EduMa: Mathematics Education Learning And Teaching
December 2024, Vol 13 No 2 Page 131 – 144
https://syekhnurjati.ac.id/jurnal/index.php/eduma/eduma/article/view/18136
p-ISSN: 2086-3918, e-ISSN: 2502-5209
Published by Tadris Matematika IAIN Syekh Nurjati cirebon

EduMa
MATHEMATICS EDUCATION LEARNING AND TEACHING

Rasch Model Analysis: Validity and Reliability of Context-Based Geometry Performance Assessment Instruments Jakarta Cultural Heritage

Bayu Gunawan [1*], Wardani Rahayu [2], Tian Abdul Aziz[3]

1, 2, 3, Mathematics Education, Universitas Negeri Jakarta
*Corresponding author: Jln.Rawamangun Muka Pulogabung Jakarta, Indonesia;
e-mail addresses: bayugwan@gmail.com

article info

How to cite this article:

Gunawan, B., Rahayu, W., & Aziz, T.A. (2024). Rasch Model Analysis: Validity and Reliability of Context-Based Geometry Performance Assessment Instruments Jakarta Cultural Heritage. *Eduma: Mathematics Education Learning And Teaching*, 13(2), 131 - 144. doi:http://dx.doi.org/10.24235/eduma.v13i2.18136

Article history:

Received: 07 06, 2024

Accepted: 12 10, 2024

Published: 12, 2024

abstract

The objective of this study is to examine the validity and reliability of performance evaluation rubrics when applied to Jakarta's cultural heritage structures, specifically focusing on the geometrical material. The Rasch Model, namely the Winstep version 3.73, will be utilized for analysis. The Rasch model offers benefits due to its ability to provide more precise estimations and identify flaws within the model. The rubric-based performance evaluation instrument is utilized to evaluate students' problem-solving abilities, with each problem being assigned a distinct rubric. A total of 381 students were sampled from two distinct schools located in North Jakarta and East Jakarta. The study's results indicate that the Cronbach's alpha value attained is 0.96, which is classified as excellent. The individual's dependability score is 0.91, which falls inside the excellent range. Similarly, the item's reliability score is 0.94, also classified as excellent. The item fit test conducted on the instrument indicates that there is no need to reject any items. The components in this performance evaluation instrument have strong validity and reliability, rendering it appropriate for evaluating student performance.

Keywords:
Rasch Model; Performance Assessment; Geometry, Problem-Solving Abilities

INTRODUCTION

The assessment process is the most important area of learning. Assessment can be interpreted as a series of activities to obtain information from the beginning to the end of learning to evaluate and diagnose what needs to be improved to achieve learning objectives so that teachers and students are able to plan review and then apply the steps that must be taken (Purnomo, 2013). The role of teachers as educators needs to realize that educational progress is very dependent on the creativity and dedication of teachers, especially understanding that their students will enter an era full of competition, there should be an effort to always improve their teaching methods. Evaluation tools or assessment methods in learning are needed to support improvements in teaching methods. Then the results of the evaluation can be used as a design and implementation of good mathematics learning activities in schools. Assessment or evaluation of these students can be used to determine the quality of students, it can also be used to measure the psychomotor domain of students which is directed through the stages of student performance (Ratumanan & Laurens, 2016).

Based on the results of an interview with one of the teachers at junior high school in East Jakarta, it was found that some students lack knowledge of basic mathematical concepts, in fact, knowledge of these mathematical concepts is very much needed in everyday life. This can be caused by a lack of student learning motivation, students not being very actively involved in learning activities, and conventional learning models still being applied in learning activities, as well as in the assessment aspect. One effective step that a teacher can use to increase insight and improve performance is to participate in the Subject Teacher Deliberation program (MGMP). MGMP activities aim to discuss problems experienced in the learning process or activities, and then find joint solutions to overcome them. Experts also make efforts to overcome education through the development of student performance assessments.

Previous research developed an instrument to evaluate student performance in orthographic projection drawing practicals. This indicates that the results of data analysis using the developed instrument are more accurate and reliable, as well as consistent in relevant categories. Most students are competent based on the available performance assessments (Anis, Rohendi, & Sukandar, 2019). Then the research on developing Project-Based Psychomotor Assessment Instruments to Improve Students' Psychomotor Competence in Blended Learning during the New Normal Era. The research results show that the developed project-based psychomotor assessment instruments can be used in blended learning methods because these instruments are an urgent need for teachers as an alternative to measuring students' skills. Therefore, it is hoped that the availability of these project-based psychomotor assessment instruments can provide options for teachers in designing and conducting psychomotor assessments both online and offline (blended learning) (Ningsih & Rahayu, 2021).

Many students in Indonesia, a cosmopolitan country, exhibit a lack of care for traditional values in this age of technology. The development of culture-based performance evaluation tools is thought to be a way to solve this problem and improve the younger generation's understanding of their culture (Khoriyah & Oktiningrum, 2021). The instrument's integration of cultural elements, including Jakarta's heritage buildings, will expose pupils to novel stimuli they have never experienced before (Alfiatin & Oktiningrum, 2019).

The Rasch Model approach was used by the researchers in contrast to earlier instrument development studies that employed conventional test theory. The ability to identify model errors and provide more precise estimates are two benefits of this approach (Taufiq, Yudha, Md, & Suryana, 2021). Because Rasch Model analysis evaluates measurement items using a probabilistic approach, it is non-deterministic and may more precisely identify the measured objects (Indihadi, Suryana, & Ahmad, 2022). Measurement using the Rasch Model can describe the interaction between respondents and test items (Putra, Hermita, & Alim, 2021). Rasch Model analysis is used to examine the validity of the instrument. The quality of the instrument in the Rasch Model analysis is measured through several aspects, namely unidimensionality, Wright map analysis, item analysis, participant ability analysis, and instrument analysis (Muslihin, Suryana, Ahman, Suherman, & Dahlan, 2022).

LITERATURE REVIEW

Performance assessment techniques involve student activities, such as observing or performing tasks, to assess the achievement of competencies that require performing specific tasks (Kartikasari & Rahayu, 2018). Performance assessment instruments require students to use their knowledge and skills in various disciplines to demonstrate proficiency and perform tasks. Performance assessment requires students to handle or respond to certain tasks (Brookhart, 2015). Performance assessment can also be defined as an activity in which students respond to a concept, endeavor to create a product, or conduct a demonstration (Oberg, 2010). In assessing performance, it is essential for teachers to understand the assessment criteria in order to provide accurate and valid assessments (Panadero & Jonsson, 2013). Guidelines for student performance assessment can utilize rubrics as the primary tool to enhance the reliability, validity, and transparency of the assessment. Rubrics are very helpful because they provide assessment criteria in a structured format. Rubrics serve as more than just a tool to assist assessors in creating summative assessments. Teachers also use rubrics as a way to provide feedback information (Nordrum, Evans, & Gustafsson, 2013).

The rubric-based performance evaluation instrument is utilized to evaluate students' problem-solving abilities, with each problem being assigned a distinct rubric. Problem-solving abilities should be prioritized in the assessment of mathematics learning since they inspire students to engage in wide and creative thinking to address obstacles (Zakiah, Hendriana, & Hidayat, 2022). The mathematical problem can be resolved by following the procedures below: (1) Comprehending the issue, which entails the identification of the known and unknown elements, as well as the assessment of their suitability for problem-solving; (2) creating a mathematical model of the problem by connecting the known and unknown elements; (3) selecting a solution strategy, elaborating and conducting calculations, or finalizing a mathematical model; and (4) interpreting the original problem's results and reevaluating the validity of the proposed solution.The assessment process of mathematics learning should prioritize problem-solving skills, as this will motivate students to think creatively and extensively in order to address the obstacles they encounter (Nasir & Syartina, 2021; Polya, 1973)

The culture of a region reflects the history experienced during that period. In general, culture is a way of life that guides each individual to understand how to act and behave when interacting with others. One form of culture is human-made objects that hold significant value and meaning for life, known as cultural heritage (Faturrahmann et al., 2022). One form of application that can be observed from culture is historical buildings. Masjid Al-Alam Marunda, located in the Cilincing district of North Jakarta, is an example of a historical building. The authenticity of the mosque's architectural form has been preserved until now. The Al-Alam Mosque can be analyzed and studied using mathematical concepts, especially geometry, due to its structure being closely related to geometry. With a culturally contextual approach to geometry material, it is hoped that students can further explore their metacognitive abilities, critical thinking, and problem-solving skills. Thus, the cultural contextual approach can be seen from everyday real objects, making it easier for students to visualize them directly in their minds (Sarwoedi, Marinka, Febriani, & Wirne, 2018).

METHODS

This study used a quantitative approach involving 381 students from two MTsN in Jakarta as respondents. The instrument for assessing student performance on geometry material developed consisted of 5 questions. This instrument has been constructed and tested empirically in this study, with a working time only two hours of lessons (80 minutes). Students were given a worksheet containing a performance assessment instrument that included two main components: contextual-based questions about Jakarta's cultural heritage, as many as five items, and an assessment rubric with three assessment criteria. There are a total of 15 items, denoted as B11, B12, B13, B21, B22, B23, B31, B32, B33, B41, B42, B43, B51, B52, and B53, and each question ranges from 1 to 5 using the assessment rubrics 1, 2, and 3. This was followed by the analysis of the field data using Winsteps version 3.73. Item validity and instrument reliability were assessed through the application of the Rasch model. This step is important to improve the quality of the instrument so that it can measure a construct more accurately. The instrument used is the student worksheet. The following is a grid of the performance assessment question sheets used in Table 1.

Table 1
Performance assessment question sheets

No	Competencies	Assessment Indicators
1	Identifying the properties of cubes, cuboids, prisms, and pyramids and their parts	Identifying and determining the elements of a rectangular pyramid
2	Solving problems related to the surface area and volume of flat-sided shapes (prisms and pyramids).	Solving problems related to the surface area of a hexagonal prism
3	Solving problems related to the surface area and volume of flat-sided shapes (prisms and pyramids).	Solving problems related to the surface area of a pyramid
4	Solving problems related to the surface area and volume of flat-sided shapes (prisms and pyramids)	Solving problems related to the volume of a hexagonal prism
5	Solving problems related to the surface area and volume of flat-sided shapes (prisms and pyramids)	Solving problems related to the volume of a cube and pyramid

The Rasch model can be applied to generate instruments that are both valid and reliable (Khatimin, Aziz, Zaharim, & Mat Yasin, 2013). The Rasch model definition of reliability is a modification of the classic Cronbach's alpha estimator. A new logistic alpha estimator is proposed. Ultimately, the Rasch model is employed to simulate the estimator's properties (Martinková & Zvára, 2007). In order to evaluate the functional items of reliability and separation of items and respondents, polarity and item fit of measurement constructs, and standardized residual correlation values, the Rasch measurement model approach is employed with the assistance of Winsteps software (Yasin, Yunus, Rus, Ahmad, & Rahim, 2015). The Rasch measurement model is capable of evaluating the accuracy of the test, the reliability of the item and respondent, and the consistency of construct interpretation. Furthermore, the Rasch measurement model framework underscores the importance of evaluating the appropriateness of the measurement scale.

Teachers use rubrics as assessment guidelines to clarify the criteria they want to use when evaluating student work results. This rubric includes a list of criteria that are expected to appear in student work, complete with guidelines for evaluating each of these criteria. The purpose of the assessment rubric is so that students can clearly understand the basis for the assessment that will be used to measure student performance (Yudha, 2020). Thus, both teachers and students will have clear, shared guidelines regarding the expected performance demands. The following is an example of a rubric for questions related to determining the surface area of prisms and pyramids. Each student's responses are compared to the descriptions given in the rubric. Scores are given according to the level of achievement in each criterion. Add up the scores from all criteria to get each student's total score. The maximum and minimum scores depend on the number of criteria and the assessment scores in the rubric.

Table 2
Performance Assessment Rubric

No	Criteria	Score			
		1	2	3	4
1	Problem Solving Approach	There is no solution that uses a precise formula to solve every part of the problem	There is a solution that uses a precise formula to solve at least one part of the problem	There is a solution that uses several precise formulas to solve several parts of the problem	The solution uses precise formulas to solve all parts of the problem
2	Calculation Accuracy	Not systematic and there are many calculation errors	Some calculation steps are not systematic, miscalculations and wrong results	Systematic calculation steps, slight miscalculations and wrong results	Systematic calculation steps, correct calculations and correct results
3	Drawing Sketches	No sketches	There are drawing sketches but do not support the solution	There are drawing sketches but do not support the solution	The sketches strongly support the solution

(Adapted from Yudha, 2020)

One factor that can reduce the validity of performance assessment is bias. When a teacher interprets student performance incorrectly, it's because they are assessing a group of students with different criteria or characteristics (Yudha, 2020). In assessing student performance, a teacher must choose and use procedures that are fair to all students, regardless of their cultural background, language, or gender. Additionally, if the teacher fails to include or provide an appropriate assessment of student performance, it can also affect the validity of the performance assessment. Therefore, more than one teacher is needed to ensure agreement on values and avoid bias. The validation process of student performance instruments in geometry subjects relies on expert assessment through statistical measurements using the Aiken V model. Inter-rater agreement is assessed based on the position or ranking of the observed subjects. The assessors who will be involved in the assessment process are three certified mathematics teachers. The use of certified teachers aims to minimise the level of subjectivity in the assessment process.

In assessing the quality of research instruments, it is important to pay attention to the reliability of the performance measurement tool so that the items can produce consistent decisions according to the measurement objectives. Reliability aims to determine the level of measurement consistency (Fitri, 2017). Reliability refers to the consistency of the measuring instrument in assessing a construct. This means that an excellent measurement will show a high level of reliability if the resulting scores are consistent. A measurement that is able to assess a variable consistently is said to have high reliability. On the other hand, measurements that produce varying scores for the same construct are considered inconsistent and have low reliability (Fitri, 2017; Loewenthal & Lewis, 2001).

Furthermore, the Rasch Model technique was used to analyze the research data. The Rasch model offers accurate data during instrument assessment (Fitri, 2017). In Rasch model analysis, item fit is assessed to determine how well the item contributes to the measurement of the underlying construct. When conducting Rasch model analysis, item fit is usually evaluated using fit statistics such as infit statistics and outfit mean square (MNSQ) (Ocy, Rahayu, & Makmuri, 2023). The order of item fit in Rasch model analysis is determined by examining the fit statistics of each item (Sumintono, 2018). Then, the results of the performance assessment instrument are reviewed according to the following criteria: unidimensionality, basic item analysis, and instrumental analysis.

The results of the validity analysis obtained from the Winstep® program are analysis in the form of construct validity and content validity, where the content validity analysis includes the level of suitability of the questions which function to see the quality of the level of suitability of the question items with the mode. The data provided is in the form of information on the suitability of the question items with the criteria, namely by looking at the outfit mean square value (MNSQ), outfit z-standard (ZSTD), and point measure correlation. Question items are said to be valid or accepted if they have met at least 2 criteria and are corrected if they meet one of the three criteria, and are discarded if none meet the criteria. The suitability value of the items is greatly influenced by the amount of data, the larger the sample used, the better the suitability level. The criterion value used to see the level of suitability of the items (content validity)(Sumintono & Widhiarso, 2015).

Separation is the combination of students and their items. The higher the separation value, the better the quality of the instrument. The formulation used to prove the combination of students and items is in equation (1) (Misbach & Sumintono, 2014):

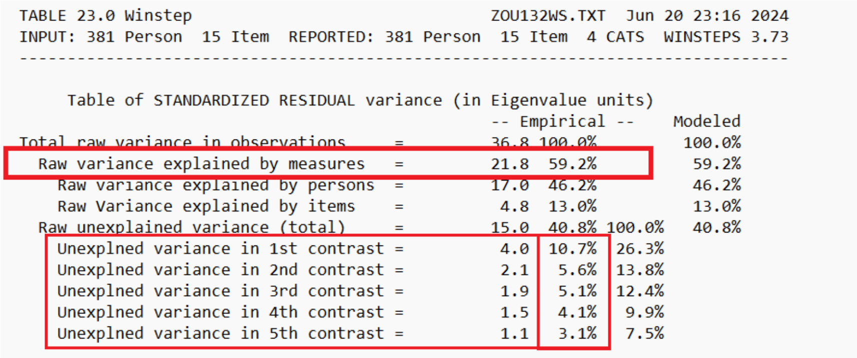
$$\mathrm{H} = \frac{(4 \times separation) + 1}{3} \tag{1}$$

RESULT AND DISCUSSION

In this section, the researcher presents the results and discussion of the study. The results of the student performance assessment using the Rasch Model are reviewed based on several aspects, namely unidimensionality, item analysis (including the level of difficulty and the level of suitability of the item), and instrument analysis in detail explained as follows.

Unidimensionality

Unidimensionality analysis identifies several aspects measured by the instrument. The Dimensionality Map menu output table 23 of Winsteps software version 3.73 displays the requirements for unidimensionality, highlighting the values of raw variance explained by measures and unexplained variance in the 1st to 5th contrast. It is possible to show that a set is one-dimensional if the variance that can be explained by the measures is less than 20%, with the range being sufficient (20–40%), good (40–60%), and very good (more than 60%), and if the variation in the residuals from the first to fifth contrasts is less than 15% for each (Salsabila, Nurihsan, & Sunarya, 2023).



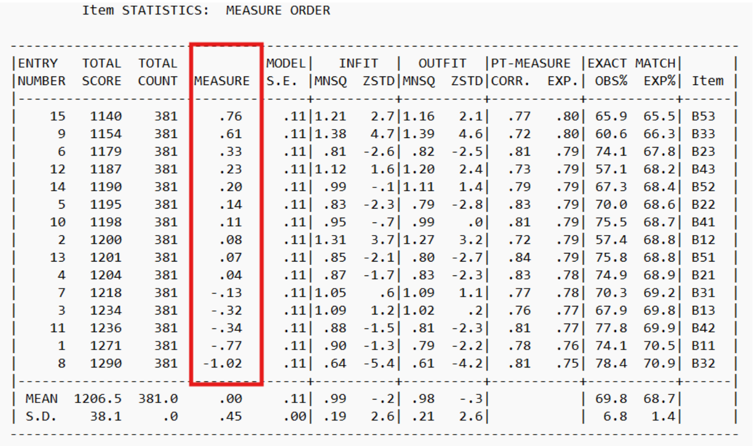
```
TABLE 23.0 Winstep                                ZOU132WS.TXT  Jun 20 23:16 2024
INPUT: 381 Person  15 Item  REPORTED: 381 Person  15 Item  4 CATS  WINSTEPS 3.73
--------------------------------------------------------------------------------

     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                                 -- Empirical --    Modeled
Total raw variance in observations      =       36.8 100.0%         100.0%
  Raw variance explained by measures    =       21.8  59.2%          59.2%
    Raw variance explained by persons   =       17.0  46.2%          46.2%
    Raw Variance explained by items     =        4.8  13.0%          13.0%
  Raw unexplained variance (total)      =       15.0  40.8% 100.0%   40.8%
    Unexplned variance in 1st contrast =        4.0  10.7%  26.3%
    Unexplned variance in 2nd contrast =        2.1   5.6%  13.8%
    Unexplned variance in 3rd contrast =        1.9   5.1%  12.4%
    Unexplned variance in 4th contrast =        1.5   4.1%   9.9%
    Unexplned variance in 5th contrast =        1.1   3.1%   7.5%
```

Figure 1
Unidimensional Test Results

The unidimensional test in Figure 1 yielded a 59.2% explanation of the observed raw variance by measures, placing it in the good category. The unexplained variance in the 1st contrast is 10.7%, whereas it is 5.6% in the 2nd contrast, 5.1% in the 3rd contrast, 4.1% in the 4th contrast, and 3.1% in the 5th contrast. None of the unknown variance values or unexplained variance exceed 15%. This means that the instrument is unidimensional, meaning that the instrument used really tests one variable or is said to be able to measure student performance.

Item Analysis

The item fit and difficulty level (item measure) are then included in the item analysis. The Winstep application's item measure (Figure 2) can be used to examine the item's degree of difficulty.



```
        Item STATISTICS:  MEASURE ORDER

-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT  |  OUTFIT  |PT-MEASURE |EXACT MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item |
|------------------------------------+----------+----------+-----------+-----------+------|
|    15   1140    381      .76    .11|1.21   2.7|1.16   2.1|  .77   .80| 65.9  65.5| B53  |
|     9   1154    381      .61    .11|1.38   4.7|1.39   4.6|  .72   .80| 60.6  66.3| B33  |
|     6   1179    381      .33    .11| .81  -2.6| .82  -2.5|  .81   .79| 74.1  67.8| B23  |
|    12   1187    381      .23    .11|1.12   1.6|1.20   2.4|  .73   .79| 57.1  68.2| B43  |
|    14   1190    381      .20    .11| .99   -.1|1.11   1.4|  .79   .79| 67.3  68.4| B52  |
|     5   1195    381      .14    .11| .83  -2.3| .79  -2.8|  .83   .79| 70.0  68.6| B22  |
|    10   1198    381      .11    .11| .95   -.7| .99    .0|  .81   .79| 75.5  68.7| B41  |
|     2   1200    381      .08    .11|1.31   3.7|1.27   3.2|  .72   .79| 57.4  68.8| B12  |
|    13   1201    381      .07    .11| .85  -2.1| .80  -2.7|  .84   .79| 75.8  68.8| B51  |
|     4   1204    381      .04    .11| .87  -1.7| .83  -2.3|  .83   .78| 74.9  68.9| B21  |
|     7   1218    381     -.13    .11|1.05    .6|1.09   1.1|  .77   .78| 70.3  69.2| B31  |
|     3   1234    381     -.32    .11|1.09   1.2|1.02    .2|  .76   .77| 67.9  69.8| B13  |
|    11   1236    381     -.34    .11| .88  -1.5| .81  -2.3|  .81   .77| 77.8  69.9| B42  |
|     1   1271    381     -.77    .11| .90  -1.3| .79  -2.2|  .78   .76| 74.1  70.5| B11  |
|     8   1290    381    -1.02    .11| .64  -5.4| .61  -4.2|  .81   .75| 78.4  70.9| B32  |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN  1206.5  381.0      .00    .11| .99   -.2| .98   -.3|           | 69.8  68.7|      |
| S.D.    38.1     .0      .45    .00| .19   2.6| .21   2.6|           |  6.8   1.4|      |
-------------------------------------------------------------------------------------------
```

Figure 2
Item Difficulty Level

Figure 2 reveals a standard deviation of 0.45. Combining the standard deviation value with the average logit value allows for the classification of the item difficulty level into several levels. The guidelines for evaluating the item are categorized into four groups. (Sumintono & Widhiarso, 2015):

Table 3

Difficulty level criteria

Measure Value	Criteria
< -1	item is very easy
-1 s.d. 0	easy items
0 s.d. 1	difficult item
> 1	item is very difficult

(Adapted from Sumintoro & Widhiarso, 2015)

The category of difficult objects includes ten items: B53, B33, B23, B43, B52, B22, B41, B12, B51, and B21. Four items, B31, B13, B42, and B11, have received the "easy" classification. Item B32 is classified with ease. The degree of item match determines the interpretation of the item. Item fit indicates whether the item functions as intended for the purpose of conducting measurements. A suitable item (fit) exhibits behavior that is consistent with the model's expectations. Students may harbor misconceptions regarding the item when it fails to suit (Karami, 2015).

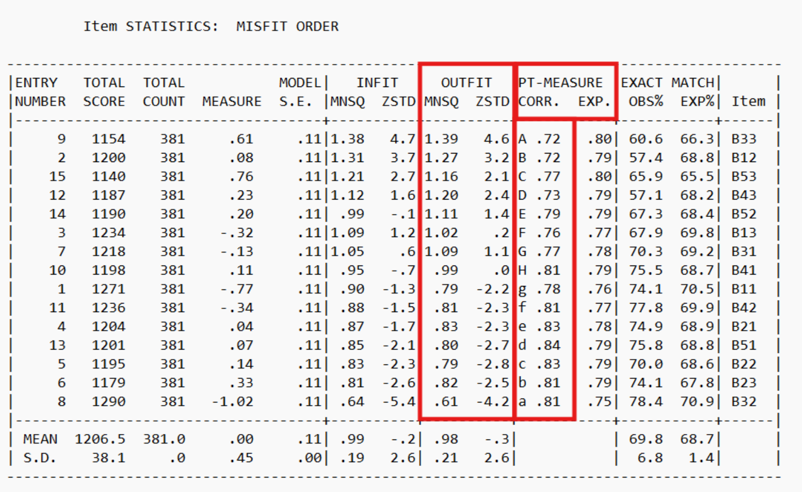
The Rasch Model test, which ascertains the item fit value or item fit order, includes several significant values, such as Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), and Point Measure Correlation (Pt. Mean Corr). The model determines that an item is fit if its behavior is consistent with its anticipated behavior, which suggests that respondents do not misinterpret the item (Suryani, 2018). The following criteria are used to identify items that are suitable: The fit criteria requirements for the Rasch Model Winstep software are detailed in the following table (Boone, Staver, & Yale, 2014).

Table 4
Item Fit

Criteria	Measure Point
Outfit MNSQ	$0,5 < MNSQ < 1,5$
Outfit ZSTD	$-2,0 < ZSTD < +2,0$
Point Measure Correlation	$0,4 < Pt\ Measure\ Corr < 0,85$

(Adapted from Boone, Staver, & Yale, 2014)

The optimal value for Outfit MNSQ is approximately 1, whereas the optimal value for Outfit ZSTD is approximately 0. Questions that do not align may indicate an error in the answer key, random responses from participants, or questions with insufficient discrimination power. Point Measure Correlation (Pt. Mean Corr) is utilized to assess discriminatory power. A Pt. Measure Corr value of 1.0 signifies that respondents with high ability answer the item accurately. A negative result signifies that the item is deceptive, as those with low ability may react accurately while those with high ability may respond inaccurately. The subject of analysis pertains to data processing utilizing the Winstep menu item fit order as presented in Figure 3.



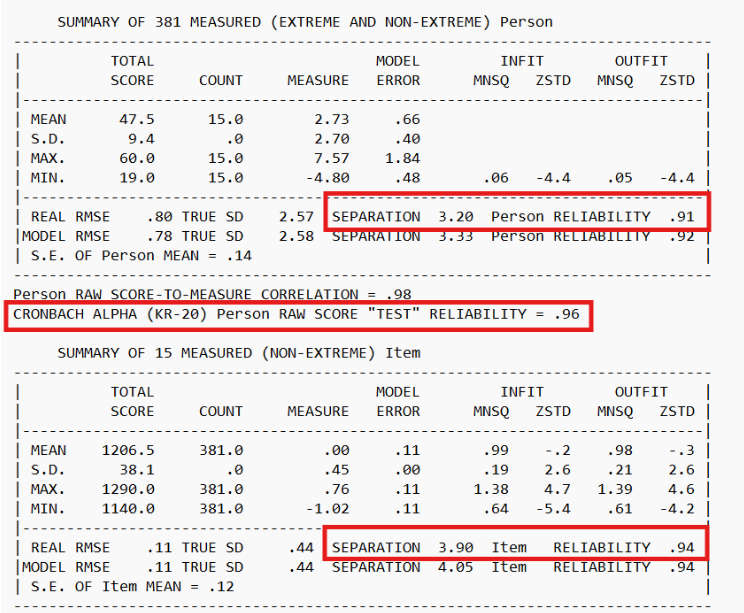
```
        Item STATISTICS:  MISFIT ORDER

-------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT    OUTFIT   |PT-MEASURE |EXACT  MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| Item |
|------------------------------------+----------+----------+-----------+-----------+------|
|     9   1154    381      .61    .11|1.38   4.7|1.39   4.6|A .72   .80| 60.6  66.3| B33  |
|     2   1200    381      .08    .11|1.31   3.7|1.27   3.2|B .72   .79| 57.4  68.8| B12  |
|    15   1140    381      .76    .11|1.21   2.7|1.16   2.1|C .77   .80| 65.9  65.5| B53  |
|    12   1187    381      .23    .11|1.12   1.6|1.20   2.4|D .73   .79| 57.1  68.2| B43  |
|    14   1190    381      .20    .11| .99   -.1|1.11   1.4|E .79   .79| 67.3  68.4| B52  |
|     3   1234    381     -.32    .11|1.09   1.2|1.02    .2|F .76   .77| 67.9  69.8| B13  |
|     7   1218    381     -.13    .11|1.05    .6|1.09   1.1|G .77   .78| 70.3  69.2| B31  |
|    10   1198    381      .11    .11| .95   -.7| .99    .0|H .81   .79| 75.5  68.7| B41  |
|     1   1271    381     -.77    .11| .90  -1.3| .79  -2.2|g .78   .76| 74.1  70.5| B11  |
|    11   1236    381     -.34    .11| .88  -1.5| .81  -2.3|f .81   .77| 77.8  69.9| B42  |
|     4   1204    381      .04    .11| .87  -1.7| .83  -2.3|e .83   .78| 74.9  68.9| B21  |
|    13   1201    381      .07    .11| .85  -2.1| .80  -2.7|d .84   .79| 75.8  68.8| B51  |
|     5   1195    381      .14    .11| .83  -2.3| .79  -2.8|c .83   .79| 70.0  68.6| B22  |
|     6   1179    381      .33    .11| .81  -2.6| .82  -2.5|b .81   .79| 74.1  67.8| B23  |
|     8   1290    381    -1.02    .11| .64  -5.4| .61  -4.2|a .81   .75| 78.4  70.9| B32  |
|------------------------------------+----------+----------+-----------+-----------+------|
| MEAN  1206.5  381.0      .00    .11| .99   -.2| .98   -.3|           | 69.8  68.7|      |
| S.D.    38.1     .0      .45    .00| .19   2.6| .21   2.6|           |  6.8   1.4|      |
-------------------------------------------------------------------------------------------
```

Figure 3
Item Fit Order

Figure 3 shows that all items fall into the fit criteria Outfit MNSQ with a range of 0.61 to a maximum of 1.39. Based on the second criterion of Outfit ZSTD, there are 11 items, namely items B33, B12, B53, B43, B11, B42, B21, B51, B22, B23, and B32 that do not fall into the fit criteria. Based on the third criterion, all items have Pt measure corr values ranging from 0.73 to 0.84. All items are said to be valid or accepted because they have met at least 2 criteria. In general, it can be concluded that all items in the student performance assessment rubric are considered valid, which means that the items are functional and can be understood by students..

Instrument Analysis

The Winstep application's table 3.1 Summary Statistics provides the information for the instrument analysis.



```
     SUMMARY OF 381 MEASURED (EXTREME AND NON-EXTREME) Person

-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT         OUTFIT    |
|          SCORE      COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|------------------------------------------------------------------------------|
| MEAN      47.5       15.0        2.73     .66                                |
| S.D.       9.4         .0        2.70     .40                                |
| MAX.      60.0       15.0        7.57    1.84                                |
| MIN.      19.0       15.0       -4.80     .48       .06   -4.4    .05   -4.4 |
|------------------------------------------------------------------------------|
| REAL RMSE    .80 TRUE SD    2.57  SEPARATION  3.20  Person RELIABILITY  .91 |
|MODEL RMSE    .78 TRUE SD    2.58  SEPARATION  3.33  Person RELIABILITY  .92 |
| S.E. OF Person MEAN = .14                                                    |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .96

     SUMMARY OF 15 MEASURED (NON-EXTREME) Item

-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT         OUTFIT    |
|          SCORE      COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|------------------------------------------------------------------------------|
| MEAN    1206.5      381.0         .00     .11       .99    -.2    .98    -.3 |
| S.D.      38.1         .0         .45     .00       .19    2.6    .21    2.6 |
| MAX.    1290.0      381.0         .76     .11      1.38    4.7   1.39    4.6 |
| MIN.    1140.0      381.0       -1.02     .11       .64   -5.4    .61   -4.2 |
|------------------------------------------------------------------------------|
| REAL RMSE    .11 TRUE SD     .44  SEPARATION  3.90  Item   RELIABILITY  .94 |
|MODEL RMSE    .11 TRUE SD     .44  SEPARATION  4.05  Item   RELIABILITY  .94 |
| S.E. OF Item MEAN = .12                                                      |
-------------------------------------------------------------------------------
```

Figure 4
Summary statistic person & item

The person measure shows the average score of all respondents when they complete the student performance assessment instrument. If the person average is greater than the item average (with the item average being 0.00 logit), this indicates that the respondent's ability is generally higher than the level of difficulty of the items in the instrument.

In figure 4, it is known that the separation test shows that the respondent separation index is 3.20 and the item separation is 3.90, meaning that the quality of the respondents and the items is both good. The higher the separation value, the better the overall quality of the person and instrument. The strata separation equation (H) is an additional equation to determine more specific grouping and using the formula (1):

$$\mathrm{H} = \frac{(4 \times separation) + 1}{3} \tag{1}$$

$$\mathrm{H}_{respondent} = \frac{(4 \times 3{,}20) + 1}{3} = \frac{13{,}8}{3} = 4{,}6$$

$$\mathrm{H}_{item} = \frac{(4 \times 3{,}90) + 1}{3} = \frac{16{,}6}{3} = 5{,}53$$

Using this formula, the respondent strata separator is 4.6, rounded to 5, and the item strata separator is 5.53, rounded to 6. This suggests that the respondents' abilities can be classified into five categories, spanning from extremely low to very high ability. Meanwhile, six categories, ranging from very easy to very difficult, distribute the items' level of difficulty.

William P. Fisher, Jr. developed Table 4 based on the Rasch literature and his extensive experience conducting Rasch analysis in various settings (Fisher, 2007).

Table 5
Rating Scale Instrument Quality Criteria

Criteria	**Poor**	**Fair**	**Good**	**Very Good**	**Excellend**
Person and Item Strata Separated	< 2	2-3	3-4	4-5	>5
Person and Item Measurement Reliability	< 0,67	0,67-0,80	0,81-0,90	0,91-0,94	>0,94

In figure 5, the person's reliability value of 0.91 indicates the consistency of the respondents' answers in the very good category. The item reliability value of 0.94 indicates that the quality of the items in the instrument is also in the excellent category. The Cronbach Alpha reliability value has 4 categories: excellent (0.80 to 1.00), excellent (0.70 to 0.80), sufficient (0.60 to 0.70), and poor (0.00 to 0.60) (Bond, Yan, & Heene, 2020). The Cronbach Alpha value of 0.96 signifies the overall interaction between the individual and the item, placing it in the very good category. Overall, the validity and reliability of the performance assessment instrument are strong, enabling it to accurately measure student performance on performance assessment questions that contextualize Jakarta's cultural heritage buildings in geometry.

CONCLUSION AND IMPLICATION

Conclusion

The accuracy, objectivity, and consistency of the data obtained by the researcher can be ensured by examining the student performance evaluation instrument in the context of Jakarta's cultural heritage buildings and applying the Rasch model to the data analysis. The Rasch model measurement can describe the interaction between the respondents and the items. Based on the results of the study, the performance assessment instrument developed has been proven valid by meeting at least 2 criteria, namely the Outfit MNSQ criteria ranging from 0.61 to a maximum of 1.39 and Pt measure corr ranging from 0.73 to 0.84. The person reliability value is 0.91, which indicates a very good level of consistency in respondents' answers, and the item reliability value is 0.94, which indicates the quality of the items in the instrument is also very good. The Cronbach Alpha value indicating the interaction between the person and the item items as a whole is 0.96, which is included in the very good category. Therefore, mathematics teachers can use this instrument to measure student performance in geometry.

The limitation of this study lies in its instrument, which combines components of Jakarta's cultural heritage, so students from other areas may have difficulty using it because they are not familiar with the cultural background. Further research is necessary to develop student performance assessment instruments that incorporate cultural elements from various other areas, making them more familiar to students in the area.

Implication

The researchers provide recommendations to several parties based on the results of the study. The high validity and reliability of the assessment instrument, as demonstrated by the Rasch Model analysis, imply that educators can make more accurate and fair evaluations of student performance. This guarantees the assessment of students' true abilities and understanding of geometry within Jakarta's cultural heritage. With reliable data from performance assessments, teachers can better identify areas where students excel or struggle. This allows for targeted interventions and instructional adjustments to meet individual student needs, ultimately improving overall learning outcomes. Educational policymakers can use the findings from this analysis to set standards and guidelines for performance assessments. The proven reliability and validity of this instrument can serve as a benchmark for developing similar assessments in other subjects and areas.

ACKNOWLEDGMENTS

Thank you to class VIII MTsN 28 and MTsN 15 Jakarta for the academic year 2023/2024, who have actively participated in working on students' worksheets given by researchers.

REFERENCES

Alfiatin, A. L., & Oktiningrum, W. (2019). Pengembangan Soal Higher Order Thinking Skills Berbasis Budaya Jawa Timur untuk Mengukur Penalaran Siswa SD. *Indiktika: Jurnal Inovasi Pendidikan Matematika*, *2*(1), 30–43. https://doi.org/10.31851/indiktika.v2i1.3395

Anis, D., Rohendi, D., & Sukandar, A. (2019). Pengembangan Instrumen Penilaian Kinerja Siswa Pada Praktikum Gambar Proyeksi Ortogonal. *Journal of Mechanical Engineering Education*, *6*(2), 161–167. https://doi.org/10.17509/jmee.v6i2.21787

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Rating Scale Surveys. In *Rasch Analysis in the Human Sciences* (pp. 21–46). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-6857-4_2

Brookhart, S. M. (2015). *Performance Assessment: Showing What Students Know and Can Do*. Learning Sciences International.

Faturrahmann, A., Fachly, A. F. R., Bhakti, T. S. D., Putri, E. N., Puspitasari, C., & Arni, M. (2022). ADAPTATIVE REUSE DAN PENDEKATAN CONTEXTUAL JUXTAPOSITION PADA STASIUN JATINEGARA, JAKARTA. *Lakar: Jurnal Arsitektur*, *5*(2), 115. https://doi.org/10.30998/lja.v5i2.14470

Fisher, W. P. (2007). Rating scale instrument quality criteria. In *Rasch measurement transactions*: *Vol. 21(1)*, 1095.

Fitri. (2017). Analisis Validitas dan Reliabilitas Instrumen Kinerja Akuntan Menggunakan Pendekatan Rasch Model. *Jurnal Ilmiah Akuntansi Peradaban*, *3*(1), 34–45.

Indihadi, D., Suryana, D., & Ahmad, A. B. (2022). The Analysis of Construct Validity of Indonesian Creativity Scale Using Rasch Model. *Creativity Studies*, *15*(2), 560–576. https://doi.org/10.3846/cs.2022.15182

Karami, H. (2015). Book review: Rasch Analysis in the Human Sciences. *Language Testing*, *32*(4), 545–548. https://doi.org/10.1177/0265532214567642

Kartikasari, D. A., & Rahayu, W. (2018). Pemahaman Matematika : Penerapan Model Arcs. *Jurnal Evaluasi Pendidikan*, *9*, 6–15. Retrieved from doi.org/10.21009/JEP.091.02%0D

Khatimin, N., Aziz, A. A., Zaharim, A., & Mat Yasin, S. H. (2013). Development of Objective Standard Setting Using Rasch Measurement Model in Malaysian Institution of Higher Learning. *International Education Studies*, *6*(6). https://doi.org/10.5539/ies.v6n6p151

Khoriyah, M., & Oktiningrum, W. (2021). Pengembangan soal higher order thinking Skills (HOTS) berbasis budaya lokal blitar untuk mengukur dimensi pengetahuan matematika siswa kelas V sekolah dasar. *Bina Gogik: Jurnal Ilmiah Pendidikan Guru Sekolah Dasar*, *8*(1), 2579–4647.

Loewenthal, K. M., & Lewis, C. A. (2001). *An Introduction to Psychological Tests and Scales*. Psychology Press. https://doi.org/10.4324/9781315782980

Martinková, P., & Zvára, K. (2007). *Reliability in the Rasch model*. Kybernetika. Retrieved from http://eudml.org/doc/33860

Misbach, I. H., & Sumintono, B. (2014). Pengembangan dan Validasi Instrumen Persepsi Siswa Tehadap Karakter Moral Guru di Indonesia dengan Model Rasch. *PROCEEDING Seminar Nasional Psikometri*, *148–162.*

Muslihin, H. Y., Suryana, D., Ahman, A., Suherman, U., & Dahlan, T. H. (2022). Analysis of the Reliability and Validity of the Self-Determination Questionnaire Using Rasch Model. *International Journal of Instruction*, *15*(2), 207–222. https://doi.org/10.29333/iji.2022.15212a

Nasir, A. M., & Syartina, S. (2021). The Effectiveness of the Polya Model Problem Solving Method on Student Learning Outcomes in Solving Math Story Problems. *Eduma : Mathematics Education Learning and Teaching*, *10*(2), 127.

https://doi.org/10.24235/eduma.v10i2.8700

Ningsih, G., & Rahayu, W. P. (2021). Pengembangan Instrumen Penilaian Psikomotor Berbasis Proyek Untuk Meningkatkan Kompetensi Psikomotor Siswa Dalam Pembelajaran Blended Learning Di Era New Normal. *Jurnal Ekonomi, Bisnis Dan Pendidikan*, *1*(34), 418–424. https://doi.org/10.17977/um066v1i52021p418-424

Nordrum, L., Evans, K., & Gustafsson, M. (2013). Comparing student learning experiences of in-text commentary and rubric-articulated feedback: strategies for formative assessment. *Assessment & Evaluation in Higher Education*, *38*(8), 919–940. https://doi.org/10.1080/02602938.2012.758229

Oberg, C. (2010). Guiding Classroom Instruction Through Performance Assessment. *Journal of Case Studies in Accreditation and Assessment*, *1*, 1–11.

Ocy, D. R., Rahayu, W., & Makmuri, M. (2023). RASCH MODEL ANALYSIS: DEVELOPMENT OF HOTS-BASED MATHEMATICAL ABSTRACTION ABILITY INSTRUMENT ACCORDING TO RIAU ISLANDS CULTURE. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *12*(4), 3542. https://doi.org/10.24127/ajpm.v12i4.7613

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, *9*, 129–144. https://doi.org/10.1016/j.edurev.2013.01.002

Polya, G. (1973). *How To Solve (2nd Ed)*. Princeton: University Press.

Purnomo, Y. W. (2013). Komputasi Mental untuk Mendukung Lancar Berhitung Operasi Penjumlahan dan Pengurangan pada Siswa Sekolah Dasar. *Seminar Nasional Matematika Dan Pendidikan Matematika*, (November), 657–662.

Putra, Z. H., Hermita, N., & Alim, J. A. (2021). Analisis Pengetahuan Matematika, Didaktika, dan Teknologi Calon Guru Sekolah Dasar Menggunakan Rasch Model. *Mosharafa: Jurnal Pendidikan Matematika*, *10*(3), 345–356. https://doi.org/10.31980/mosharafa.v10i3.1042

Ratumanan, T. G., & Laurens, T. (2016). Analisis Penguasaan Objek Matematika. *Jurnal Pendidikan Matematika Raflesia*, *1*(2), 146–154. https://doi.org/10.33369/jpmr.v1i2.4005

Salsabila, F., Nurihsan, J., & Sunarya, Y. (2023). Pengujian Validitas dan Reliabilitas Instrumen Manajemen Diri Remaja:Rasch Model Analysis. *Jurnal Bimbingan Dan Konseling Terapan*, *07*(01), 86–102. Retrieved from https://ojs.unpatti.ac.id/index.php/bkt/article/view/234/158

Sarwoedi, S., Marinka, D. O., Febriani, P., & Wirne, I. N. (2018). Efektifitas etnomatematika dalam meningkatkan kemampuan pemahaman matematika siswa. *Jurnal Pendidikan Matematika Raflesia, 3(2),* 171–176. https://doi.org/https://doi.org/10.33369/jpmr.v3i2.7521

Sumintono, B. (2018). Rasch Model Measurements as Tools in Assesment for Learning. *Proceedings of the 1st International Conference on Education Innovation (ICEI 2017)*. Paris, France: Atlantis Press. https://doi.org/10.2991/icei-17.2018.11

Sumintono, B., & Widhiarso, W. (2015). Aplikasi pemodelan rasch pada assessment pendidikan. In *Trim komunikata*. Trim komunikata.

Suryani, Y. E. (2018). Aplikasi Rasch Model dalam Mengevaluasi Intelligenz Structure Test (IST). *Psikohumaniora: Jurnal Penelitian Psikologi*, *3*(1), 73. https://doi.org/10.21580/pjpp.v3i1.2052

Taufiq, A., Yudha, E. S., Md, Y. H., & Suryana, D. (2021). Examining the Supervision Work Alliance Scale: A Rasch Model Approach. *The Open Psychology Journal*, *14*(1), 179–184. https://doi.org/10.2174/1874350102114010179

Yasin, R. M., Yunus, F. A. N., Rus, R. C., Ahmad, A., & Rahim, M. B. (2015). Validity and Reliability Learning Transfer Item Using Rasch Measurement Model. *Procedia - Social and Behavioral Sciences*, *204*, 212–217. https://doi.org/10.1016/j.sbspro.2015.08.143

Yudha, R. P. (2020). Validity And Reliability Rubric Of Performance Assessment Geometry Study In Junior High School Using The Many Facet Rasch Model Approach. *Eduma : Mathematics Education Learning and Teaching*, *9*(2), 26. https://doi.org/10.24235/eduma.v9i2.7100

Zakiah, I., Hendriana, H., & Hidayat, W. (2022). The Effect of Contextual Learning Trough Teaching Materials Application-Based on Problem Solving Ability. *Eduma : Mathematics Education Learning and Teaching*, *11*(1), 20. https://doi.org/10.24235/eduma.v11i1.9604