JHSS (Journal of Humanities and Social Studies)
https://journal.unpak.ac.id/index.php/jhss
Volume 09, Number 01, Page 053-056
e-ISSN: 2598-120X; p-ISSN: 2598-117X

ANALYSIS OF PHYSICS TEST ITEMS FOR GRADE XI USING THE RASCH MODEL AT SMAN 1 SAPE

Rudy Sumiharsono [a*)]

[a)] *Universitas PGRI Adi Buana Surabaya, Surabaya, Indonesia*

[*)]*Corresponding Author: rudy.sumi8@gmail.com*

Article history: *received 21 January 2025; revised 02 February 2025; accepted 04 March 2025* *DOI: https://doi.org/10.33751/jhss.v9i1.11383*

Abstract. Assessment is an integral part of the learning process, aiming to measure the achievement of educational objectives. High-quality assessment instruments require evaluation of item validity, reliability, and difficulty level. This study aims to analyze the quality of Physics test items for grade XI at SMAN 1 Sape, Bima Regency, using the Rasch Model. This approach was selected because it provides detailed information on item characteristics, including validity, reliability, and difficulty, in a quantitative manner. The study employed a descriptive quantitative method. Data were collected from midterm examination results of grade XI students from purposively selected schools. A total of 25 multiple-choice items were analyzed using WINSTEP software with the Rasch Model approach. The analysis was conducted to identify items that were valid, reliable, and appropriately difficult according to students' ability profiles. The results showed that most items lacked adequate validity and required revision. The reliability of the instrument was categorized as moderate, indicating a sufficient level of consistency in measuring students' abilities. Regarding item difficulty, most questions were classified as difficult, with only a few categorized as easy or very easy. These findings underscore the importance of regular evaluation of test items to ensure the quality of assessment instruments.

Keywords: Rasch Model; item analysis; validity; reliability; difficulty level; Physics grade XI.

I. INTRODUCTION

In the learning process, the interaction between students and teachers is a primary factor in achieving educational goals. Teachers play a vital role not only as educators but also as evaluators who ensure the quality of learning through effective assessments [1][2]. Assessment, as an integral part of education, serves to evaluate the extent to which learning objectives have been achieved [3]. A robust evaluation system not only supports successful learning but also forms the basis for improving overall educational quality [4][5].

Assessment tools are used to measure students' abilities in achieving specific competencies. These tools can take the form of tests and non-tests, with tests being the primary means for quantitatively evaluating learning outcomes [6][7]. A well-constructed test should meet criteria for validity, reliability, difficulty level, and item discrimination [8]. Multiple-choice tests, one of the most commonly used assessment formats, offer efficiency in measuring a wide range of student competencies. Thus, the quality of test items becomes a critical aspect of educational assessment.

Interviews with educators revealed that many test items used have not undergone empirical testing, leading to unverified quality. This may result in assessments that fail to accurately reflect students' abilities. Classical Test Theory (CTT) is commonly used for item analysis, but it has limitations, such as dependence on the respondent population's characteristics [9]. These limitations highlight the need for more modern and precise analytical approaches. The modern Rasch Model offers a solution by providing stable item characteristics independent of the respondent population [10]. This model enables the measurement of students' abilities and item quality on the same metric scale. Additionally, the Rasch Model provides detailed information on item difficulty, discrimination, and reliability, making it highly valuable for developing high-quality test instruments [11][12].

This study aims to analyze Physics test items for grade XI at SMAN 1 Sape, Bima Regency, using the Rasch Model. By focusing on validity, reliability, difficulty level, and discrimination, this study seeks to provide guidelines for developing improved test items. The analysis also offers insights for educators to enhance the quality of learning assessments, particularly in Physics. Moreover, this study supports efforts to improve data-driven assessment systems that are more meaningful and equitable. The novelty of this research lies in its application of the Rasch Model to analyze test items at the senior high school level in Bima Regency, which remains underexplored. The findings are expected to contribute significantly to the development of test instruments at both local and national levels. Additionally, this study

Volume 09, Number 01, Page 053-056
e-ISSN: 2598-120X; p-ISSN: 2598-117X

offers valuable input for policymakers in improving the quality of education, particularly in data-based learning evaluations. With more valid and reliable test items, assessment results will more accurately represent students' abilities, thereby supporting efforts to enhance educational quality sustainably.

II. RESEARCH METHODS

This study employed a descriptive quantitative method to analyze Physics test items for grade XI. The study aimed to evaluate the quality of test items in terms of validity, reliability, difficulty level, and discrimination power. This research design was chosen as it provides a detailed understanding of item characteristics based on student responses.

The subjects consisted of grade XI students at SMAN 1 Sape, Bima Regency. The participants were purposively selected to represent variations in student ability levels, school locations, and educational backgrounds. The data used were collected from midterm exam results in Physics. A total of 25 multiple-choice test items with three response options were used as the research instrument. Multiple-choice questions were chosen due to their structured nature, which facilitates the analysis of student responses.

Data collection was conducted through direct testing of students in selected schools. The testing was carried out under controlled conditions to ensure data validity. The responses obtained from student answer sheets were processed and evaluated using the Rasch Model approach. The analysis was conducted using WINSTEP software to gather information on item quality, including difficulty level, discrimination power, and item reliability. Data analysis in this study was performed quantitatively. The processed data provided insights into the performance of each test item based on Rasch Model parameters. The results were used to evaluate the quality of test items and provide recommendations for improving items that did not meet standard criteria. This analysis is expected to result in a more valid and reliable test instrument that supports high-quality learning assessments.

III. RESULTS AND DISCUSSION

Validity Testing Results

The analysis of item quality was conducted to evaluate the ability of test items to distinguish between parallel and series circuits based on validity, reliability, and difficulty levels. Item validity was assessed using the Rasch Model, which encompasses three main criteria: Outfit MNSQ (Mean Square), Outfit ZSTD (Z-Standard), and Point Measure Correlation (Pt Measure Corr) (Sumintono & Widhiarso, 2015). Items were considered valid if they met the following criteria: Outfit MNSQ values between 0.5 and 1.5, Outfit ZSTD values between -2.0 and +2.0, and Pt Measure Corr values between 0.4 and 0.85. If any criterion was not met, the item was classified as invalid.

Using WINSTEP software, the validity of each test item was analyzed based on these three criteria. Table 1 presents the results of the validity testing for the 25 analyzed items.

Table 1. Validity Analysis Results Using the Rasch Model

No	Result	Item Numbers
1	Valid	1,2,3,5,8,10,11,12,14,15,17,19,20,21,22,23,24,25
2	Invalid	4,6,7,9,13,16,18

The results revealed that 18 test items met all validity criteria and were deemed valid. Conversely, 7 items were classified as invalid because they failed to meet one or more criteria, such as Outfit MNSQ, Outfit ZSTD, or Pt Measure Corr. These findings indicate that a significant portion of the test items requires revision to be used as accurate assessment tools in evaluating learning outcomes. The Rasch Model provided more detailed and accurate results compared to conventional methods. Test items that met all three criteria were shown to consistently measure students' abilities and align with the intended constructs. Therefore, the Rasch Model is highly recommended for developing and evaluating assessment instruments in education.

Reliability of Test Items

The reliability analysis was conducted to evaluate the consistency of test items in measuring students' abilities using the Rasch Model. Reliability in the Rasch Model is assessed through two main indicators: Item Reliability and Person Reliability, which reflect the consistency of test items and students' responses, respectively (Sumintono & Widhiarso, 2015). Table 2 provides the interpretation criteria for reliability values based on the Rasch Model.

Table 2. Reliability Criteria in the Rasch Model

Nilai Reliability (person/item)	Interpretation
>0,94	Excellent
0,91-0,94	Very Good
0.81-0,90	Good
0,67-0,80	Fair
<0,67	Poor

The reliability analysis results, as shown in Table 3, indicate that the Item Reliability of the test items is 0.87. According to the Rasch Model reliability criteria, this value falls into the "Good" category, demonstrating an acceptable level of item consistency.

Table 3. Reliability Analysis Results of Test Items Using the Rasch Model

Item Reliability	Category
0,87	cukup

These findings indicate that the test items in this study exhibit a good level of reliability and consistency in measuring students' abilities. However, further development is necessary to optimize the consistency of some test items. Despite this, the current reliability level remains sufficient for analysis and further evaluation purposes.

Item Difficulty Levels

The analysis of item difficulty levels was performed using the Rasch Model. Difficulty levels were determined based on the measure logit values and standard deviation (SD) of the

Volume 09, Number 01, Page 053-056
e-ISSN: 2598-120X; p-ISSN: 2598-117X

item logits, which were subsequently categorized into four levels: very easy, easy, difficult, and very difficult. This categorization aims to identify the distribution of item difficulty levels in greater detail. Table 4 presents the criteria for item difficulty levels based on the Rasch Model.

Table 4. Criteria for Item Difficulty Levels in the Rasch Model

Nilai Measure (logit)	Difficulty Level
Measure logit > SD logit	Very Easy
-SD logit ≤ measure logit ≤ 0	Easy
0 ≤ measure logit ≤ SD logit	Difficult
Measure logit > SD logit	Very Difficult

The results of the item difficulty level analysis are shown in Table 5, illustrating the distribution of test items across each difficulty category.

Table 5. Distribution of Item Difficulty Levels Using the Rasch Model

Measure Value (Logit)	Difficulty Level	Number of Items
Measure logit > -1,02	Very Easy	2
-1,02 ≤ measure logit ≤ 0,0	Easy	3
0,0 ≤ measure logit ≤ 1,02	Difficult	15
Measure logit > 1,02	Very Difficult	3

Based on the analysis, 2 items were classified as very easy, 5 items as easy, 15 items as difficult, and 3 items as very difficult. This distribution indicates that most test items have relatively high difficulty levels. These findings suggest the need for evaluation and adjustment of items categorized as difficult and very difficult to ensure that the difficulty levels align with students' abilities. The Rasch Model provides detailed insights that support the improvement of the overall quality of assessment instruments.

This study revealed that most Physics test items for grade XI exhibit varying levels of quality in terms of validity, reliability, and difficulty. In terms of validity, only a few items met the validity criteria based on Rasch Model analysis, while the majority required improvement. These findings align [10], which highlights the Rasch Model's ability to more accurately identify valid test items compared to classical approaches. However, the presence of items that fail to meet the criteria underscores the need for revisions in item construction. Regarding reliability, the analysis showed that the instrument achieved a moderate level of consistency. This indicates that the test items have demonstrated adequate reliability in measuring students' abilities, although there remains room for improvement. [9] also emphasized the importance of reliability in instruments to ensure dependable evaluation results. While the current reliability is sufficient, efforts to enhance it through better item development are still necessary.

The difficulty level analysis revealed that most test items were categorized as difficult. This distribution is consistent [13], which indicated that items with high difficulty levels are often unsuitable for assessing students with varying levels of understanding. Additionally, overly difficult items may decrease students' motivation to respond [14]. Therefore, adjustments to the difficulty levels are needed to align the items with students' ability profiles. This study underscores the importance of using the Rasch Model to evaluate the quality of test items [15]. The model provides comprehensive insights into the validity, reliability, and difficulty of test items. These findings are relevant for educators and item developers to improve the quality of assessment instruments. However, the study also highlights the need for a solid understanding of techniques and data interpretation when employing the Rasch Model [11]. Thus, the results of this study can serve as a foundation for the development of better assessment instruments in the future.

IV. CONCLUSIONS

The results of this study indicate that, in terms of validity, only a small proportion of test items did not meet the criteria, while the rest require revisions. In terms of reliability, the instrument demonstrated moderate consistency in measuring students' abilities. The difficulty level analysis showed an uneven distribution, with the majority of items categorized as difficult. These findings highlight the need for further evaluation of the quality of assessment instruments to ensure their alignment with students' abilities and learning objectives. The implications of this study emphasize the necessity of training educators in constructing and analyzing test items using modern approaches such as the Rasch Model. Educators and policymakers are encouraged to prioritize validity, reliability, and difficulty levels in the development of assessment instruments. Additionally, invalid test items should be revised or replaced to enhance the quality of learning evaluation. This study also contributes to the advancement of educational assessment, particularly in utilizing Rasch Model-based data analysis to ensure that instruments used are more accurate and meaningful. This study has limitations, including its focus on a single school, which restricts the generalizability of its findings. Furthermore, the analysis only examined multiple-choice items without considering other types of questions that may be relevant. Future research can expand the scope to include multiple schools and diverse question types for a more comprehensive understanding. Developing technology-based applications that support Rasch Model analysis also presents an opportunity for future research to enhance the efficiency and effectiveness of test item evaluations in education.

REFERENCES

[1] Mardapi, D. (2017). Pengukuran Penilaian dan Evaluasi Pendidikan Edisi 2. Yogyakarta: Parama Publishing.
[2] Widyaningsih, S. W., & Yusuf, I. (2018). Analisis Soal Modul Laboratorium Fisika Sekolah I Menggunakan Racsh Model. Gravity: Jurnal Ilmiah Penelitian Dan Pembelajaran Fisika, 4(1).
[3] Sumarni, S. (2019). Designing Ict Competences-Integrated Assessment Instruments Of Practical Key Teaching Competences For English Language Education Study Program. Ijlecr-International Journal Of Language Education And Culture Review, 5(1), 47–55.

Volume 09, Number 01, Page 053-056
e-ISSN: 2598-120X; p-ISSN: 2598-117X

[4] Suryani, Y. E. (2017). Pemetaan Kualitas Empirik Soal Ujian Akhir Semester Pada Mata Pelajaran Bahasa Indonesia SMA di Kabupaten Klaten. Jurnal Penelitian Dan Evaluasi Pendidikan, 21(2), 142–152.

[5] Saryanto, S., Sumiharsono, R., Ramadhan, S., & Suprapto, E. (2020). The Analysis of Instrument Quality to Measure theStudents Higher Order Thinking Skill in Physics Learning. Turkish Journal of Science Education, 17(4), 520–527. https://doi.org/10.36681/tused.2020.42

[6] Wahyudi, W. (2010). Assesment Pembelajaran Berbasis Portofolio di Sekolah. Jurnal Visi Ilmu Pendidikan, 2(1).

[7] Meisya, R., Jannah, R., & Ramadhan, S. (2023). Analisis Kualitas Butir Soal Tematik Madrasah Ibtidaiyah Menggunakan Model Rasch. Al-Madrasah: Jurnal Pendidikan Madrasah Ibtidaiyah, 7(4), 1764. https://doi.org/10.35931/am.v7i4.2712

[8] Ramadhan, S., Mardapi, D., Kun, Z., & Budi, H. (2019). The Development of an Instrument to Measure the Higher Order Thinking Skill in Physics. European Journal of Educational Research, 8(3), 743–751. https://doi.org/10.12973/eu-jer.8.3.743

[9] Retnawati, H. (2014). Teori Respons Butir Dan Penerapannya: Untuk Peneliti, Praktisi Pengukuran Dan Pengujian, Mahasiswa Pascasarjana. Yogyakarta: Nuha Medika.

[10] Sumintono, B., & Widhiarso, W. (2015). Aplikasi Pemodelan Rasch Pada Assessment Pendidikan. Trim komunikata.

[11] Talib, A. M., Alomary, F. O., & Alwadi, H. F. (2018). Assessment Of Student Performance For Course Examination Using Rasch Measurement Model: A Case Study Of Information Technology Fundamentals course. Education Research International, 2018.

[12] Huda, N., Rizki, A., Oktavia, L., & Ramadhan, S. (2023). Pengembangan Instrumen Penilaian Sikap Disiplin Menggunakan Skala Likert Untuk Mengukur Sikap Disiplin Siswa Di Madrasah Ibtidaiyah. ELEMENTARY SCHOOL JOURNAL PGSD FIP UNIMED, 13(2), 136. https://doi.org/10.24114/esjpgsd.v13i2.42178

[13] Rahayu, R., & Djazari, M. (2016). Analisis Kualitas Soal Pra Ujian Nasional Mata Pelajaran Ekonomi Akuntansi. Jurnal Pendidikan Akuntansi Indonesia, 14(1).

[14] Iskandar, A., & Rizal, M. (2018). Analisis Kualitas Soal di Perguruan Tinggi Berbasis Aplikasi TAP. Jurnal Penelitian Dan Evaluasi Pendidikan, 22(1), 12–23.

[15] Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). Analisis Kualitas Soal Kemampuan Membedakan Rangkaian Seri Dan Paralel Melalui Teori Tes Klasik Dan Model Rasch. Indonesian Journal Of Educational Research and Review, 3(1), 11–19.