Jurnal Pijar MIPA
https://jurnalfkip.unram.ac.id/index.php/JPM

ISSN 1907-1744 (Print)
ISSN 2460-1500 (Online)

Evaluation of HOTS Test on Renewable Energy Materials through Rasch Model Analysis

Nur Rahmatillah*, Abu Zainuddin

Department of Physics, Universitas Negeri Surabaya, Surabaya, Indonesia
*e-mail: run.rahmatillah@gmail.com

Received: June 25, 2025. Accepted: July 20, 2025. Published: July 30, 2025

Abstract: The growth of education in the globalisation era demands a paradigm change in the learning process, especially in terms of assessment that emphasises the importance of higher-order thinking skills (HOTS), which are no longer only focused on memorising and understanding basic concepts. This study aims to evaluate HOTS tests on renewable energy materials using the Rasch model as the main analysis technique. This research uses an evaluative design with a quantitative approach that aims to analyse the psychometric quality of HOTS test instruments objectively and systematically. The research subjects consisted of 71 learners selected through a purposive sampling technique from two classes in senior high school. The research instrument used was 10 multiple-choice questions with five answer options and analysed using Ministep software. The analysis results through the Wright map show that the distribution of students' abilities is close to a normal distribution. In contrast, the distribution of item difficulties has an uneven pattern, so that it does not cover the entire range of students' HOTS abilities. In addition, most of the items had adequate psychometric quality, but two items did not fit and needed special attention. Nevertheless, further analysis using the Item Characteristic Curve (ICC) revealed that item S3 indicated misfit, bringing the total to three items that did not fit the Rasch Model predictions. In addition, item S5 is included as a bias item. Overall, the instrument shows that the hierarchy of item difficulty is consistent and reliable, but it has limitations in sensitivity in distinguishing ability levels between learners. In addition, ICC analysis provides more sensitive detection of response pattern discrepancies, indicating the need for multiple analyses for comprehensive validation. Additional items are needed to comprehensively cover the spectrum of student abilities and improve the precision of identifying individual ability differences.

Keywords: HOTS; Ministep; Physics; Rasch Model.

Introduction

In the world of education, the need for test instruments that can provide meaningful information about learners' abilities is becoming increasingly crucial. Evaluation is an integral component in the learning process to determine the level of achievement of learning objectives and the basis for making appropriate educational decisions. Evaluation in a broad sense is a process of planning, obtaining, and providing information that is needed to make various alternative decisions [1]. Evaluation in education is defined as a structured method for assessing or measuring the learning process in various aspects of the world of education [2]. In this case, evaluation does not merely serve as a technical process of awarding grades, but includes comprehensive feedback to all elements of the education system for continuous improvement.

In the context of the development of educational measurement theory, quality evaluation requires measurement instruments that are carefully designed and analysed with appropriate methodologies. In reality, many test instruments still have unknown quality, resulting in pseudo-assessment. One of them is research by Adam et al., which revealed that of the 32 items developed, only 24 items met the valid criteria [3]. Similarly, research by Mappalesye et al. showed that of the 50 questions developed, 31 items were said to be valid [4]. These findings indicate a need for thorough analysis of test quality to produce accurate assessment data.

In an independent curriculum emphasising competence and higher order thinking skills, assessing Higher Order Thinking Skills (HOTS) is an essential component in the educational ecosystem. HOTS assessment includes cognitive levels C4 (analyse), C5 (evaluate), and C6 (create) by Bloom's revised taxonomy [5]. As explained by Ayumniyya & Setyarsih, HOTS is the ability to think, not limited to remembering facts, but demands a deeper meaning to get solutions to problems by analysing, evaluating, and or creating [6].

Although the reality on the ground shows various research studies on HOTS test instruments, the attention has never waned. However, the different understanding of HOTS and the variation in the quality of education between regions indicate the need for a measurement tool that cannot only evaluate students' abilities accurately, but can also accommodate the diversity of students' characteristics. The majority of research on HOTS test instruments relies on classical test theory (CTT), such as Saddia et al., Verdiana et al., and Fitriana et al., where, according to Sumintono & Widhiarso, CTT has limitations on sample characteristics [7], [8], [9], [10]. This means that the test group used for analysis will influence parameters such as the difficulty level of the questions. This makes the results of the analysis difficult to generalise to other populations.

How to Cite:
N. Rahmatillah and A. Zainuddin, "Evaluation of HOTS Test on Renewable Energy Materials through Rasch Model Analysis", *J. Pijar.MIPA*, vol. 20, no. 5, pp. 933–939, Jul. 2025. https://doi.org/10.29303/jpm.v20i5.9458

In recent decades, although classical test theory has aided test development and analysis, item response theory (IRT) has quickly become mainstream as the theoretical basis for measurement [11]. The main advantage of IRT compared to CTT-IA is that the item response theory model provides test and item statistics that are invariant to the population [12]. This was also expressed by Saepuzaman et al. that item response theory is a solution to the weaknesses of classical test theory because it has the concept of releasing the relationship between items and samples or test-taker subjects [13].

Various models have been developed from an IRT perspective, one of the simplest of which is often called the Rasch Model or one-parameter logistic (1PL) model [14]. Georg Rasch (1980) explained that the basis of the Rasch Model is a probabilistic model that implies two parameters: difficulty for each item and ability for each person [15]. Evaluation with Rasch modelling fulfils objective measurement because the data obtained is free from the influence of subject type, rater characteristics, and measuring instrument characteristics [10].

This is where the Rasch Model, a modern measurement model based on item response theory, offers a superior methodological solution. One of the parameters in the Rasch Model is the Wright Map, which shows the distribution of learner ability and item difficulty on the same scale. The Wright Map is derived from empirical analyses of learner data on a series of assessment tasks [16]. This feature facilitates substantive interpretation of learners' abilities regarding what they know and can do, and where they have difficulty.

The Rasch model is one of the statistical approaches in item response theory that is very useful in evaluating the quality of test instruments. In recent decades, the application of the Rasch Model in education, especially in evaluating HOTS test instruments, has covered various subjects, such as research by Kamilia in chemistry, Yudha in mathematics, and Irmayanti et al. on physics subjects [17], [18], [19].

Physics, as a scientific discipline that requires an in-depth understanding of abstract concepts, requires carefully constructed test instruments to accurately measure learners' higher-order thinking skills. In this study, the Rasch Model is applied to renewable energy materials relevant in today's education, given the global urgency towards energy transition and sustainable development. Renewable energy materials, including solar, wind, hydro, and biomass, present their complexity in learning, so it is necessary to measure the difficulties faced by learners. The previous research in HOTS item analysis generally uses the Classical Test Theory (CTT) approach, which is limited to descriptive analysis, such as Anggraeni et al.'s research on the subject of fluids [20]. The CTT approach has limitations because it cannot objectively identify item characteristics, detect item bias, or provide visualisation of the distribution of students' abilities against the difficulty level of the questions.

This research focuses on analysing HOTS test instruments using the Rasch Model on existing instruments to evaluate whether the items function according to the Rasch Model. Renewable energy material was chosen because of its relevance to the SDGs and energy literacy in the Merdeka Curriculum. It also filled the research gap for HOTS item analysis on strategic topics, but still limited instrument studies. In addition, this research aims to explore the potential integration of the Rasch Model in measurement methodology in physics so that the results of this research are expected to contribute to encouraging the transformation of conventional assessment practices towards more responsive digital assessments and enrich the scientific literature on the implementation of modern measurement.

Research Methods

This research uses an evaluative design with a quantitative approach that focuses on applying the Rasch Model as the main analysis technique in evaluating the quality of test instruments based on higher-order thinking skills. The Rasch model is an application of item response theory developed by Georg Rasch, which offers a probabilistic approach to analysing respondents' abilities and the difficulty level of test items on the same logit scale. The technique used to take the subjects of this research was purposive sampling. The subjects of this study were 2 grade X students in one of the public high schools in Surabaya, totalling 71 people as respondents to the HOTS test instrument.

The material of the higher-order thinking ability test tested on students is renewable energy, which amounts to 10 multiple-choice questions. The test instrument used has undergone a validation process conducted by three experts to ensure the items are completely at the Higher-Order Thinking Skills level. The validation process includes an assessment of the aspects of content, construct, and language clarity, which is carried out comprehensively. Validity was assessed using a Likert scale score of 1-4, with assessment criteria as in the following table [21].

Table 1. Validity Assessment Criteria

Score	Criteria
4	Very Valid
3	Valid
2	Moderately Valid
1	Invalid

The results of instrument validation by experts were then analysed to determine the percentage of instrument validity through the following equation.

$$p = \frac{f}{n} \times 100\% \quad \ldots(1)$$

Description:
p = percentage number of the questionnaire data
f = number of scores obtained
n = maximum number of scores

Based on the feasibility percentage obtained, it is then interpreted into several criteria as in the table below [22].

Table 2. Criteria for validity

Percentage (%)	Criteria
0-20	Invalid
21-40	Less Valid
41-60	Moderately Valid
61-80	Valid
81-100	Very Valid

The test results are in the form of scores, which are then analysed using Ministeps software, namely Wright Map analysis, item measure, item fit, item DIF, and summary

statistics, which are interpreted with the following provisions:

1. The Wright Map shows the distribution of learner ability and item difficulty on the same scale (logit).
2. Item measure is a grouping of question difficulty levels by combining the mean and standard deviation values based on the following table.

Table 3. Grouping Problem Difficulty Levels Based on the Rasch Model

Longit Range	Problem Difficulty Category
Measure logit $<$ $-$SD logit	Very easy
$-$SD logit $\leq$ Measure logit ≤ 0	Easy
$0 \leq$ Measure logit $\leq$ SD logit	Hard
SD logit $<$ Measure logit	Very hard

3. Item fit reveals the level of item fit based on three value criteria presented in the following table.

Table 4. Three Criteria for Item Fit Validity Values

Criteria	Accepted Value
Outfit MNSQ	$0.5 <$ MNSQ < 1.5
Outfit ZSTD	$-2 <$ ZSTD < 2
Pt Measure Correlation	$0.4 <$ Pt Measure Corr < 0.85

4. Item DIF or Differential Item Functioning detects the presence of biased items if the probability value is less than 0.05 or 5%.
5. Summary statistics show an overview of the instrument's complex yet informative qualities based on the following table.

Table 5. Instrument Quality Evaluation Guidelines

Criteria	Value	Category
Person Measure	Mean $>$ 0.0 logit	Participant's ability is greater than the difficulty level of the question.
	Mean $<$ 0.0 logit	Participant's ability is less than the difficulty level of the question
Alpha Cronbach	$\alpha < 0.5$	Bad
	$0.5 \leq \alpha \leq 0.6$	Poor
	$0.6 < \alpha \leq 0.7$	Fair
	$0.7 < \alpha \leq 0.8$	Good
	$\alpha > 0.8$	Excellent
Person-Item Realibility	Value $<$ 0.67	Weak
	0.67 $\leq$ Value $<$ 0.80	Fair
	0.80 $\leq$ Value $<$ 0.91	Good
	0.91 $\leq$ Value $\leq$ 0.94	Excellent
	Value $>$ 0.94	Special
INFIT & OUTFIT MNSQ (Person-Item)	1.0	Ideally
INFIT & OUTFIT ZSTD (Person-Item)	1.0	Ideally
Person-Item Separation	$H_{Person-item} = [(4 \times Separation) + 1]/3$	Person-item clustering

Results and Discussion

The HOTS test instrument used in this research refers to instruments that have been used previously in previous research. The HOTS test instrument in this research underwent a selection process to obtain 10 multiple-choice questions most representative of the original instrument. The selected questions were then modified by maintaining the characteristics of HOTS, which measure higher-order thinking skills, including (C4) analysing and (C5) evaluating skills. Modifications were made to ensure the questions remained at the HOTS level. The validation results show a very valid level of validity as shown in Figure 1. The high validation value indicates that the instrument has met the psychometric standards and is suitable for use.

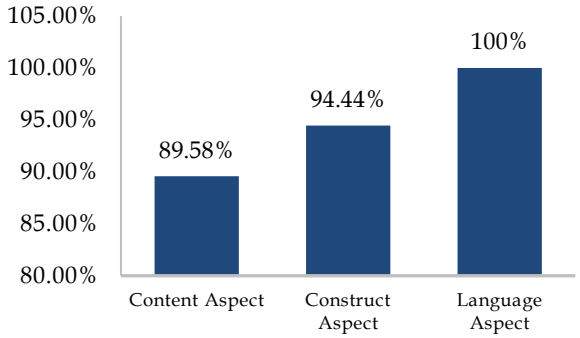


Figure 1. Result of HOTS Test Instrument Validation

Using the Rasch Model in item quality analysis offers a comprehensive approach to examine the characteristics of measurement instruments from two main perspectives, namely item functioning and respondent ability. This model allows researchers to examine in depth how each item measures the intended construct, including item difficulty, item fit to the measurement model, and item consistency in differentiating respondent ability.

One of the advantages of the Rasch Model is its ability to present the analysis results visually through a Wright map of the distribution of learner ability and item difficulty on the same logit continuum as shown in the following figure.

Based on Figure 2, there are 71 learners and 10 items spread on the logit bar from -3 to 3. Learners 21L and 48L have the highest ability, where both are outside the standard deviation limit (T). Learners 02P, 08P, 27L, 28L, 33L, 35P, 38L, and 56L have the lowest ability, as seen from their position at the bottom of the latent continuum. From Figure 2, the average logit person value is below the average logit item, which is below logit 0.0. Where the logit 0.0 value is formulated as the average item value [23]. This means that

the average HOTS ability of students is below the average difficulty level of the questions. The results of the Wright map visualisation show that the distribution of students is close to a normal distribution with a slight negative skew tendency, where it appears that most students are in moderate ability (around the mean person score).

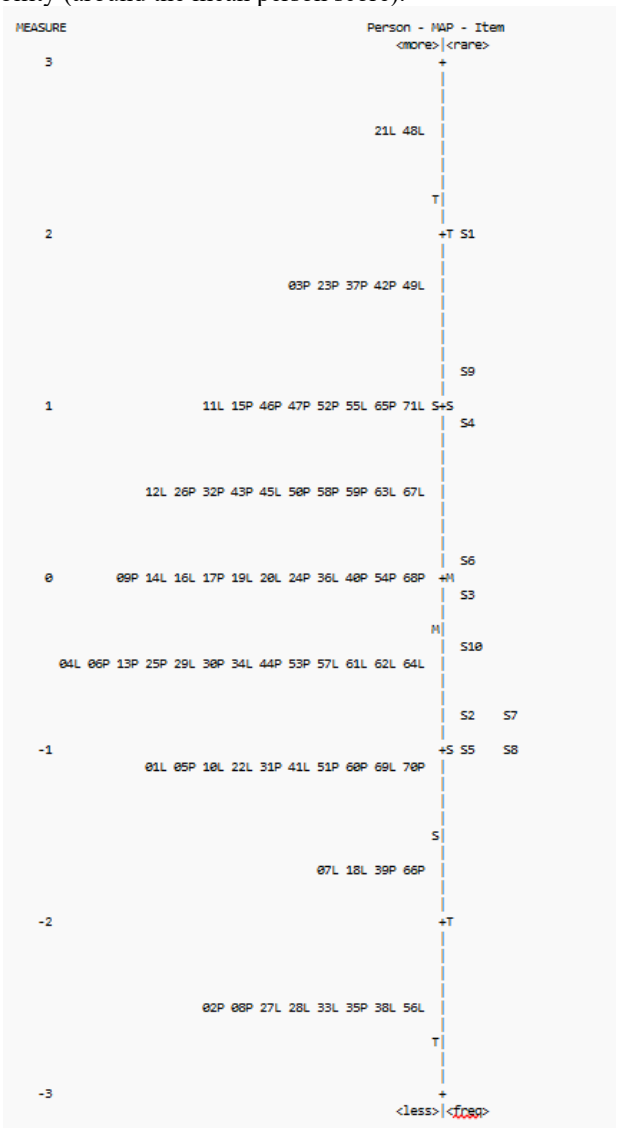
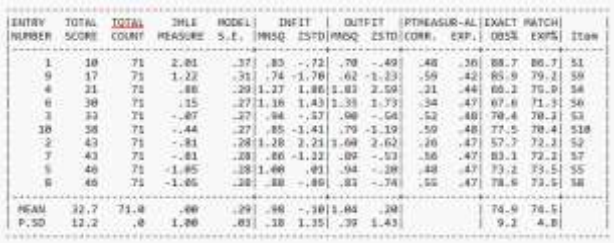


Figure 2. Wright Map Visualisation Results

Item S1 is the item with the highest level of ability, while items S5 and S8 are the items with the lowest level of ability. Although S1 was the most difficult item, the logit scores of learners 21L and 48L showed higher values. This indicates that the two learners did not find the items difficult enough to differentiate their HOTS ability. Likewise, items S5 and S8, as the easiest items, did not reach some low-ability learners. This indicates that the distribution of the ten items seems uneven because it does not reach the entire range of learners' HOTS abilities.



ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
1	10	71	2.01	.37	.85	-.72	.70	-.49	.48	.36	88.7	86.7	S1
9	17	71	1.22	.31	.74	-1.70	.62	-1.23	.59	.42	85.9	79.2	S9
4	21	71	.88	.29	1.17	1.86	1.03	2.59	.31	.44	80.3	76.9	S4
6	30	71	.15	.27	1.16	1.43	1.35	1.73	.34	.47	67.6	71.3	S6
3	33	71	-.07	.27	.94	-.57	.90	-.58	.52	.48	70.4	70.2	S3
10	38	71	-.44	.27	.85	-1.41	.79	-1.19	.59	.48	77.5	70.4	S10
2	43	71	-.81	.28	1.28	2.21	1.60	2.62	.26	.47	57.7	72.1	S2
7	43	71	-.81	.28	.86	-1.32	.89	-.53	.56	.47	83.1	72.1	S7
5	46	71	-1.05	.28	1.00	.01	.94	-.20	.48	.47	73.2	73.5	S5
8	46	71	-1.05	.28	.88	-.86	.83	-.74	.55	.47	78.6	73.5	S8
MEAN	32.7	71.0	.00	.29	.98	-.10	1.04	.20			74.9	76.5	
P.SD	12.2	.0	1.00	.03	.18	1.35	.39	1.43			9.2	4.8	

Figure 3. Measure Item Variation Distribution Results

Table 6. Grouping of Measure Item Variation

Longit Range	Problem Difficulty Category
Measure logit < -1	Very easy
$-1 \leq$ Measure logit ≤ 0	Easy
$0 \leq$ Measure logit ≤ 1	Hard
$1 <$ Measure logit	Very hard

However, the item measure can classify the difficulty level of questions from very easy to very difficult by combining the mean and standard deviation values. Based on Figure 3, the grouping of item measure variations is shown in Table 6 so that the level of item difficulty with each category can be seen in Table 7 below.

Table 7. Item Difficulty Level Results

Problem Difficulty Category	Question Item Code	Total
Very easy	S5, S8	2
Easy	S2, S3, S7, S10	4
Hard	S4, S6	2
Very hard	S1, S9	2

Based on Table 7, out of the 10 items analysed, 2 are very easy, indicating that most learners can answer the question correctly. Then, the other four questions are in the easy category, indicating that the items are still relatively mild but more challenging than the very easy items. Furthermore, there were two items categorised as difficult, which means that only a small proportion of learners were able to answer the question correctly. The items in this difficult category indicate that this question is above the HOTS ability of students. Furthermore, there are two items in the very difficult category, which indicates that both items can only be answered by students who have high HOTS abilities. The results of this item measure show that the instrument has a fairly good variation in difficulty levels. However, it is necessary to pay attention to the balance of proportions between categories of difficulty of questions that dominate in the 'easy' category. This is because balancing the question difficulty level is important to produce an instrument that can measure learners' abilities thoroughly and provide accurate information about learning achievement.

In addition to the question difficulty level, the Rasch Model analysis also provides information related to misfit items sorted from the most misfit, shown in the following table.

Table 8. Item Fit and Item Misfit Results

No Item	Outfit MNSQ	Outfit ZSTD	Pt Measure Corr	Explanation
S1	0.70	-0.49	0.48	Fit Item
S2	1.60	2.62	0.26	Misfit Item
S3	0.90	-0.54	0.52	Fit Item
S4	1.83	2.59	0.21	Misfit Item
S5	0.94	-0.20	0.48	Fit Item
S6	1.35	1.73	0.34	Fit Item
S7	0.89	-0.53	0.56	Fit Item
S8	0.83	-0.74	0.55	Fit Item
S9	0.62	-1.23	0.59	Fit Item
S10	0.79	-1.19	0.59	Fit Item

Based on Table 8, items S2 and S4 do not comply with the three criteria for the MNSQ outfit value, ZSTD outfit, and Pt measure corr, meaning that the two items are not predicted by the Rasch Model. While item S6 does not fulfil the Pt measure corr. criterion value so that it still tends to be said to be a fit item, so that the item is still retained. In contrast, items S2 and S4 must be replaced or deleted because they are unfit. This is supported by the results of the item characteristic curve in the following figure.

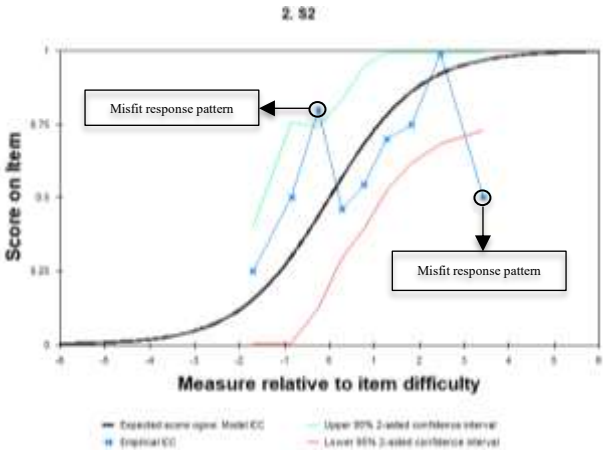


Figure 4. Item Characteristic Curves for Questions S2

The Item Characteristic Curve, often abbreviated as ICC, shows a graphical representation of the relationship between the participant's ability and the probability of answering an item correctly [24-25]. Based on Figure 4, the black sigmoid S curve is the ideal model line curve that illustrates how the probability of answering correctly increases as the participant's ability level increases. The results of the ICC graph on item S2 show the presence of 2 empirical points indicating a misfit response pattern. In addition, the fluctuation pattern up and down indicates the inconsistency of students' answers, which is in line with the results in Table 8 that item S2 is considered not fit with the Rasch Model predictions. This result can be seen from the black circles in Figure 4, which show that for high ability learners (around 3-4 logits), the probability of answering correctly is far below the ideal line curve. This means that this result contradicts the basis of Rasch modelling, which assumes that high-ability learners should have a high probability of answering correctly.

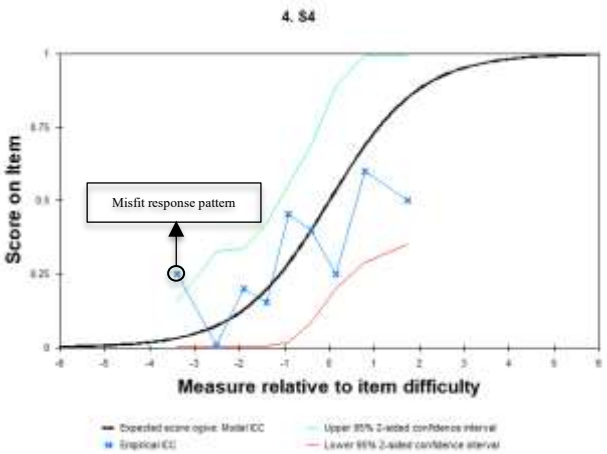


Figure 5. Item Characteristic Curves for Questions S4

In another case, the results of the ICC graph for item S4 show the most striking fluctuations across almost the entire ability range, with an up-and-down pattern that is highly inconsistent with model expectations. This is also evident in Figure 5 above, where the empirical point is outside the thin green top line, indicating that the item is unfit for Rasch modelling. This is consistent with the item fit and item misfit results in Table 8, which show item S4 is a misfit item.

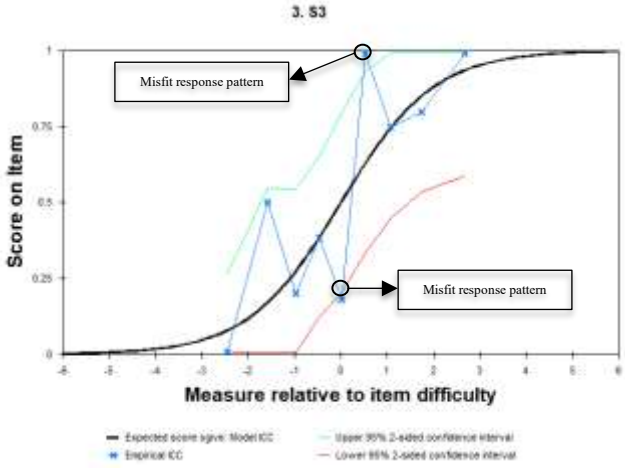


Figure 6. Item Characteristic Curves for Questions S3

The analysis results showed that 2 of the 10 multiple-choice items tested were identified as unfit items. However, through a more in-depth Item Characteristic Curve (ICC) analysis, one additional item showed a misfit pattern, namely item S3. The statistical analysis results on item fit stated that item S3 was fit with Rasch modelling. Despite this, the pattern of empirical data showed irregular fluctuations, and the indication of misfit indicated a problem with item S3 because it did not match the predictions of the Rasch Model. This finding indicates the need to revise or replace these items to improve the overall quality of the test instrument.

Another component of Rasch Model analysis is Differential Item Function (DIF), which refers to the situation when members of different groups (e.g., age, gender, culture) at the same level of a latent trait have different probabilities of responding to a particular item [26]. The research sample involved learners from two different classes, so through analysis, it was possible to identify items that showed differential functioning based on classmates.

DIF class/group specification is: DIF=$S4W1

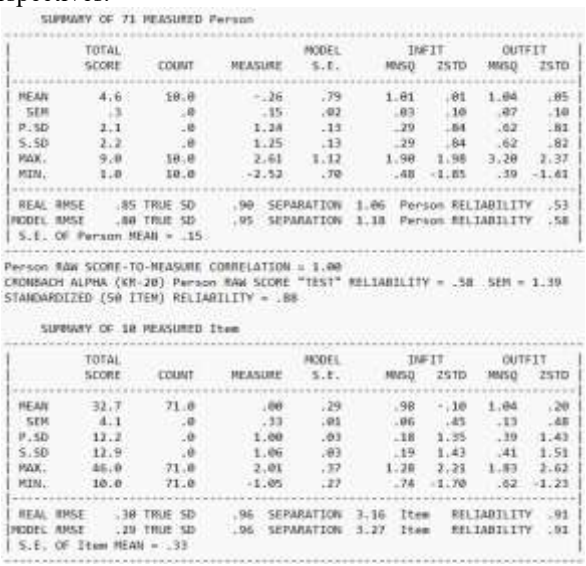
Person CLASSES	SUMMARY DIF CHI-SQUARED	D.F.	PROB.	BETWEEN-CLASS/GROUP Item UNWTD MNSQ	ZSTD	Number Name
2	.0119	1	.9130	.0118	-1.17	1 S1
2	1.6760	1	.1955	1.7323	.90	2 S2
2	.6019	1	.4378	.6186	.15	3 S3
2	1.5057	1	.2198	1.5541	.81	4 S4
2	5.0200	1	.0250	5.3786	2.07	5 S5
2	.3001	1	.5838	.3066	-.22	6 S6
2	.1181	1	.7311	.1206	-.60	7 S7
2	.0000	1	1.0000	.0028	-1.35	8 S8
2	.4830	1	.4871	.4930	.81	9 S9
2	.2224	1	.6372	.2252	-.36	10 S10

Figure 7. Output Items DIF between 2 Classes

Based on Figure 7, the Rasch Model analysis detected item S5 as biased because it has a probability value of 0.025 or less than 5%. In addition, items S2 and S4, which are unfit items but not biased items, can be indicated because both items are consistently problematic in both classes. Since both items were equally problematic in both classes, they did not

show bias towards a particular group. In contrast, item S5 fits the Rasch Model as a whole, but shows bias between classes. This indicates that item S5 measures HOTS ability well and consistently within each group, but the two classes have different levels of difficulty or interpretation. Therefore, this result indicates that the item favours one of the classes, so item S5 needs to be revised.

Overall, the following summary statistics show an overview of the Rasch Model analysis of measurement characteristics from both the learner (person) and item perspectives.



SUMMARY OF 71 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	4.6	10.0	-.26	.79	1.01	.01	1.04	.05
SEM	.3	.0	.15	.02	.03	.10	.07	.10
P.SD	2.1	.0	1.24	.13	.29	.84	.62	.81
S.SD	2.2	.0	1.25	.13	.29	.84	.62	.82
MAX.	9.0	10.0	2.61	1.12	1.98	1.98	3.20	2.37
MIN.	1.0	10.0	-2.52	.70	.48	-1.85	.39	-1.41

REAL RMSE	.85 TRUE SD	.90	SEPARATION	1.06	Person RELIABILITY	.53
MODEL RMSE	.80 TRUE SD	.95	SEPARATION	1.18	Person RELIABILITY	.58
S.E. OF Person MEAN = .15						

Person RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .58 SEM = 1.39
STANDARDIZED (50 ITEM) RELIABILITY = .88

SUMMARY OF 10 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	32.7	71.0	.00	.29	.98	-.10	1.04	.20
SEM	4.1	.0	.33	.01	.06	.45	.13	.48
P.SD	12.2	.0	1.00	.03	.18	1.35	.39	1.43
S.SD	12.9	.0	1.06	.03	.19	1.43	.41	1.51
MAX.	46.0	71.0	2.01	.37	1.28	2.23	1.83	2.62
MIN.	10.0	71.0	-1.05	.27	.74	-1.70	.62	-1.23

REAL RMSE	.30 TRUE SD	.96	SEPARATION	3.16	Item RELIABILITY	.91
MODEL RMSE	.29 TRUE SD	.96	SEPARATION	3.27	Item RELIABILITY	.91
S.E. OF Item MEAN = .33						

Figure 8. Output Summary Statistic

Based on Figure 8, the person measure results obtained a logit value of -0.26. This means that the average HOTS ability of students is below the average level of item difficulty. On the other hand, the Cronbach's alpha value of 0.58, which is in the poor category, indicates that the interaction between students' ability and item difficulty level is weak. This is confirmed by the person reliability value of 0.53 and person separation of 1.746, indicating that the instrument is less than ideal in consistently measuring students' HOTS ability.

Nevertheless, the item reliability result of 0.91 is included in the excellent category, which means that the items' difficulty level can be estimated well. This is consistent with the item separation result of 4.546, which shows that the items can be reliably divided into 4 to 5 levels. These results are supported by the INFIT and OUTFIT MNSQ values, which are close to 1, and the INFIT and OUTFIT ZSTD values are close to 0, indicating that the instrument is still of fairly good quality in terms of damage to the Rasch Model.

These psychometric results have interesting implications for interpreting and using instruments in measurement contexts. On the one hand, fit statistics showing ideal results and high item reliability will likely produce excellent measurements. On the other hand, low item reliability indicates that the instrument is insufficient to accurately measure learner ability differences. This condition reflects that most learners have abilities that are far below the range of item difficulty, which aligns with the results of the Wright map, where most learners' abilities are below the average level of item difficulty (mean person of -0.26 logit). Therefore, although this instrument fulfils the assumptions of the Rasch Model, it is limited in practical terms because it has limitations in measuring learners' abilities precisely.

Conclusion

Based on the research results, it can be summarised that the instrument has met the psychometric standards, where the items function by the assumptions of the Rasch Model, characterised by the high value of item reliability. Despite the high item reliability, which indicates good internal consistency between items, the low person reliability indicates that this instrument is less than optimal for individual ability measurement. This can be caused by the distribution of the difficulty level of the questions that are not yet optimal, so that they do not reach the entire range of students' HOTS abilities. However, there are items in the very difficult to very easy category. In reality, there are learners with high abilities that cannot be measured accurately because there are not enough challenging questions. Likewise, low-ability learners, those with logit scores that are significantly below the mean (shown on the Wright map), do not have items that can accurately reflect their position on the ability spectrum. Therefore, although the instrument shows items with well-estimated difficulty levels, it has weaknesses in the sensitivity of measuring individual ability. In addition, the Item Characteristic Curve (ICC) analysis results show a varied pattern of student responses to renewable energy HOTS items, providing in-depth insights into the quality of item discrimination and the suitability of the difficulty level to the target population's ability. Based on the results of Item Characteristic Curve (ICC) analysis and HOTS cognitive level evaluation, systematic revision of misfit items is required.

Implementing this revision is expected to optimise the instrument's ability to distinguish students based on their HOTS level of mastery more accurately and reliably. In addition to revising items that experience misfit, adding new items to cover the entire spectrum of student abilities, from low to high ability, is necessary. These items aim to create a finer and more comprehensive gradation of difficulty levels, so that the instrument can identify differences in student abilities with more precision. The expanded item distribution will allow for assessments more sensitive to variations in students' abilities and provide more detailed information.

Author's Contribution

Nur Rahmatillah: Conceptualisation of the research, data collection and analysis, writing the overall article; Abu Zainuddin: Academic supervision, methodology review, validation of result analysis, and quality check of manuscripts.

Acknowledgements

All praise and gratitude are due to Allah SWT for His mercy and grace so that this research can be completed properly. Thank you to the principal and teachers who have given permission and facilitated the process of research data acquisition. All students who have participated as respondents in this study are also acknowledged. Special appreciation also goes to the family for the moral and encouragement during the research and article writing process.

References

[1] R. Febriana, *Evaluasi Pembelajaran*. Jakarta: PT Bumi Aksara, 2019.

[2] M. Astuti, *Evaluasi Pendidikan*. Sleman: Penerbit Deepublish, 2022.

[3] D. A. Adam, Khaeruddin, and K. Arafah, "Pengembangan Instrumen Tes Hasil Belajar Kognitif Fisika Kelas XI MIPA SMA Negeri 2 Majene," *J. Sains dan Pendidik. Fis.*, vol. 19, no. 2, pp. 183–193, 2023, [Online]. Available: https://ojs.unm.ac.id/JSdPF/article/view/34499/23335

[4] N. Mappalesye, S. S. Sari, and K. Arafah, "Pengembangan Instrumen Tes Kemampuan Berpikir Kritis Dalam Pembelajaran Fisika," *J. Sains dan Pendidik. Fis.*, vol. 17, no. 1, pp. 69–82, 2021.

[5] L. W. Anderson and D. R. Krathwohl, *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Addison Wesley Longman, 2001.

[6] L. Ayumniyya and W. Setyarsih, "Profil Kemampuan Berpikir Tingkat Tinggi Siswa SMA dalam Pemecahan Masalah pada Materi Hukum Newton," *IPF Inov. Pendidik. Fis.*, vol. 10, no. 1, pp. 50–58, 2021, doi: 10.26740/ipf.v10n1.p50-58.

[7] A. Saddia, Sutrisno, M. Saldi, and M. N. Agriawan, "Analisis Kemampuan Menyelesaikan Soal Hots Fisika Siswa Sma Di Kota Majene," *J. Fis. dan Pembelajarannya*, vol. 4, no. 1, pp. 1–5, 2021, doi: 10.31605/phy.v4i1.1275.

[8] V. Verdiana, H. Munawaroh, and F. Fatiatun, "Analisis Peningkatan Hasil Belajar Siswa pada Mata Pelajaran Fisika pada penerapan Model Pembelajaran PBL Menggunakan Soal HOTS," *Biochephy J. Sci. Educ.*, vol. 4, no. 1, pp. 70–74, 2024, doi: 10.52562/biochephy.v4i1.981.

[9] Fitriana, S. S. Sitompul, and M. M. S. H., "Analisis Kemampuan Kognitif Peserta Didik dalam Menyelesaikan Soal HOTS Fisika Materi Getaran Harmonis," *J. Dunia Pendidik.*, vol. 5, no. 2, pp. 553–565, 2024, [Online]. Available: http://jurnal.stokbinaguna.ac.id/index.php/JURDIP/article/view/2083

[10] B. Sumintono and W. Widhiarso, *Aplikasi Model Rasch untuk Penelitian Ilmu-ilmu Sosial*, Edisi Revi. Cimahi: Trim Komunikata Publishing House, 2014.

[11] S. E. Embretson and S. P. Reise, *Item Response Theory For Psychologists*. New Jersey: Lawrence Erlbaum Associates, Inc, 2000.

[12] K. S. Shultz, D. J. Whitney, and M. J. Zickar, *Measurement Theory in Action Case Studies and Exercises*, Third. New York: Routledge, 2021.

[13] D. Saepuzaman, E. Istiyono, H. Haryanto, H. Retnawati, and Y. Yustiandi, "Analisis Karakteristik Butir Soal Fisika Dengan Pendekatan IRT Penskoran Dikotomus dan Politomus," *Radiasi J. Berk. Pendidik. Fis.*, vol. 14, no. 2, pp. 62–75, 2021, doi: 10.37729/radiasi.v14i2.1200.

[14] R. M. Furr and V. R. Bacharach, *Psychometrics An Introduction*. California: Sage Publications, Inc, 2008.

[15] N. L. A. Kassim, *The Rasch Model for Psychometric Testing Rater-Mediated Assessment and Standard Setting*. Kuala Lumpur: IIUM Press, 2024.

[16] M. Wilson and K. Draney, "A Strategy for the Assessment of Competencies in Higher Education: The BEAR Assessment System," in *Modeling and Measuring Competencies in Higher Education*, Rotterdam: Sense Publishers, 2013.

[17] L. Kamilia, "Analisis Soal HOTS Materi Asam Basa untuk Mengukur Keterampilan Berpikir Kritis dengan Rasch Model," *Pentagon J. Mat. dan Ilmu Pengetah. Alam*, vol. 3, no. 1, pp. 99–111, 2025, doi: https://doi.org/10.62383/pentagon.v3i1.411.

[18] R. P. Yudha, "Higher Order Thinking Skills (HOTS) Test Instrument: Validity and Reliability Analysis with The Rasch Model," *EduMa Math. Educ. Learn. Teach.*, vol. 12, no. 1, pp. 21–38, 2023, doi: 10.24235/eduma.v12i1.9468.

[19] R. Irmayanti, M. Rusdi, and Yusnaidar, "The Rasch Model: Implementation of Physics Learning Evaluation Instrument Based on Higher Order Thinking Skills," *Integr. Sci. Educ. J.*, vol. 4, no. 2, pp. 62–68, 2023, doi: 10.37251/isej.v4i2.325.

[20] C. D. Anggraeni, M. Junus, and P. Damayanti, "Analisis Soal Latihan pada Buku Soal Fisika Kelas XI Berdasarkan Taksonomi Bloom Revisi Dilihat dari Prespektif Higher Order Thinking Skill pada Pokok Bahasan Fluida," *J. Literasi Pendidik. Fis.*, vol. 4, no. 1, pp. 40–51, 2023, doi: 10.30872/jlpf.v4i1.1945.

[21] Sugiyono, *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*. Bandung: CV. Alfabeta, 2017.

[22] Riduwan, *Dasar-dasar Statistika*. Bandung: Alfabeta, 2015.

[23] B. Sumintono and W. Widhiarso, *Aplikasi Pemodelan Rasch Pada Assessment Pendidikan*. Cimahi: Penerbit Trim Komunikata, 2015.

[24] J. Irawan, S. Hadi, Z. Zulandri, J. Jamaluddin, A. Syukur, and S. Hadisaputra, "Validating metacognitive awareness inventory (MAI) in chemistry learning for senior high school: A rasch model analysis," *J. Pijar MIPA*, vol. 16, no. 4, pp. 442–448, 2021. [Online]. Available: https://doi.org/10.29303/jpm.v16i4.2603

[25] S. E. Stemler and A. Naples, "Rasch Measurement v. Item Response Theory: Knowing When to Cross the Line," *Pract. Assessment, Res. Eval.*, vol. 26, pp. 1–16, 2021, doi: 10.7275/v2gd-4441.

[26] S. P. Reise and D. A. Revicki, *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. New York: Routledge, 2015.