achieved a separation index of 3.19 in test product development.

Various studies using the Rasch model to analyze measurement instrument quality have provided 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). Riza, Zain, Fozee, and Nasser (2020) explored reliability and validity using a superitem test while Saldi 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 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 currently limited. The advantage 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, Jaiiani, & 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 as well as 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 & Budiarso, 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 & Arshih, 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, & Nurunnisa, 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, 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>
<thead>
<tr>
<th>School</th>
<th>Average school exam score in mathematics</th>
<th>Sample</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public Junior High School 7 Bima City</td>
<td>36.58</td>
<td>50</td>
<td>High</td>
</tr>
<tr>
<td>Public Junior High School 1 Bima City</td>
<td>36.51</td>
<td>68</td>
<td>High</td>
</tr>
<tr>
<td>Public Junior High School 6 Bima City</td>
<td>35.62</td>
<td>54</td>
<td>Medium</td>
</tr>
<tr>
<td>Public Junior High School 2 Bima City</td>
<td>35.28</td>
<td>60</td>
<td>Medium</td>
</tr>
<tr>
<td>Public Junior High School 4 Bima City</td>
<td>34.50</td>
<td>65</td>
<td>Low</td>
</tr>
<tr>
<td>Public Junior High School 11 Bima City</td>
<td>34.51</td>
<td>60</td>
<td>Low</td>
</tr>
</tbody>
</table>

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 curriculum framework of junior high school mathematics set by the standard, curriculum and educational assessment body in a standardized manner (see Table 2).
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 (Alkso, 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 assessment. 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).

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

Explanation
\[ V \]: Index of expert agreement on item validity.
\[ \sum_{i=1}^{c} s_i \]: 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 the Rasch model 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 2. Matrix of mathematics questions integrated with cultural architectural structures.

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

### Table 3. Rating scale instrument quality criteria.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Poor</th>
<th>Fair</th>
<th>Good</th>
<th>Very good</th>
<th>Excellent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item model fit mean-square range extremes</td>
<td>$&lt; 0.33$ or $&gt;3.0$</td>
<td>$0.34 - 2.9$</td>
<td>$0.5 - 2.0$</td>
<td>$0.71 - 1.4$</td>
<td>$0.77 - 1.3$</td>
</tr>
<tr>
<td>Person and item measurement reliability</td>
<td>$&lt;0.67$</td>
<td>$0.67-0.80$</td>
<td>$0.81-0.90$</td>
<td>$0.91-0.94$</td>
<td>$&gt;0.94$</td>
</tr>
<tr>
<td>Person and item strata separated</td>
<td>2 or less</td>
<td>2-3</td>
<td>3-4</td>
<td>4-5</td>
<td>$&gt;5$</td>
</tr>
<tr>
<td>Ceiling effect: % maximum extreme scores</td>
<td>$&gt;5%$</td>
<td>$2-5%$</td>
<td>$1-2%$</td>
<td>$0.5-1%$</td>
<td>$&lt;0.5%$</td>
</tr>
<tr>
<td>Floor effect: % minimum extreme scores</td>
<td>$&gt;5%$</td>
<td>$2-5%$</td>
<td>$1-2%$</td>
<td>$0.5-1%$</td>
<td>$&lt;0.5%$</td>
</tr>
</tbody>
</table>
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.

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 4. Content validity with the Aiken validity index.

<table>
<thead>
<tr>
<th>Items</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>s1</th>
<th>s2</th>
<th>s3</th>
<th>s4</th>
<th>Σs</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>9</td>
<td>0.75</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>9</td>
<td>0.75</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>0.83</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>0.75</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>0.50</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>10</td>
<td>0.83</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>0.83</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>10</td>
<td>0.83</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>11</td>
<td>0.92</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>0.83</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>11</td>
<td>0.92</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>10</td>
<td>0.83</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>0.50</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>9</td>
<td>0.75</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0.67</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>11</td>
<td>0.92</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>10</td>
<td>0.83</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>9</td>
<td>0.75</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>0.50</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>0.42</td>
</tr>
<tr>
<td>22</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>12</td>
<td>1.00</td>
</tr>
<tr>
<td>23</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>11</td>
<td>0.92</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>0.83</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>9</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Average: 0.76

Note: The column $R_s (s-1)$ 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.

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

<table>
<thead>
<tr>
<th>Items</th>
<th>Difficulty</th>
<th>Std. error</th>
<th>WMS</th>
<th>UMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>item1</td>
<td>-0.73</td>
<td>0.14</td>
<td>0.88</td>
<td>0.68</td>
</tr>
<tr>
<td>item2</td>
<td>-0.87</td>
<td>0.15</td>
<td>1.53</td>
<td>0.55</td>
</tr>
<tr>
<td>item3</td>
<td>-0.10</td>
<td>0.13</td>
<td>1.21</td>
<td>1.20</td>
</tr>
<tr>
<td>item4</td>
<td>0.50</td>
<td>0.13</td>
<td>0.75</td>
<td>0.65</td>
</tr>
<tr>
<td>item5</td>
<td>-0.51</td>
<td>0.13</td>
<td>1.42</td>
<td>1.39</td>
</tr>
<tr>
<td>item6</td>
<td>0.95</td>
<td>0.13</td>
<td>0.75</td>
<td>0.83</td>
</tr>
<tr>
<td>item7</td>
<td>-0.31</td>
<td>0.13</td>
<td>0.86</td>
<td>0.75</td>
</tr>
<tr>
<td>item8</td>
<td>-0.92</td>
<td>0.13</td>
<td>1.29</td>
<td>1.29</td>
</tr>
<tr>
<td>item9</td>
<td>1.31</td>
<td>0.15</td>
<td>2.17</td>
<td>1.42</td>
</tr>
<tr>
<td>item10</td>
<td>0.54</td>
<td>0.13</td>
<td>0.74</td>
<td>0.65</td>
</tr>
<tr>
<td>item11</td>
<td>-1.09</td>
<td>0.14</td>
<td>0.66</td>
<td>0.44</td>
</tr>
<tr>
<td>item12</td>
<td>0.14</td>
<td>0.13</td>
<td>0.78</td>
<td>0.65</td>
</tr>
<tr>
<td>item13</td>
<td>-1.13</td>
<td>0.15</td>
<td>0.88</td>
<td>0.95</td>
</tr>
<tr>
<td>item14</td>
<td>-0.48</td>
<td>0.13</td>
<td>0.78</td>
<td>0.63</td>
</tr>
<tr>
<td>item15</td>
<td>-0.31</td>
<td>0.13</td>
<td>0.86</td>
<td>0.68</td>
</tr>
<tr>
<td>item16</td>
<td>0.42</td>
<td>0.13</td>
<td>1.39</td>
<td>1.34</td>
</tr>
<tr>
<td>item17</td>
<td>-0.48</td>
<td>0.13</td>
<td>0.79</td>
<td>0.65</td>
</tr>
</tbody>
</table>
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.

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.
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.

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 9th-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 focused on the reliability of mathematics questions from previous studies (Hamimi, Zanharirah, & Rusdy, 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, Balut, Guo, & Zhang, 2017). Meanwhile, Supandi and Farikiah (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 0.83 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 architectural 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


Asian Online Journal Publishing Group is not responsible or answerable for any loss, damage or liability, etc. caused in relation to/ arising out of the use of the content. Any queries should be directed to the corresponding author of the article.