Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 3, 346-356
2026
Publisher: Learning Gate
DOI: 10.55214/2576-8484.v10i3.12387

Item analysis of a reading comprehension test using the Rasch model for EFL learners

Zakiyah Zakiyah[1*], Nur Mukminatien[2], Mirjam Anugerahwati[3], Nunung Suryati[4]
[1]Doctorate Program in ELT, Faculty of Letters, Universitas Negeri Malang, Indonesia; zakiyah.2202219@students.um.ac.id (Z.Z.).
[2,3,4]Department of English, Faculty of Letters, Universitas Negeri Malang, Indonesia; nur.mukminatien.fs@um.ac.id (N.M.) mirjam.anugerahwati.fs@um.ac.id (M.A.) nunung.suryati.fs@um.ac.id (N.S.).

Abstract: This study aimed to analyze the quality of a reading comprehension test using Rasch measurement. A quantitative method with a descriptive design was employed, involving 200 eleventh-grade students from a senior high school in Sukoharjo, Central Java, Indonesia. The instrument consisted of 29 multiple-choice reading comprehension items and was piloted from June to July 2025. Data were collected through test administration and analyzed using Quest software based on the Rasch model. The results showed that the instrument demonstrated good reliability, with item reliability of 0.89 and person reliability of 0.87. Item fit analysis indicated that 13 items fit the Rasch model, while 16 items were misfitting and required revision. Most items were classified as having a medium level of difficulty (93.10%), and the majority of students were in the high reading comprehension ability category (70%). Overall, the instrument showed satisfactory measurement quality; however, several items need refinement to improve the validity and precision of the test. This study provides empirical support for the utility of Rasch analysis in test evaluation and offers a basis for future improvement of reading comprehension assessments.

Keywords: *Item reliability, Measurement quality, Rasch model, Reading comprehension, Test item analysis.*

1. Introduction

Every learning process undertaken by a teacher over a period of time must be assessed to determine the extent to which it improves students' abilities or provides added value. Tests are a measurement tool, gathering information on the characteristics of an object. This object can be a student's abilities, attitudes, interests, or motivation [1]. In the educational world, these tests are called learning outcome tests. Learning outcome tests are one of the most widely used measurement tools to determine an individual's learning outcomes in the teaching and learning process or an educational program. The term "test" is often defined as a specific assessment tool, simply expressed in the form of questions that reveal behavior, potential, or are related to learning outcomes [2]. Tests are standardized and objective tools, allowing them to be widely used to assess and compare individuals' psychological states or behaviors. For example, every test given by a teacher to students requires a response from the subject being assessed.

As a tool for measuring student abilities after participating in educational activities for a specific period of time, the existence of tests is crucial [3]. A good test will reveal a student's true condition, while a poor test will fail to reveal their true abilities. Therefore, it is important to identify each item before it is used to measure student ability. Teachers routinely identify and analyze items to improve, revise, and refine them. This way, teachers can ensure that future learning outcome tests they design can truly serve as high-quality learning outcome measurement tools. Furthermore, item analysis is conducted to determine whether a question is functioning effectively.

History: Received: 9 January 2026; Revised: 16 February 2026; Accepted: 20 February 2026; Published: 10 March 2026
* Correspondence: zakiyah.2202219@students.um.ac.id

Item analysis is generally conducted in two ways: qualitative and quantitative [4]. Qualitative item analysis is conducted based on the rules of item writing. This review is usually conducted before the questions are used or tested. Aspects considered in qualitative review include material, construction, language or culture, and answer keys. Quantitative item analysis is the examination of test items based on empirical evidence. One of the main objectives of empirically testing test items is to determine the extent to which each test item differentiates between high-ability and low-ability students.

In the field of educational measurement, there are two approaches frequently used to analyze test quality: Classical Test Theory (CTT) and Item Response Theory (IRT). However, test analysis using the CTT approach has been largely abandoned due to its numerous weaknesses. According to Istiyono et al. [5], CTT has at least two weaknesses: 1) measurement results depend on the characteristics of the test used; 2) item parameters depend on the test taker's abilities; and 3) measurement errors can only be detected for groups, not individuals. Regarding the weaknesses of CTT, Wardhani and Putra [6] stated that CTT is less able to describe or reflect the actual abilities of test takers. This assumption is based on student abilities being measured solely from the total score obtained, ignoring the correlation between test-taker abilities and item characteristics. This differs from the IRT approach, which assumes that the probability of test takers answering each item correctly depends on their abilities. Thus, test takers with high ability have a greater chance of answering correctly than test takers with low ability [7, 8] stated that three assumptions form the basis of IRT, namely, unidimensionality, local independence, and parameter invariance. According to Hambleton et al. [8], unidimensionality means that each test item only measures one ability. For example, a reading achievement test. This means that the test can only determine the test taker's ability in reading and not other abilities. However, in practice, this assumption is difficult to implement due to several influencing factors such as cognitive factors, personality, environment, or anxiety. The assumption of local independence in this case states that there is no relationship between the test taker's response to different test items. Meanwhile, parameter invariance states that the characteristics of the test items do not depend on the distribution of test-taker parameters, and the parameters that characterize the test-takers do not depend on the characteristics of the items [9].

Hambleton and Jones [10] mention several advantages of IRT, namely: 1) the score describes the ability of the test taker and does not depend on the difficulty of the test, 2) it can be used to relate the test items to the ability of the test taker, and 3) it does not require parallel tests to determine the reliability coefficient. According to Istiyono, et al. [5] one of the simple models and has been widely used by experts in developing a test is the Rasch model, with 1 parameter (1-P) In addition, the selection of the Rasch model because this model has at least fulfilled the principles of the measurement model, namely; 1) this model can provide linear measurements with the same interval, 2) can overcome the problem of missing data, 3) can provide more precise estimates, 4) can detect the inaccuracy of a model, and 5) provides a measurement instrument that is independent of the parameters studied [11].

Furthermore, Istiyono et al. [5] stated that measurement in the Rasch model is a direct comparison between individuals and items. In this case, the individual is the test taker's ability, and the item is the parameter of the level of difficulty. Therefore, under certain conditions, for example, if the test taker's ability increases, the chance of answering the test item correctly becomes greater. Therefore, the chance of answering the test item correctly refers to two things: the test taker's ability and the item's difficulty level. Regarding the basic concept of the Rash model, Sumintono [11] also stated the same thing, that the Rash model is a measurement model that determines the relationship between the test taker's ability and the level of difficulty of the test item. Furthermore, Sumintono [11] illustrates this relationship; for example, if a test taker can answer 80% of the questions correctly, it certainly has a higher ability than a test taker who is only able to answer 60% of the questions correctly. The Rash model in the application of test instrument analysis has advantages over other models, such as CTT. According to Sumintono [11], the Rash model can predict missing data based on individual response patterns. This is what makes the Rash analysis model more accurate. In addition, the Rash model can also produce a standard error measurement score for the instrument used. In addition, the Rash model also calibrates three

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 3: 346-356, 2026
DOI: 10.55214/2576-8484.v10i3.12387

things, namely, the measurement scale, test participants (person), and items. Calibration is intended to ensure the validity and reliability of measurement results so that the test can provide comprehensive information [12].

According to Istiyono et al. [5], the analysis of test instruments using the Rash model can be done through the following steps: 1) assessing item fit statistics. This stage is the stage to determine items that match the Rasch model. If there are items that do not match, they can be removed. 2) assessing person fit statistics. This stage determines which test participants match the Rasch model. 3) Determining which items and test participants (person) match the Rasch model through goodness of fit analysis. Research on test quality analysis has been widely conducted, especially in the field of education. From the results of the search, most are more dominant using the CTT approach. As explained above, the CTT approach still has weaknesses. So the CTT approach has begun to be abandoned and switched to the IRT approach. Based on the description above, this study aims to describe the characteristics of reading comprehension test items, including reliability, item fit, item difficulty level, and student ability.

2. Literature Review

2.1. Reading Comprehension in EFL Learning

Reading comprehension is an essential skill in EFL classes. In this process, students are not only required to recognize individual words but also to understand the meaning contained in a text as a whole. According to Grabe [13], reading comprehension involves complex cognitive processes, including word decoding, syntactic structure recognition, and integration of information with background knowledge. Decoding is the process of converting written symbols into sounds and meanings, which are the basis for understanding a text Nation [14]. Mastery of syntactic structures helps readers recognize the relationships between words in sentences, so that they can understand the meaning of a text as a whole. In addition, the integration of information with background knowledge allows readers to connect new information with previous experiences or understandings, which can improve memory and a deeper understanding of the text [15].

In the EFL context, challenges in reading comprehension become more significant because students have to overcome linguistic and cultural barriers that are different from their mother tongue. One of the main barriers is limited vocabulary, which causes difficulties in understanding key words in the text [16]. In addition, the syntactic structure in English is often different from the students' mother tongue, which can lead to errors in understanding the relationship between ideas in the text. Not only that, but cultural differences also contribute to the level of reading difficulties of EFL students. According to Koda [17], EFL readers often face difficulties in accessing different cultural schemas, which can hinder in-depth understanding of the text. For example, texts containing certain cultural references may be difficult to understand for readers who do not have the same cultural background. Therefore, reading comprehension in EFL requires not only linguistic skills but also the ability to navigate cultural differences and contexts. This suggests that cognitive and affective aspects must be considered in developing EFL students' reading comprehension skills.

2.2. The Rasch Model

The Rasch Model is a measurement model within IRT that examines the relationship between test takers' ability and item difficulty on a common scale. The model assumes that the probability of a correct response depends on the interaction between a learner's ability and the difficulty level of a test item [18]. A key assumption of the Rasch Model is unidimensionality, which means that all items in a test are expected to measure a single underlying construct. In reading comprehension assessment, this construct refers to learners' ability to understand written texts. Another important assumption is local independence, indicating that responses to individual items are independent of one another once the latent trait has been accounted for [19].

The Rasch Model is widely used in item analysis because it provides detailed information about item quality, including item difficulty estimates, item fit statistics, and person and item reliability. According

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 3: 346-356, 2026
DOI: 10.55214/2576-8484.v10i3.12387

to Bond and Sumintono [20] that do not fit the Rasch Model may weaken the validity of measurement and therefore should be reviewed or revised. In the context of EFL assessment, Rasch analysis is particularly valuable for evaluating reading comprehension tests, as it allows researchers to determine whether test items function consistently across different levels of learner ability. This contributes to the development of valid and reliable reading assessment instruments for EFL learners [21].

3. Research Methodology

This study used a quantitative method with a descriptive design to analyze the quality of teacher-created test items. The study was conducted with 200 students from a high school in Sukoharjo, Central Java, Indonesia. The students participating in the study were in grade 11. The pilot test was conducted from June to July 2025. The instrument used in this study was a learning outcome test covering English, specifically reading comprehension. The instrument consisted of 29 multiple-choice items. The research procedure began with data collection through the administration of tests to all students in the sample. After the data was collected, analysis was conducted using Quest software. Adam and Khoo [22] explain that Quest is an application that offers comprehensive Rasch analysis, including item estimation, case estimation, and fit statistics. The results of the analysis in the Quest application can be accessed through various informative tables and maps. Additional analyses are available, including item and person reliability. Data analysis using the Rasch model is divided into six parts, namely 1. Validity Instrument, 2. Estimating reliability, 3. Estimating the suitability of items and persons, 4. Estimating passing items, 5. Estimating the level of difficulty, 6. Estimating the ability level of test participants [9, 23].

4. Results

4.1. Reliability Estimation

Reliability estimation is a crucial aspect of test analysis, as it indicates the consistency and stability of an instrument in measuring the intended construct. In the context of educational assessment, reliable instruments ensure that the results obtained are dependable and can be interpreted with confidence. Within the Rasch measurement framework, reliability is examined from two perspectives, namely person reliability and item reliability, which reflect the consistency of respondents' performance and the quality of the test items, respectively. To interpret the reliability values obtained from the analysis, specific criteria are required. Therefore, this section refers to the reliability criteria proposed by Sumintono [11] as presented in Table 1, which outlines the standards for evaluating person reliability and item reliability.

Table 1.
Person reliability criteria and item reliability.

Reliability Score	Category
<0.667	Weak
0.67-0.80	Average
0.81-0.90	Good
0.91-0.94	Very good
>0.94	Excellent

Reliability analysis was conducted to measure the reliability of the research instrument used to measure the variables. The results of the reliability analysis are presented in Table 2. The item and personal reliability scores were good, with values of 0.89 and 0.87, respectively. The Cronbach's Alpha reliability coefficient showed good values, indicating that the items in this research instrument consistently measure the same construct.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 3: 346-356, 2026
DOI: 10.55214/2576-8484.v10i3.12387

Table 2.
The result of person reliability and item reliability.

Subject	Estimates	Mean	SD	$SD_{(adj)}$	Reliability
Reading comprehension	Item	0.00	0.68	0.64	0.89
	Case	1.39	1.98	1.85	0.87

4.2. Item Fit Estimation

4.2.1. Infit Mean Square Criteria

Evaluating item fit is an essential step in Rasch model analysis to determine whether each test item functions as expected in measuring the intended construct. Item fit analysis helps identify items that are consistent with the model as well as those that may distort measurement due to unexpected response patterns. One commonly used statistic for assessing item fit is the Infit Mean Square (MNSQ) value, which is sensitive to response patterns from respondents whose ability levels are close to the item difficulty. By examining the Infit MNSQ values, researchers can assess the extent to which each item conforms to the assumptions of the Rasch model. To determine the suitability of each item, the obtained Infit Mean Square values are compared with established criteria. Therefore, this study refers to the Infit Mean Square criteria proposed by Setyawarno [24] as presented in Table 3, which provides guidelines for evaluating item fit within the Rasch measurement framework.

Table 3.
Infit mean square criteria.

Infit MNSQ	Interpretation
>1.33	Misfit
0.77-1.33	Fit
<0.77	Misfit

The results of the item analysis are summarized in Table 4. Based on the analysis using the Mean Squared Infit (MNSQ) value, it is clear that 13 items (items 2-5, 10, 11, 13-15, 18, 20, 22, 28, and 29) are within the acceptable fit range of 0.77 to 1.33. There are 16 items (items 1, 2, 6-9, 12, 16, 17, 19, 21, and 23-27) that require attention. Therefore, 16 items need improvement.

Table 4.
Infit mean square estimation results.

Items	Infit MNSQ	Criterion Infit MNSQ	Items	Infit MNSQ	Criterion Infit MNSQ	Items	Infit MNSQ	Criterion Infit MNSQ
1	1.77	Misfit	11	1.02	Fit	21	1.4	Misfit
2	1.49	Misfit	12	0.63	Misfit	22	0.93	Fit
3	1.23	Fit	13	0.95	Fit	23	0.66	Misfit
4	1.07	Fit	14	0.90	Fit	24	0.65	Misfit
5	0.88	Fit	15	1.12	Fit	25	0.70	Misfit
6	0.67	Misfit	16	0.52	Misfit	26	0.62	Misfit
7	0.62	Misfit	17	0.41	Misfit	27	1.46	Misfit
8	0.40	Misfit	18	1.32	Fit	28	1.15	Fit
9	1.63	Misfit	19	0.69	Misfit	29	1.21	Fit
10	1.16	Fit	20	0.75	Fit			

This analysis can be more clearly understood through a visual representation, as presented in Figure 1, which was generated using the Quest program. The figure provides a comprehensive illustration of the distribution of item difficulty and respondent ability, allowing for a clearer interpretation of how well the test items align with the ability levels of the students. By examining this figure, readers can identify patterns related to item fit, the spread of item difficulty, and the consistency of respondents' performance across the measurement scale. Consequently, Figure 1 serves as an

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 3: 346-356, 2026
DOI: 10.55214/2576-8484.v10i3.12387

important supporting component of the Rasch analysis, as it helps to strengthen the interpretation of the statistical results obtained from the Quest software.

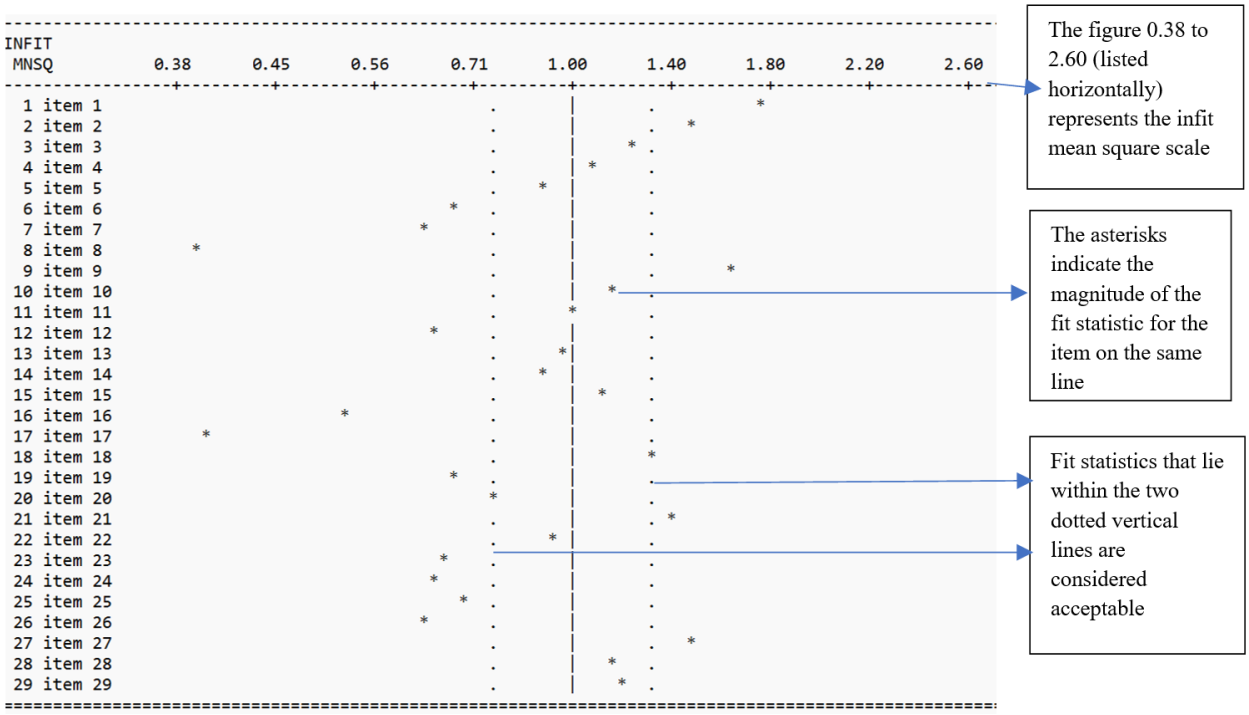

Figure 1.
Item fit map for the reading comprehension test.

As seen in Figure 1, 13 items fit the reading comprehension test because the asterisk is between two vertical dashed lines [22, 25]. This shows that 13 test items (44.83%) fit or match the Rasch Model (one-parameter logistic model) with an acceptance limit of > 0.77 to <1.30 [4].

4.2.2. Outfit T Criteria

Evaluating whether a test item passes or fails the Rasch model requirements is an important step in item analysis to ensure the accuracy of measurement. One statistical indicator commonly used to determine item acceptance is the Outfit t value, which is sensitive to unexpected response patterns, particularly from respondents whose ability levels are far from the item difficulty. By examining the Outfit t values, researchers can identify items that function appropriately as well as those that show irregular response behavior. To determine whether each item meets the model criteria, the obtained Outfit t values are compared with established standards. Therefore, this study refers to the Outfit t criteria proposed by Setyawarno [24] as presented in Table 5.

Table 5.
Outfit t criteria.

Outfit t	Interpretation
-2.00 ≤ Outfit t ≤ 2.00	Passed
Outfit t < -2.00	Failed
Outfit t > 2.00	Failed

The results of the analysis regarding the estimation of items that pass and fail the Rasch model criteria are comprehensively summarized in Table 6 below. This table presents a clear classification of

each item based on its statistical performance, allowing readers to easily identify which items meet the required standards and which items require further review or revision.

Table 6.
Passing and failing items.

Items	Outfit t	Criterion Outfit t	Items	Outfit t	Criterion Outfit t	Items	Outfit t	Criterion Outfit t
1	3.0	Failed	11	-1.0	Passed	21	3.6	Failed
2	2.1	Failed	12	1.3	Passed	22	-0.3	Passed
3	1.8	Passed	13	2.6	Failed	23	-1.2	Passed
4	0.5	Passed	14	-2.1	Passed	24	-2.2	Failed
5	-1.3	Passed	15	-0.7	Passed	25	-1.6	Passed
6	-3.1	Failed	16	-2.8	Failed	26	-1.5	Passed
7	-1.4	Passed	17	-3.3	Failed	27	3.9	Failed
8	-3.2	Failed	18	1.2	Passed	28	1.4	Passed
9	2.0	Passed	19	-1.8	Passed	29	0.7	Passed
10	0.2	Passed	20	-1.9	Passed			

Based on analysis using the Outfit t value, there are 19 items (items 3-5, 7, 9-12, 14, 15, 18-20, 22, 23, 25, 26, 28, and 29) that meet the Rasch model criteria, with Outfit t values ranging from -1.9 to 2.0. There are 10 items (items 1, 2, 6, 8, 13, 16, 17, 21, 24, and 27) that do not meet the t Outfit criteria.

4.3. Item Difficulty Level

To determine the difficulty level of each test item, the threshold value obtained from the Rasch analysis is examined and then compared with the established difficulty criteria. The threshold values for all items are matched with the classification standards proposed by Setyawarno [24] as presented in Table 7, which provides clear guidelines for categorizing items into different levels of difficulty and facilitates a more accurate interpretation of item difficulty within the measurement framework.

Table 7.
Criteria for item difficulty level.

Threshold Value	Interpretation
$b>2$	Very Difficult
$1<b\leq 2$	Difficult
$-1\leq b\leq 1$	Moderate
$-1>b\geq -2$	Easy
$b<-2$	Very Easy

Table 8.
Results of item difficulty levels.

Items	Threshold Value	Interpretation	Items	Threshold Value	Interpretation	Items	Threshold Value	Interpretation
1	-0.04	Moderate	11	0.95	Moderate	21	1.05	Difficult
2	0.18	Moderate	12	-0.28	Moderate	22	-0.28	Moderate
3	0.84	Moderate	13	-0.23	Moderate	23	-0.38	Moderate
4	-0.77	Moderate	14	0.66	Moderate	24	-0.49	Moderate
5	-0.77	Moderate	15	0.74	Moderate	25	-0.49	Moderate
6	0.70	Moderate	16	-0.66	Moderate	26	-0.66	Moderate
7	-0.81	Moderate	17	-0.71	Moderate	27	1.66	Difficult
8	-0.71	Moderate	18	0.09	Moderate	28	0.51	Moderate
9	-0.04	Moderate	19	0.23	Moderate	29	-0.83	Moderate
10	0.39	Moderate	20	0.18	Moderate			

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 3: 346-356, 2026
DOI: 10.55214/2576-8484.v10i3.12387

Analysis of the difficulty level of the questions shows that there are 27 items with a moderate level of difficulty, and 2 items are difficult.

4.4. Student ability

The analysis of ability levels is based on specific estimate value criteria, which serve as guidelines for classifying respondents' abilities. These criteria, as proposed by Setyawarno [24] and presented in Table 9, provide a clear framework for interpreting the estimated ability values and categorizing participants into different ability levels in a systematic and meaningful manner.

Table 9.
Criteria for student ability using the Rasch model.

Estimate Value	Description
>1.00	High Ability
-1.00 - +1.00	Moderate Ability
<-1.00	Low Ability

Based on the Quest results, student ability can be assessed using the estimated ability score, which is divided into three categories: high (>1.00), medium (-1.00 to 1.00), and low (<-1.00). In English, the estimated ability was high for 140 students, medium for 24 students, and low for 18 students. A summary of the abilities of the 200 students in this study is available in Table 10.

Table 10.
Summary of the respondent's abilities.

Ability Level	Total Students	%
Low	36	18
Moderate	24	12
High	140	70

5. Discussions

5.1. Reliability Estimation

The item and personal reliability scores were good, with values of 0.89 and 0.87, respectively. High item reliability indicates that the test items fit the model well, ensuring that most items are appropriate and provide the expected information [24]. Several factors can influence reliability levels, including the number of items, score distribution, item difficulty level, objectivity, the psychological and health conditions of students, teacher/grader subjectivity in scoring, and having too many items, leading to fatigue and haste in answering [18]. These factors should be considered for improving or revising the test items.

5.2. Item Fit Estimation

Based on the analysis using the Mean Squared Infit (MNSQ) value, it is clear that 13 items are within the acceptable fit range of 0.77 to 1.33. 16 items require attention because of a misfit. Based on analysis using the Outfit t value, 19 items meet the Rasch model criteria, with Outfit t values ranging from -1.9 to 2.0. 10 items do not meet the Outfit t criteria.

This is consistent with Azizah's [26] study, which analyzed the quality of Physics questions for 10th grade. In her model fit analysis results, 18 questions were found to fit the Rasch model, and two did not. The Rasch model fit criteria in Azizah's [26] study were an infit mean square value between 0.77 and 1.33 and an outfit t value ranging from -2 to 2. Additionally, Maulana et al. [27] research showed that 20 diagnostic Science items met the Rasch model criteria. Ngadi's [28] study on the analysis of questions on the subject of motorcycle electrical system maintenance showed that 30 items met the Rasch model criteria. The Rasch model criteria used by Maulana et al. [27] and Ngadi [28]

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 3: 346-356, 2026
DOI: 10.55214/2576-8484.v10i3.12387

were: 1) Outfit Mean Square (MNSQ): 0.5 < MNSQ < 1.5, 2) Outfit Z. Standard (ZSTD): -2.0 < ZSTD < +2.0, and 3) Point Measure Correlation (Pt Mean Corr): 0.4 < Pt Mean Corr < 0.85.

5.3. Item Validity

The validity of tests in the Rasch model can be seen from model fit [26]. The criteria for the Rasch model can be observed from the infit mean square and outfit t values. In this study, the infit mean square criteria use a range of 0.77 to 1.33, and the outfit t criteria use a range of -2 to 2 [24]. Items that fall within these ranges are considered valid, while items that do not fit the model are considered invalid.

5.4. Item Difficulty Level

Based on the threshold values for the reading comprehension, there are 27 items (93.10%) with a medium difficulty level, and 2 items (6.90%) with a difficult level. The study by Mutakin, et al. [29] shows the distribution of mathematics difficulty levels as follows: X > M + 2SD (very difficult), M + 2SD ≤ X < M + 1SD (difficult), M + 1SD ≤ X < 0 (medium), 0 ≤ X < M - 1SD (easy), X < M - 1SD (very easy). Out of 26 questions that fit the Rasch model, there are 2 very difficult items, 0 difficult items, 13 medium items, 10 easy items, and 1 very easy item. The studies by Rahmayani and Istiyono [30] and Li et al. [31] use the criteria b > 2 very difficult, 1 < b < 2 difficult, -1 < b < 1 medium, -1 > b > -2 easy, and b < -2 very easy. Rahmayani and Istiyono [30] found 2 difficult items, 1 very easy item, and 7 medium items. Meanwhile, Li et al. [31] found 4 difficult items (33.3%), 4 medium items (33.33%), 3 easy items (25%), and 1 very easy item (8.3%). The difficulty levels in the Li et al. [31] study show an ideal composition.

5.5. Student Ability

In the reading comprehension, 140 students (70%) are in the high ability category, 24 students (12%) are in the moderate ability category, and 36 students (18%) are in the low ability category. This aligns with Maulana et al. [27] study, which showed that in the diagnostic test for science, there were 17 participants with high ability, 14 participants with low ability, and 52 participants with medium ability. Ngadi [28] study shows that in the Electrical System Maintenance subject test, out of 49 students, 44.9% had a high ability with logits from 2.85 to 4.83, 46.94% had a medium ability (-1.78 to 1.33 logits), 4.08% had the low ability (-2.42 logits), and 4.08% had the very low ability (-3.65 logits).

6. Conclusion

The results of the Rasch analysis indicate that the instrument demonstrates good measurement quality, as reflected by high item reliability (0.89) and person reliability (0.87). Analysis of item fit shows that 13 items meet the Rasch model criteria based on MNSQ values, while 16 items exhibit misfit and require revision. Similarly, the Outfit t analysis reveals that 19 items fit the model, whereas 10 items do not meet the required criteria. In terms of item difficulty, threshold analysis indicates that the majority of items are of medium difficulty, with 27 items (93.10%), and only 2 items (6.90%) classified as difficult. Furthermore, analysis of students' reading comprehension ability shows that most students are in the high ability category (140 students; 70%), followed by the low ability category (36 students; 18%) and the moderate ability category (24 students; 12%). Overall, while the instrument shows satisfactory reliability and an appropriate distribution of item difficulty, several items require refinement to enhance the overall validity and measurement precision of the instrument.

7. Recommendations

Based on the research conclusions, several recommendations can be proposed. First, for researchers and instrument developers, items that show misfit in the Rasch analysis should be revised or improved by re-examining the alignment of indicators, the clarity of item construction, and the level of item difficulty to enhance the instrument's validity and measurement precision. Second, for teachers and educational practitioners, the instrument, which demonstrates good reliability and is dominated by

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 3: 346-356, 2026
DOI: 10.55214/2576-8484.v10i3.12387

items of moderate difficulty, can be used as an assessment tool to evaluate students' reading comprehension and to identify students across high, moderate, and low ability levels. Third, for future researchers, it is recommended to test the instrument on larger and more diverse samples and to develop a wider range of item difficulties to improve item discrimination and the generalizability of the findings.

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Copyright:

References

[1] D. Rosana, *Science learning evaluation*. Yogyakarta: UNY Press, 2025.
[2] Suparwoto, *Learning outcome measurement and assessment techniques*. Yogyakarta: Pustaka Pelajar, 2007.
[3] J. Lababa, "Item analysis with classical test theory: An introduction," *Jurnal Ilmiah Iqra'*, vol. 2, no. 2, pp. 29–37, 2018. http://dx.doi.org/10.30984/jii.v2i2.538
[4] B. Subali and P. Suyata, *A guide to analyzing educational measurement data to obtain empirical evidence of validity using the Quest program*. Yogyakarta, Indonesia: Lembaga Penelitian dan Pengabdian pada Masyarakat, Universitas Negeri Yogyakarta, 2011.
[5] E. Istiyono, D. Mardapi, and S. Suparno, "Development of a high-level thinking ability test in physics (pysthots) for high school students," *Jurnal Penelitian dan Evaluasi Pendidikan*, vol. 18, no. 1, pp. 1-12, 2014.
[6] D. F. Wardhani and A. P. Putra, "Development of standard cognitive test instruments for science subjects for grade 7 junior high school students in Banjar regency," *Indonesian Journal of Physics Education*, vol. 13, no. 1, pp. 75–82, 2016.
[7] R. K. Lord and R. K. Hambleton, *Estimation of ability*. Dordrecht: Martinus Nijhoff, 1985.
[8] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of item response theory*. Newbury Park, CA, USA: Sage, 1991.
[9] W. F. Hanna and H. Retnawati, "Analysis of the quality of mathematics test items using the Rasch model with the help of Quest software.," *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, vol. 11, no. 4, pp. 3695-3704, 2022.
[10] R. K. Hambleton and R. W. Jones, "Comparison of classical test theory and item response theory and their applications to test development," *Educational Measurement: Issues and Practice*, vol. 12, no. 3, pp. 38-47, 1993.
[11] B. Sumintono, "The Rasch model for quantitative social research," 2014. http://deceng3.wordpress.com
[12] D. Ardiyanti, "Application of the Rasch model to the development of a self-efficacy scale in students' career decision making," *Jurnal Psikologi UGM*, vol. 43, no. 3, p. 131733, 2016. https://doi.org/10.22146/jpsi.17801
[13] W. Grabe, *Reading in a second language: Moving from theory to practice*. Cambridge, UK: Cambridge University Press, 2009.
[14] I. S. P. Nation, *Learning vocabulary in another language*. Cambridge: Cambridge University Press, 2001.
[15] W. Kintsch, "Book review," 2000. www.elsevier.nldocate/pragma
[16] B. Laufer and G. C. Ravenhorst-Kalovski, "Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension," *Reading in a Foreign Language*, vol. 22, no. 1, pp. 15–30, 2010.
[17] K. Koda, "Reading and language learning: Crosslinguistic constraints on second language reading development," *Language Learning*, vol. 68, no. 1, pp. 1–35, 2007.
[18] E. Avinç and F. Doğan, "Digital literacy scale: Validity and reliability study with the rasch model," *Education and information Technologies*, vol. 29, no. 17, pp. 22895-22941, 2024. https://doi.org/10.1007/s10639-024-12662-7
[19] S. E. Embretson, *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.
[20] T. Bond and W. W. Sumintono, "Model rasch untuk penelitian sosial kuantitatif. Surabaya," 2014. http://deceng3.wordpress.com
[21] P. Susongko, C. Yuenyong, and A. Zainudin, "Buddhist critical thinking assessment using Rasch model," *Kasetsart Journal of Social Sciences*, vol. 43, no. 2, pp. 285-292, 2022.
[22] R. J. Adam and S.-T. Khoo, "Acer Quest: The interactive test analysis system," 1996.
[23] H. H. Dewi, S. M. Damio, and S. Sukarno, "Item analysis of reading comprehension questions for English proficiency test using Rasch model," *Research and Evaluation in Education*, vol. 9, no. 1, pp. 24-36, 2023. https://doi.org/10.21831/reid.v9i1.53514

[24] D. Setyawarno, *Efforts to improve the quality of test items by analyzing the Quest application*. Yogyakarta: FPMIPA UNY, 2017.
[25] S. Suwarto, S. Suyahman, S. Meidawati, Z. Zakiyah, and H. Arini, "The COVID-19 pandemic and the characteristic comparison of English achievement tests," *Перспективы науки и образования*, vol. 2, no. 62, pp. 307-329, 2023. https://doi.org/10.32744/pse.2023.2.18
[26] I. Azizah, "Analysis of the quality of the daily joint assessment questions for physics for class X of SMA Negeri 1 Patikraja," *Jurnal Pendidikan Fisika*, vol. 10, no. 2, pp. 90–104, 2023.
[27] S. Maulana, A. Rusilowati, S. E. Nugroho, and E. Susilaningsih, "Implementation of the Rasch model in the development of diagnostic test instruments," presented at the Prosiding Seminar Nasional Pascasarjana Universitas Negeri Semarang (pp. 748–756). Semarang, Indonesia: Universitas Negeri Semarang, 2023.
[28] N. Ngadi, "Rasch model analysis to measure the knowledge competency of students of SMKN 1 Kalianget in the subject of motorcycle electrical system maintenance," *Jurnal Pendidikan Vokasi Otomotif*, vol. 6, no. 1, pp. 1-20, 2023.
[29] T. Z. Mutakin, B. Tola, and B. Hayat, "Rasch model to analyze item quality and ability of fourth elementary school students," *International Journal of Innovation, Creativity and Change*, vol. 16, no. 2, pp. 181–193, 2022.
[30] F. Rahmayani and E. Istiyono, "Affective assesment instrument to assess student attitudes towards science, technology, engineering and mathematics," *Journal of Education Research and Evaluation*, vol. 6, no. 4, pp. 637-644, 2022. https://doi.org/10.23887/jere.v6i4.47681
[31] S. Li, W. Hanafiah, A. Rezai, and T. Kumar, "Interplay between brain dominance, reading, and speaking skills in English classrooms," *Frontiers in Psychology*, vol. 13, p. 798900, 2022. https://doi.org/10.3389/fpsyg.2022.798900

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 3: 346-356, 2026
DOI: 10.55214/2576-8484.v10i3.12387