25 (2), 2024, 671-685

Jurnal Pendidikan MIPA

e-ISSN: 2685-5488 | p-ISSN: 1411-2531

http://jurnal.fkip.unila.ac.id/index.php/jpmipa/

Development and Validation of Physics Multiple-Choice Tests on the Nature of Physics Using Rasch Modelling Analysis

Sul Daeng Naba [1,*] , Edi Istiyono [1], Aris Kurniawan [1], & Nanang Adrianto [2]

[1]Department of Physics Education, Universitas Negeri Yogyakarta, Indonesia

[2]Division of Materials Science, Nara Institute of Science and Technology, Japan

Abstract: Assessment and evaluation of student learning outcomes are crucial aspects of education. Effective assessment requires well-designed evaluation test questions, one of them is multiple-choice questions. However, the quality of the multiple-choice questions made must be high. Quality questions are those that have undergone item analysis to assess the validity and reliability of the assessment results for students during the learning process. This study employed an evaluative research methodology with a descriptive quantitative approach. The research subjects involved 251 tenth grade students from two schools in Bombana Regency, Southeast Sulawesi Province. The evaluation was conducted using an instrument consisting of 10 physics questions that were created and validated with the assistance of teachers from each school. These questions covered the topics of the nature of physics, the scientific method, and work safety. The purpose of this study is to analyze the quality of the physics questions to determine whether they meet the criteria or not. Data collection was carried out using the QUEST software and analyzed using the Rasch model. The analysis results showed that the 10 questions had INFIT MNSQ values ranging from 0.77 to 1.33, which are consistent with the Rasch model. Furthermore, the OUTFIT t values for all item questions were obtained at values less than or equal to 2.00, indicating that the questions can be considered usable and passable. The reliability estimate score for the items is 0.98, indicating that this test has high consistency and can be relied upon to accurately measure students' learning outcome.

Keywords: cognitive assessment, rasch modelling, Item response theory.

▪ INTRODUCTION

Physics learning is considered successful and effective when students achieve good grades. Learning isn't solely determined by the intended curriculum; it's shaped by students' active engagement (Fischer et al., 2024; Goodyear et al., 2021). Learning outcomes refer to changes in students' abilities that occur after the learning process (Fadlilah et al., 2020; Anh & Phong, 2023). Learning outcomes are the results obtained by students after engaging in the learning process (Araujo et al., 2021; Benly et al., 2020). Based on these explanations, it can be concluded that learning outcomes are the scores given after the learning process to assess changes in students' abilities. However, in reality, the assessment of students' physics learning outcomes is not optimal (Chweu et al., 2024). Popham (2008), asserts that value assessments in higher education are often inadequate, vague, and disorganized.

Effective learning and student success depend on thorough evaluation and assessment (Anderson & Krathwohl, 2001). Quality assessments provide valuable insights into student mastery and teaching strategies (Popham, 2008; Kurniawan et al., 2024). Comprehensive item analysis is key to accurate assessment (Bond & Fox, 2007; Kurniawan et al., 2024). Item trials help create high-quality test questions, leading to more reliable and accurate measurements.

Sul Daeng Naba
*Email: suldaeng.2023@student.uny.ac.id

DOI: http://dx.doi.org/10.23960/jpmipa/v25i2.pp671-685
Received: 01 September 2024
Accepted: 23 September 2024
Published: 27 September 2024

There are several crucial aspects in education, one of which is learning outcomes assessment. One crucial aspect of education is evaluating students' learning outcomes, which serves to measure how well students understand the taught material (Achadah, 2019). Effective assessment requires good evaluation tools, one of which is multiple-choice tests (Hu et al., 2023; Justice et al., 2019). However, the quality of the test items used in multiple-choice tests is often not well analyzed, which can impact the validity and reliability of assessment results (Hedgeland et al., 2018). Some public school in Bombana, although multiple-choice tests are frequently used in assessing physics subjects, there have been no systematic efforts to analyze the quality of test items using current technology. This is supported by online interview results conducted with physics teachers, indicating that the evaluation of physics multiple-choice test items has not been done in a modern way by teachers. Typically, teachers only assess the creation of test items based on the alignment with the material and learning objectives (Abate & Mishore, 2024). Especially in physics topics related to broader concepts (Stojanovic & Maksimovic, 2022). This aligns with Land (2013), which show that a good grasp of physics concepts is related to students' ability to solve physics problems effectively. Understanding concepts is a crucial skill for students and must be mastered. By understanding concepts, students can expand their learning abilities and apply the concepts learned to real-life situations (Pennington, 2010). Conceptual understanding is a vital component of learning. It enables students to construct a more intricate cognitive structure and student learning outcome, facilitating the connection between concepts (Ozarslan & Cetin, 2018). One of these topics in the X grade physics material includes the nature of physics, scientific methods, and work safety. The importance of producing high-quality test items is one of the driving factors for students to be able to answer and evaluate the material that has been studied optimally (Ibnu et al., 2019). One supporting software for checking the quality of test items is QUEST software (Samila et al., 2019).

Item Response Theory (IRT) has proven to be an effective tool in item analysis as it can provide more detailed and accurate information compared to classical methods (Osterlind, 2006). IRT is a theoretical framework used to design, analyze, and assess the quality of tests and assessment instruments. In contrast to classical theory that focuses on total scores, IRT focuses on the relationship between individual abilities and question item characteristics (Hambleton et al., 1991). IRT is based on several important assumptions, they are unidimensionality, local independence, and parameter invariance (Baker & Kim, 2004). Rasch model is one popular model in IRT, which provides a strong framework for analyzing item characteristics. The model assumes that the probability of a correct answer depends only on the difference between individual ability and item difficulty (Rash, 1960; Embretson & Reise, 2013). The Rasch model is one of the simplest and most frequently used IRT models. In the context of grain quality analysis, the model provides important information such as item difficulty and item match to the model (Bond & Fox, 2007).

Previous research has employed the Rasch model to examine item bias in physics tests. DIF analysis using Rasch can identify biased items and inform the development of fairer assessments (Glamočić et al., 2022). Kurniawan et al (2024), found that the Rasch model was an effective tool for analyzing items measuring conceptual understanding in the domain of electromagnetic waves. Vera et al (2023), demonstrated the effectiveness of the Rasch model in developing an assessment instrument for electromagnetic wave problem-solving skills. Rahman et al (2023), found that the Rasch model was effective in determining the characteristics of the PABMMSB instrument. Zaidi et al (2023), showed the effectiveness of the Rasch model, using Winsteps software, to measure students' literacy abilities in the topic of global warming. Additionally, the Rasch model has been used to validate measurement instruments for computational thinking skills in physics education (Purnami et al., 2023; Hofer & Rubin, 2017). Research has shown the Rasch model's effectiveness in identifying non-compliant items and providing insights into student understanding (Planinic et al., 2010; Syadiah & Hamdu, 2020). The Rasch model has also been used to assess the quality of test items and detect bias in gender and domicile (Nisa et al., 2023; Tarigan et al., 2022). Based on several previous studies, none have conducted an evaluation assessment of the learning outcome test instrument on the material of the nature of physics, scientific methods, and occupational safety using the Rasch model with the support of QUEST software.

This study utilized the Rasch model to examine the learning outcome test for physics, scientific methods, and occupational safety. The Rasch model's ability to provide objective and reliable measurements made it ideal for analyzing student achievement. The model effectively differentiates item difficulty from student ability, offering a clearer understanding of students' conceptual grasp (Maulana et al., 2023). Additionally, the Rasch model excels in diagnostic analysis, as demonstrated by Wright maps, which visually represent student ability and item difficulty, enabling the identification of knowledge gaps (Popham, 2008; Sumintono & Widhiarso, 2015).

The use of QUEST software in item analysis is a progressive step that can assist teachers and researchers in evaluating and enhancing the quality of their assessment instruments (Safitri & Retnawati, 2020; Nugraha et al., 2020). QUEST is software designed for item analysis using the Rasch model, a model within the Item Response Theory (IRT) that provides a more in-depth and accurate approach compared to classical models (Pratama, 2020; Hofer & Rubin, 2017). The Rasch model enables the identification of invalid items, determines the difficulty level of items, and ensures that the test is reliable and valid (Ashraf & Author, 2020). At its core, the Rasch model is capable of evaluating both the items comprising a measurement instrument and the individuals being measured (Khairani & Razak, 2015; Matore, 2018). Through this model, teachers gain several benefits including aiding in proving the validity of the instrument used and providing more accurate measurements of students' abilities (Amelia et al., 2021; Bond & Fox, 2007; Simonetto, 2011). The lack of proper item analysis at at some public school in Bombana results in a lack of accurate information about the quality of test items and students' abilities. This has a negative impact on the effectiveness of assessment and efforts to improve learning because teachers cannot determine which items are too difficult or too easy, and which items may not measure physics concepts well.

The use of QUEST software and the Rasch model can provide a comprehensive solution to address this issue (Yilmaz, 2019) With detailed item analysis, teachers can obtain more accurate feedback on the quality of questions and students' abilities (Meyer & Zhu, 2013). This analysis also aids in developing better test items, which ultimately can improve the quality of physics learning at some public school in Bombana. This study aims to analyze the quality of multiple-choice physics questions on the topics of the nature of physics, scientific methods, and work safety for X grade students at some public school in Bombana using the Rasch model with the assistance of QUEST software. By conducting this analysis, it is expected to provide a clear overview of the quality of the test items used, identify items that need revision, and provide recommendations for future improvements.

▪ METHOD

Participants

The population of this study consists of all 10th-grade students in public high schools in Bombana. A random sampling technique was employed to select two classes as the sample: 10th grade at SMA Negeri 1 Bombana and 10th grade at SMA Negeri 3 Bombana. A total of nine classes, comprising 251 students, were included in the sample. With a student population distribution that is more heavily skewed towards females than males.

Research Design and Procedure

This study used an evaluative research methodology with a descriptive quantitative approach. Descriptive research with a quantitative approach aims to explain phenomena by systematically collecting and analyzing numerical data (Ali et al., 2022). The test instrument development process followed a modified version of the models proposed (Mardapi, 2008; Mahfudi & Istyono, 2024). The steps involved in creating the test instrument included designing the test, validating its content, conducting a trial run, and analyzing the collected data.

Instrument

The instrument used was a cognitive test with multiple choice format to measure students' learning outcomes on the topics of the nature of physics, scientific methods, and laboratory safety. The instrument consisted of 10 multiple-choice items. The instrument was validated by three content experts, and all test items were found to be valid. The result of the content validation are presented in Table 1.

Table 1. Content expert validation analysis

Item Soal	Validator			V	Kategori
	I	II	III		
Item 1	5	5	4	0.92	Extremely high
Item 2	5	5	4	0.92	Extremely high
Item 3	4	3	4	0.67	High
Item 4	5	5	5	1	Extremely high
Item 5	4	4	4	0.75	High
Item 6	5	5	5	1	Extremely high
Item 7	5	5	5	1	Extremely high
Item 8	4	4	4	0.75	High
Item 9	5	5	5	1	Extremely high
Item 10	4	5	4	0.83	Extremely high

The blueprint of the developed instrument, specifically the cognitive assessment item matrix, can be showh in Table 2.

Table 2. Indicators and cognitive level for test items

No.	Indicator	Item Numbers at Each Cognitive Level				Total Question
		C1	C2	C3	C4	
1	Explaining the concept of physics as a process, product, and attitude.	1. 4. 5				3
2	Applying physics concepts in the appropriate context.		6. 9			2
3	Analyzing the steps of the scientific method accurately.			3. 8		2
4	Identifying and explaining safety procedures in physics.				2. 7. 10	3

Data Analysis Techniques

The data analysis technique used the QUEST software with the Rasch model in Item Response Theory. Data analysis was conducted using the QUEST software, and the questions were considered to be of good quality if they met the criteria for item evaluation, which included the following stages: 1) item fit estimation; 2) difficulty level estimation; 3) item fit (conformity) estimation; and 4) reliability estimation (Hanna & Retnawati, 2022). According Heri Retnawati (2016), the stages of this research involved eight steps: 1) deciding how the instrument will be prepared; 2) finding theories relevant to the subject matter; 3) preparing item indicators for the instrument; 4) composing instrument items; 5) validating the instrument; 6) revising the instrument based on validator feedback; 7) conducting a test trial with respondents; and 8) performing the analysis.

▪ RESULT AND DISSCUSSION

The Content of Each Assessment Item that was Developed

The stages of this research have been conducted in steps 1-4, then continued to step 5 with validation by three validators. Step 6 involves revision based on feedback from validators and calculation of the item validity index (Aiken). Based on the Aiken index calculation, the research instrument shows moderate validity for the 10 test items used. The results of this calculation are interpreted using the criteria that the item validity index between 0.6-1 high to very high validity (Retnawati, 2016). Based on the Aiken V analysis, it can be concluded that all test items are highly suitable for assessing students' cognitive abilities.

Development of a multiple-choice cognitive test instrument to measure students' learning outcomes on the material of the nature of physics, the scientific method, and laboratory safety. The test consists of 10 items with cognitive levels ranging from C1 to C4. Items 1, 4, and 5 are at the C1 cognitive level with the indicator of explaining the concept of physics as a process, product, and attitude. Items 6 and 9 are at the C2 cognitive level with the indicator of applying physics concepts in the appropriate context. Items 3 and 8 are at the C3 cognitive level with the indicator of analyzing the steps of the scientific method accurately. Meanwhile, items 2, 7, and 10 are at the C4 cognitive level with the indicator of identifying and explaining safety procedures in physics. Based on the explanation, it can be concluded that the 10 test items fulfill the indicators of cognitive levels C1 to C4 with an even distribution across these levels. This finding aligns with Istiyono's (2020) assertion that cognitive levels C1 to C4 are designed to assess students' knowledge, comprehension, application, and analysis. Furthermore, cognitive assessment can be utilized to identify students' strengths and weaknesses. As Kolmos and Holgard (2007) suggest, feedback is crucial in the learning process. Consequently, the assessment results can be used to improve the quality of instruction.

Validity Analysis Using Rasch Modelling

Subsequently, in step 7, a trial was conducted with students. The data collection for this test item was carried out through a Google Form distributed to students via their respective class WhatsApp groups. The research was conducted with a sample of 9 classes consisting of 251 respondents, who were X grade students from SMAN 1 Bombana and SMAN 3 Bombana. Using QUEST software, the Rasch model was employed to evaluate response patterns of the respondents. The Rasch model was used to estimate variables such as item suitability, difficulty level, item fit, and reliability, with the aim of determining the quality within the Rasch model.

Estimation of Item Fit

As stated by Setyawarno (2017), INFIT MNSQ can be used to compare the determination of each item or item with model criteria where if the INFIT MNSQ value falls in the score range >1.33, it is considered irrelevant to the Rasch model, 0.77-1.33 is relevant to the Rasch model, and a value <0.77 is not relevant to the Rasch model. This range of values is used to assess item fit with the Rasch model using the QUEST software (Suryani, 2018;). Figure 1 shows the summary of the QUEST program's results for INFIT MNSQ values.

Item Estimates (Thresholds) In input Order
all on all (N = 251 L = 10 Probability Level= .50)

	ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1	item 1	215	250	-.76 .20	.92	.77	-.5	-1.1
2	item 2	195	250	-.17 .17	.98	.89	-.2	-.7
3	item 3	200	250	-.30 .17	1.08	1.00	.8	.0
4	item 4	210	250	-.59 .19	.99	.99	.0	.0
5	item 5	205	250	-.44 .18	1.02	.80	.2	-1.1
6	item 6	220	250	-.96 .22	1.03	.94	.2	-.2
7	item 7	75	250	2.04 .15	1.23	1.22	3.2	1.9
8	item 8	215	250	-.76 .20	.89	.64	-.8	-1.8
9	item 9	65	250	2.25 .15	1.05	1.08	.6	.7
10	item 10	200	250	-.30 .17	.92	.69	-.8	-2.1
Mean				.00	1.01	.90	.3	-.4
SD				1.16	.10	.18	1.2	1.2

Figure 1. Item recapitulation

In the previous description, it was explained that items relevant to the Rasch model and meeting the requirements fall within the range of INFIT MNSQ values between 0.77 and 1.33. Based on the analysis results, it is observed in Figure 1 that all item scores meet the requirements for the Rasch model. Another way to assess the fit of item scores with the Rasch model is illustrated in Figure 2, which shows the item fit map of the Rasch model.

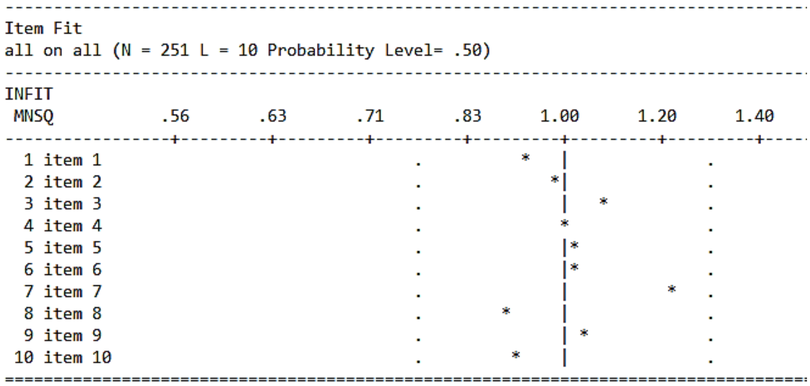

Figure 2. Fit map model rasch

Figure 2 shows that the points on the left lead to a value of 0.77 and the points on the right lead to a value of 1.33. Figure 2 shows that the points on the left lead to a value of 0.77 and the points on the right lead to a value of 1.33. These two points show the limits of suitability of the items included in the Rasch model. It can be noticed that none of the items cross the bricks of these points, and none are even in the same position as these boundary points. This shows that the item values that have fit the Rasch model are all 10 item values. This is because all item items are in the INFIT MNSQ value range.

Difficulty Level Estimation

The difficulty level of item scores can be measured using the QUEST software (Sarmila et al., 2019; Suyata et al., 2014). The range of criteria for checking the appropriateness of the difficulty level ranges from -2.0 to 2.0. An item score will be indicated as easy if the range value or student distribution is less than -2.0. Conversely, if the item score distribution is greater than 2.0, the item can be classified as difficult. The distribution of item difficulty levels is shown in Figure 3.

Figure 3 shows that the two most difficult item scores are questions numbers 9 and 7, while questions number 6 is the easiest item score but still falls within the moderate category. To determine the difficulty level of items using the QUEST program, you can refer to the range of item estimate thresholds (Pratama, 2020). he threshold value criteria are as follows: if b > 2 it is considered as very difficult criteria; if 1 < b ≤ 2 2 it is considered as difficult criteria, -1 < b ≤ 1 2 it is considered as moderate criteria, -1 < b "≥" -2 2 it is considered as easy criteria and b < -2 2 it is considered a very easy criteria (Heru & Suparno, 2019). The overall summary of the difficulty level for each item based on these criteria is shown in Table 3.

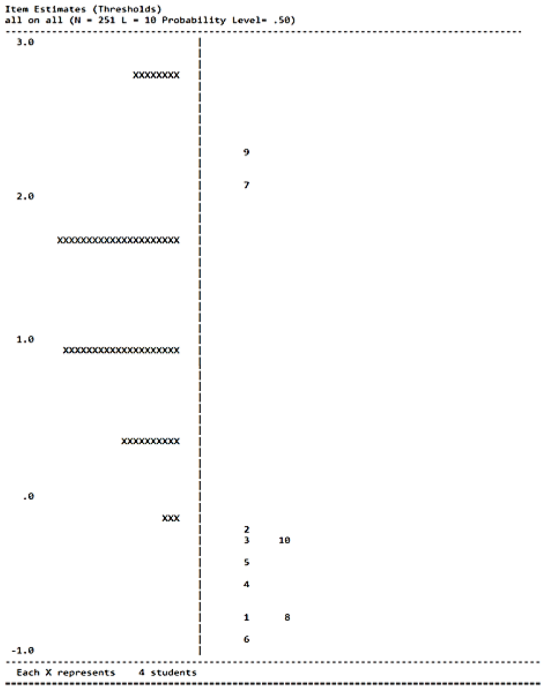

Table 3. Recapitulation of difficulty level in rasch model

Item	Threshold Value	Interpretation
1	-0.76	Moderate
2	-0.17	Moderate
3	-0.30	Moderate
4	-0.59	Moderate
5	-0.44	Moderate
6	-0.96	Moderate
7	2.04	Very difficult
8	-0.76	Moderate
9	2.25	Very difficult
10	-0.30	Moderate

Based on Table 3, which contains a summary of the difficulty level in the Rasch model, we can interpret the level of difficulty for each item based on the calculated threshold values. These threshold values depict the relative difficulty level of each item in the test, with higher values indicating more difficult items. In this table, most items have negative threshold values, interpreted as a "moderate" difficulty level. Items with threshold values of -0.76, -0.17, -0.30, -0.59, -0.44, -0.96, and -0.76 are all categorized as having moderate difficulty, indicating that these items are relatively easier for test takers. However, two items stand out with significantly positive threshold values, namely item 7 and item 9, with threshold values of 2.04 and 2.25, respectively. Both of these items are interpreted as “very difficult”. This indicates that these items are significantly more challenging compared to others and require further review to ensure they align with the test's objectives and the test-taker population. Extreme difficulty in item questions can influence the overall test results, and it’s important to adjust them to be more balanced with other items in the test.

Estimation of Passed (Fit) Items

The t OUTFIT value in the QUEST program is used to assess whether items in the test conform to the expected model. According to the research by (Langenfeld et al., 2020), an item is considered successful if the t OUTFIT value is less than or equal to 2.00, and fails if the t OUTFIT value is greater than or equal to 2.00. In Figure 1, the t OUTFIT values for each item are displayed. The fit components based on the t OUTFIT values are summarized as seen in Table 4. The use of these criteria ensures that the items in the test have sufficient validity to measure students' conceptual understanding abilities.

Table 4. Item fit recapitulation

Item	OUTFIT t Value	Description
1	-1.1	Passed
2	-0.7	Passed
3	-0.0	Passed
4	-0.0	Passed
5	-1.1	Passed
6	-0.2	Passed
7	1.9	Passed
8	-1.8	Passed
9	-0.7	Passed
10	-2.1	Passed

Table 4 shows a summary of t OUTFIT values for 10 items in the test, along with their descriptions. All items in the table are labeled "Passed," indicating that each item meets the passing criteria based on the t OUTFIT value. According to the standards mentioned earlier, an item is considered to pass if the t OUTFIT value is less than or equal to 2.00. The data shows t OUTFIT values for items ranging from -2.1 to 1.9. For example, item 1 has a t value of -1.1, item 2 has -0.7, item 3 has -0.0, and so on, with the highest t value at item 7 being 1.9. Although the t value for item 7 approaches the upper limit of 2.00, all items still meet the passing criteria. (Hidayatullah et al., 2022; Putri et al., 2016), indicating that all items in the test have good fit with the Rasch model, suggesting that these items are valid and sufficiently accurate for use.

Reliability Estimation

The QUEST program was used to calculate the reliability values in the Rasch model. Figure 4 shows the reliability estimation for the items. According to Langenfeld et al., (2020), reliability is a measure of the consistency of the results obtained from a test. A high reliability value indicates that the items in the test provide stable and dependable results when repeated under the same conditions (Ghazali, 2016). This estimation is crucial to ensure that the test instrument accurately measures the intended abilities without being influenced by external factors.

The data in the figure shows the analysis results of cognitive assessment using item estimation in the Rasch model. With a sample of 251 and 10 probability levels, this analysis provides an overview of the distribution and reliability of items in the cognitive test. The average (Mean) of item estimation is 0.00, indicating that overall the difficulty level of items in this test has been measured well and balanced around the average. The

```
-----------------------------------------------------------------------
Item Estimates (Thresholds)
all on all (N = 251 L = 10 Probability Level= .50)
-----------------------------------------------------------------------

Summary of item Estimates
=========================

Mean                              .00
SD                               1.16
SD (adjusted)                    1.14
Reliability of estimate           .98
```

Figure 4. Reliability of item estimate

standard deviation (SD) of 1.16 indicates variation in the difficulty levels of items in the test, reflecting diversity in the abilities measured by each item. The adjusted standard deviation (SD adjusted) is 1.14, which is almost the same as the original SD, indicating that the adjustment made did not significantly alter the original distribution of those items. Most importantly, the estimation reliability is very high, at 0.98.

According to Gleason et al. (2010) and Marambaawang et al. (2023), a reliability value approaching 1 indicates that the items in this test are highly consistent and reliable in measuring students' cognitive abilities. This high level of reliability suggests that the test has strong validity and that its results can be trusted for use in educational or psychological decision-making. In conclusion, this analysis demonstrates that the cognitive test used has items with varying but balanced difficulty levels and is highly reliable. With a high reliability value, it can be said that this test provides a consistent and accurate measure of the cognitive abilities of the test participants. This research aligns with previous findings by (Alfarisa & Purnama, 2019; Sinta et al., 2020), which emphasize the importance of reliability in measuring abilities through the Rasch model, ensuring that the measurement tools used provide consistent and valid results.

▪ CONCLUSION

This study aims to develop multiple-choice items for assessing students' learning outcome and to analyze the quality of multiple-choice physics items using the QUEST software with the Rasch model. Based on the results of the Aiken's V index analysis, the validity of the instrument for 10 items was categorized as high to very high. Therefore, it can be stated that all items are valid based on the assessment of content experts. Meanwhile, it was found that item fit estimation revealed that all items have INFIT MNSQ values within the appropriate range for the Rasch model, indicating that all items fit well with the model. All item scores fall within the range suitable for the Rasch model. Furthermore, based on the Rasch model's fit map analysis, no item scores exceed the boundaries of the Rasch model's range. This indicates that the item scores are in line with the Rasch model. Moreover, difficulty level estimation revealed that most items fall into the "moderate" difficulty category within the threshold value range of -0.17 to -0.96, with some items categorized as "very difficult" which indicated that most items have a moderate difficulty level. However, these very difficult items require special attention to ensure they do not compromise the overall validity of the test. Item fit estimation revealed that all item scores met the passing criteria of OUTFIT t-values indicating that these items have sufficient validity for use in the test. Reliability estimation revealed a very high item estimation, indicating that this test has high consistency and can be relied upon to accurately measure students' learning outcome. This research provides evidence that the use of QUEST software and the Rasch model in item analysis is an effective method to enhance the quality of assessment instruments. Through in-depth and systematic analysis, teachers and researchers can obtain more accurate feedback regarding the quality of items and students' learning outcome, as well as make necessary improvements to ensure that the tests used are truly valid and reliable. These results are expected to contribute to the development of better item questions in the future.

▪ REFERENCES

Abate, T., & Mishore, E. (2024). Alignment analysis between teacher-made tests with the learning objectives in a selected school of central regional state of Ethiopia. Heliyon, 10(11).

Achadah, A. (2019). *Evaluasi dalam pendidikan sebagai alat ukur hasil belajar* [Educational evaluation as a metric for student learning outcomes]. Jurnal An-Nuha, 6(1), 91-92.

Alfarisa, F., & Purnama. (2019). *Analisis butir soal ulangan akhir semester mata pelajaran ekonomi sma menggunakan rasch model* [A Rasch analysis of items from the high school economics end-of-semester examination]. Jurnal Pendidikan Ekonomi, 11(2), 366-368.

Ali, Mm., Hariyati, T., Yudestia Pratiwi, M., & Afifah. (2022). *Metodologi penelitian kuantitatif dan penerapannya dalam penelitian* [The quantitative research paradigm and its application in empirical research]. In Education Journal, 2(2).

Amelia, R. N., Sari, A. R. P., & Astuti, S. R. D. (2021). Assessment of chemistry learning: how is the quality of the tests made by the teacher. Journal of Educational Chemistry (JEC), 3(1).

Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition. Addison Wesley Longman, Inc.

Anh, T. T. N., & Phong, N. T. (2023). The effects of socrative-based online homework on learning outcomes in vietnam: a case study. International Journal of Interactive Mobile Technologies, 17(5), 182–199.

Araujo, I., Espinosa, T., Miller, K., & Mazur, E. (2021). Innovation in the teaching of introductory physics in higher education: the Applied Physics 50 course at Harvard University. Revista Brasileira de Ensino de Fisica, 4(3), 1–18.

Ashraf, Z. A., & Author, C. (2020). Classical and modern methods in item analysis of test tools assistant professor of clinical psychology, IMHANS, Kozhikode. International Journal of Research and Review, 7(5), 5.

Baker, F. B., & Kim, S. H. (2004). Item response theory: Parameter estimation techniques. CRC press.

Benly, M., Kartowagiran, B., Sukariasih, L., & Fayanto, S. (2020). Application of cooperative learning type group investigation to improve physics learning outcomes in vocational schools. Universal Journal of Educational Research, 8(10), 4618–4627.

Bond, T. G., & Fox, C. M. (2007). Applying the rasch model: fundamental measurement in the human sciences (second edition). Hong Kong: University of Toledo.

Chweu, E. M., Mnisi, S., & Mji, A. (2024). A Curricular framework for discipline-specific acquisition, teaching, and assessment of values in higher education. International Journal of Assessment and Evaluation, 31(2), 65–84.

D Mardapi. (2008). *Teknik penyusunan instrumen tes dan nontes.* Mitra Cendikia Press.

Fadlilah, N., Sulisworo, D., & Maruto, G. (2020). The effectiveness of a video-based laboratory on discovery learning to enhance learning outcomes. Universal Journal of Educational Research, 8(8), 3648–3654.

Fischer, J., Bearman, M., Boud, D., & Tai, J. (2024). How does assessment drive learning? A focus on students' development of evaluative judgement. Assessment and Evaluation in Higher Education, 49(2), 233–245.

Ghazali, N. H. M. (2016). A Reliability and validity of an instrument to evaluate the school based assessment system: a pilot study. International journal of evaluation and research in education, 5(2), 148-157.

Glamočić, D. S., & Mešić, V. (2022). A Rasch modeling approach to analyzing students' incorrect answers on multiple-choice questions: an example from wave optics. Metodički Ogledi, 29(1), 217-240.

Gleason, P. M., Harris, J., Sheean, P. M., Boushey, C. J., & Bruemmer. (2010). Publishing nutrition research: validity, reliability, and diagnostic test assessment in nutrition-related research. Journal of the American Dietetic Association, 110(3), 409–419.

Goodyear, P., Carvalho, L., & Yeoman, P. (2021). Activity-Centred analysis and design (ACAD): Core purposes, distinctive qualities and current developments. Educational Technology Research and Development, 69(2), 445–464.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Sage Publications.

Hanna, W. F., & Retnawati, H. (2022). *Analisis kualitas butir soal matematika menggunakan model rasch dengan bantuan software quest* [A Rasch analysis of mathematics test items using QUEST software to assess item quality]. Aksioma: Jurnal Program Studi Pendidikan Matematika, 11(4), 36-95.

Hedgeland, H., Dawkins, H., & Jordan, S. (2018). Investigating male bias in multiple choice questions: Contrasting formative and summative settings. European Journal of Physics, 39(5), 1-6.

Heru, M., & Suparno S. (2019). The development of reasoned multiple choice test in interactive physics mobile learning media (PMLM) of work and energy material to measure high school students' HOTS. Formatif: Jurnal Ilmiah Pendidikan MIPA, 9(2), 2-5.

Hidayatullah, H., Safitri, R., & Suyanto, S. (2022). *Model analisis butir instrumen tes biologi untuk penilaian akhir tahun menggunakan item response theory* [An Item Response Theory-based item analysis model for biology end-of-year assessments]. Measurement In Educational Research (Meter), 2(1), 1.

Hofer, S. I., Schumacher, R., & Rubin, H. (2017). The test of basic Mechanics Conceptual Understanding (bMCU): using Rasch analysis to develop and evaluate an efficient multiple choice test on Newton's mechanics. International journal of STEM education, 4, 1-20.

Hu, P., Li, Y., & Singh, C. (2023). Challenges in addressing student difficulties with measurement uncertainty of two-state quantum systems using a multiple-choice question sequence in online and in-person classes. European Journal of Physics, 44(1), 2-5.

Ibnu, M., Indriyani, B., Inayatullah, H., & Guntara, Y. (2019). *Aplikasi Rasch Model: Pengembangan instrumen tes untuk mengukur miskonsepsi mahasiswa pada materi mekanika* [The application of the Rasch model in developing a test instrument to measure student misconceptions in mechanics]. Jurnal Pendidikan FKIP, 2(1), 205–210.

Istiyono, E. (2020). *Pengembangan instrumen penilaian dan analisis hasil belajar fisika dengan teori tes klasik dan modern.* Yogyakarta : UNY Press

Justice, P., Marshman, E., & Singh, C. (2019). Improving student understanding of quantum mechanics underlying the Stern-Gerlach experiment using a research-validated multiple-choice question sequence. European Journal of Physics, 40(5), 2-3.

Khairani, A. Z., & Abd Razak, N. (2015). Modeling a multiple choice mathematics test with the Rasch model. Indian Journal of Science and Technology, 8(12), 1.

Kolomos, A., & J.E. Holgaard. (2007). Alignment of PBL and assessment. International Conference on Research in Higher Education. Honolulu : American Educational Research Association, 4(2), 1-9.

Kurniawan, A., Istiyono, E., & Daeng Naba, S. (2024). Item quality analysis of physics concept understanding test with rasch model. JIPF (Jurnal Ilmu Pendidikan Fisika), 9(3), 474-477.

Land, T. (2013). Conceptual understanding: The case of electricity and magnetism. European Journal of Science and Mathematics Education, 1(1), 13–28.

Langenfeld, T., Thomas, J., Zhu, R., & Morris, C. A. (2020). Integrating multiple sources of validity evidence for an assessment-based cognitive model. Journal of Educational Measurement, 57(2), 159–184.

Mafudi, I., & Istiyono, E. (2024). Development and validation of the relativity concept inventory test using item response theory generalized partial credit model 3PL. Jurnal Pendidikan MIPA, 25(1), 142–154.

Marambaawang, D. N., Oktoviana Bano, V., Rambu, R., & Enda, H. (2023). *Analisis kualitas butir soal penilaian akhir semester gasal tahun 2021/2022 menggunakan iteman di smp negeri 1 kambera* [An item analysis of the first semester final examination at SMP Negeri 1 Kambera in 2021/2022, utilizing the Iteman software to assess item quality] . Jurnal Undhari, 4(1), 233.

Matore, M. E. E. M., Maat, S. M., Affandi, H. M., & Khairani, A. Z. (2018). Assessment of psychometric properties for Raven Advanced Progressive Matrices in measuring intellectual quotient (IQ) using Rasch model. Asian Journal of Scientific Research, 11(3), 393-400.

Maulana, S., Rusilowati, A., Nugroho, S. E., & Susilaningsih, E. (2023). *Implementasi rasch model dalam pengembangan instrumen tes diagnostik* [The implementation of the Rasch model for the development of diagnostic test instruments]. In Prosiding Seminar Nasional Pascasarjana, 6(1), 748-757.

Meyer, J. P., & Zhu, S. (2013). Fair and equitable measurement of student learning in MOOCs: an introduction to item response theory, scale linking, and score equating. Journal Research & Practice in Assessment, 8(1).

Nisa, K., & Suprapto, N. (2023). *Deteksi bias gender dan domisili menggunakan DIF (differential item functioning): analisis instrumen tes keterampilan pemecahan masalah terintegrasi etnofisika* [An analysis of gender and domicile bias in an integrated ethno-physics problem-solving skills test using Differential Item Functioning]. Inovasi Pendidikan Fisika, 12(1), 30-35.

Nugraha, D. A., Cari, C., Suparmi, A., & Sunarno, W. (2019). Analysis of undergraduate student concept understanding-three-tier test: Simple harmonic motion on mass-spring system. AIP Conference Proceedings American Institute of Physics, 2(1), 1.

Osterlind, S. J. (2006). Modern measurement: Theory, principles, and applications of mental appraisal. Prentice Hall.

Ozarslan, M., & Çetin, G. (2018). Biology students' cognitive structures about basic components of living organisms. Science Education International, 29(2).

Pennington, M. C., & Black, S. L. (2010). Conceptual Understanding and Scientific Reasoning of High School Students in Physics. International Journal of Science Education, 32(5), 567–589.

Popham, W. J. (2008). Classroom assessment: What teachers need to know. Pearson.

Purnami, W., Fauzi, A., & Naingalis, M. L. P. (2023). Computational thinking skills identification among students of physics education department using Rasch model analysis. In AIP Conference Proceedings, 2751(1).

Putri, F. S., Istiyono, E., & Nurcahyanto, E. (2016). *Instrumen pengembangan berpikir kritis* [Critical Thinking Enhancement Instrument]. Unnes Physics Education Journal In UPEJ, 5(2).

Pratama, D. (2020). *Analisis kualitas tes buatan guru melalui pendekatan item response theory (IRT) model rasch* [An application of the Item Response Theory (IRT) Rasch model to analyze the quality of teacher-made tests]. Tarbawy : Jurnal Pendidikan Islam, 7(1), 61–70.

Rahman, A., Liliawati, W., & Rusdiana, D. (2023). Performance assessment with multiple intelligence differentiation to measure communication skills: application of many facet rasch model. Jurnal Pendidikan MIPA, 24(4), 932–943.

Rash, G. (1960). Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research.

Retnawati, H. (2016). *Analisis kuantitatif instrumen penilaian.* Yogyakarta: Pramana Publishing.

Safitri, A., & Retnawati, H. (2020). The estimation of mathematics literacy ability of junior high school students with partial credit model (pcm) scoring on quantity. Journal of Physics: Conference Series, 1581(1), 1-2.

Samila, A. P., Lastivka, G. I., & Tanasyuk, Y. V. (2019). Actual problems of computer parametric identification of the NMR and NQR spectra: A review. Journal of Nano- and Electronic Physics, 11(5), 1-2.

Setyawarno, D. (2017). *Penggunaan aplikasi software iteman (item and test analysis) untuk analisis butir soal pilihan ganda berdasarkan teori tes klasik* [An application of the Iteman software (Item and Test Analysis) in analyzing multiple-choice items based on the principles of Classical Test Theory]. Jurnal Ilmu Fisika dan Pembelajarannya, 1(1).

Simonetto, A. (2011). Using Structural Equation And Item Response Models To Assess Relationship Between Latent Traits. Journal of Apllied Quantitative Methods, 6(4), 44-45.

Sinta, T., Aprilia, N., Susilaningsih, E., & Priatmoko, S. (2020). Desain Instrumen Tes Pemahaman Konsep Berbasis Hot Dengan Analisis Model Rasch [The design and development of a test instrument to assess higher-order thinking skills related to conceptual understanding, with data analysis using the Rasch model]. Journal Chemistry in Education CiE, 9(2).

Stojanović, M., & Maksimović, B. (2022). Scientific concepts related to physics from the perspective of students of biology. Journal of Physics: Conference Series, 2288(1).

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

Suryani, Y. E. (2018). *Aplikasi rasch model dalam mengevaluasi Intelligenz Structure Test (IST)* [An application of the Rasch model for evaluating the Intelligenz Structure Test (IST)]. Psikohumaniora: Jurnal Penelitian Psikologi, 3(1), 73-100.

Suyata, P., Hidayanto, N., & Widyantoro, A. (2014). *Standarisasi instrumen integrated assessment hasil belajar bahasa dengan program quest* [The standardization of integrated assessment instruments for language learning outcomes using the quest program]. LITERA, 13(2), 364-367.

Syadiah, A. N., & Hamdu, G. (2020). *Analisis rasch untuk soal tes berpikir kritis pada pembelajaran STEM di sekolah dasar* [An application of the Rasch model for analyzing critical thinking test items in primary school STEM education]. Premiere Educandum: Jurnal Pendidikan Dasar dan Pembelajaran, 10(2), 138-148.

Tarigan, E. F., Nilmarito, S., Islamiyah, K., Darmana, A., & Suyanti, R. D. (2022). *Analisis instrumen tes menggunakan rasch model dan software SPSS 22.0* [An application of the rasch model and SPSS 22.0 software for test instrument analysis]. Jurnal Inovasi Pendidikan Kimia, 16(2), 92-96.

Vera, R., Kaniawati, I., Judhistira, &, & Utama, A. (n.d.). Using the rasch model to develop a measure of students' problem solving ability in optical instruments. Jurnal Pendidikan MIPA, 24(2), 419–431.

Yilmaz, H. B. (2019). A comparison of IRT model combinations for assessing fit in a mixed format elementary school science test. International Electronic Journal of Elementary Education, 11(5), 539–545.

Zaidi, M., Amiruddin, B., Samsudin, A., Suhandi, A., Kaniawati, I., Coştu, B., … Kuniawan, F. (n.d.). Validity and reliability of the global warming instrument: a pilot study using rasch model analysis. Jurnal Pendidikan MIPA, 24(4), 911–922.