English Language Teaching Educational Journal
Vol. 8, No. 1, April 2025, pp. 25-36

ISSN 2621-6485

Measuring up: Rasch analysis of English reading comprehension test for informal education learners

Arandha May Rachmawati[a,1,*], Agus Widyantoro[b,2]

[a,b] English Language Education Department, Yogyakarta State University, Jl. Colombo No. 1, Karangmalang, Caturtunggal, Depok, Sleman, Yogyakarta, 55281 Indonesia
[1]arandhamay.2023@student.uny.ac.id *; [2]agus_widyantoro@uny.ac.id

*corresponding author

ARTICLE INFO

Article history
Received 9 February 2025
Revised 27 March 2025
Accepted 08 April 2025

Keywords
Informal Education
Learners
Rasch Analysis
Reading Comprehension

ABSTRACT

This study aims to evaluate the quality of English reading comprehension test instruments used in informal learning, especially as English literacy tests. With a quantitative approach, the analysis was carried out using the Rasch model through the Quest program on 30 multiple-choice questions given to 30 grade IX students from informal educational institutions in Bantul. The results of the analysis showed that although all questions were included in the fit category for the Rasch model, the level of reliability was relatively low, that was 0.52 for items and 0.39 for participants. In addition, 13.3% of the questions showed inconsistent results (misfit), this means that there is inconsistency in the results and quality of the questions that need to be improved. The analysis of the level of difficulty also showed that there were questions that were too easy or too difficult. These findings highlight the importance of revising the test items and the need to increase the number of participants and items to obtain more accurate measurement results. This study also provides practical implications regarding the need for continuous and planned instrument development in the context of informal education, to provide valid and reliable evaluation tools to measure students' literacy skills.

How to Cite: Rachmawati, A.M. & Widyantoro, A. (2025). Measuring up: Rasch analysis of English reading comprehension test for informal education learners. *English Language Teaching Educational Journal, 8*(1), 25-36. https://doi.org/10.12928/eltej.v8i1.12747

1. Introduction

According to Butterfuss et al. (2020), reading comprehension involves three main elements: the reader, the text, and the reading activity, all of which are situated in a broader social and cultural context. To effectively comprehend a text, readers must have various abilities, such as working memory, inference, attention, motivation, and linguistic and conceptual knowledge. In the context of foreign language learning, Srisanga and Everatt (2021) added that reading comprehension is influenced by low-level skills (such as vocabulary and grammar mastery) and high-level skills (such as making inferences and monitoring comprehension). Both support each other in forming a complete understanding of the contents of the text. In second language (L2) learning, reading comprehension becomes more complicated due to significant differences in experience, institutions, and culture between the first and second languages (Rafatbakhsh & Ahmadi, 2023). Furthermore, test evaluation can also support the development of data-based learning systems that enable personalization of learning according to the reader's ability level (Stenner, 2023).

The limited time for learning in formal schools causes the need for additional learning support outside of school. Informal learning makes a significant contribution in the context of language learning because it allows learners to experience authentic, contextual, and continuous learning processes outside the formal classroom. In informal environments, such as everyday social interactions, media consumption, and community involvement, language learners can develop communicative competence naturally and oriented towards meaning. This approach allows them to use the target language in real situations, thereby improving fluency, vocabulary, and cultural understanding (Johnson & Majewska, 2022). In addition, informal learning fosters intrinsic motivation because learners are actively and personally involved in learning experiences that suit their interests and needs (Smith & Seal, 2021). From a critical education perspective, informal learning encourages linguistic empowerment, because learners not only learn language structures but also use them as tools to negotiate meaning and strengthen identities in certain social contexts (Jones & Brady, 2022). Along with the development of technology and digital media, informal spaces such as social media and online platforms also expand opportunities for flexible language learning that is integrated with everyday life (Smith, 2021). Thus, informal learning in language learning offers a holistic, relevant, and experience-based approach, which enriches learners' linguistic and socio-cultural competencies. The concept of language learning outside the classroom or informal context has attracted great interest among educators and researchers (Chong & Reinders, 2022).

There are still limited studies evaluating the quality of test instruments in the context of informal education, especially in the Bantul area. Several previous studies have shown the effectiveness of the Rasch model in evaluating and validating various types of language tests, including reading comprehension tests. Aryadoust et al. (2021) conducted a comprehensive review of the application of the Rasch model in language assessment and emphasized the importance of reporting unidimensionality, local independence, and item and person reliability. Dunn (2024) developed the Rasch model with a Generalized Linear Mixed Model approach to analyze lexical factors that influence the level of item difficulty in vocabulary tests. Meanwhile, Morea et al. (2024) used Rasch analysis to validate a receptive vocabulary test in young language learners, showing the importance of the match between item difficulty and test taker ability. Nguyen (2022) combined Rasch analysis and CFA to validate the construct structure of a college student reading achievement test, and found that a one-factor model was the most appropriate to represent reading ability. Anggia and Habók (2023) highlighted the importance of adjusting text complexity and task difficulty in developing a Rasch-based reading test for EFL students, which can improve the fit between reader ability and text characteristics. A study by Noroozi and Karami (2024) also confirmed the relevance of the Rasch model in evaluating various aspects of validity based on the Messick framework in a university-level English proficiency test. Meanwhile, Polat (2022) and Toker and Seidel (2023) applied variants of the Rasch model such as the Many-Facet Rasch Model and the Mixture Rasch Model to evaluate rater behavior and respondent heterogeneity, highlighting the flexibility of this model in the context of complex language assessment.

Educational strategies are needed to develop students' abilities in informative learning by organizing methods and reflecting on the environment as a media space for transferring information (Tkacová et al., 2022). The concept of Rasch Analysis refers to an approach in Item Response Theory (IRT) that is used to measure individual abilities and the quality of test items objectively and accurately. This model was developed by Georg Rasch and is based on the probabilistic principle, where the probability of a person answering an item correctly is predicted by the difference between the person's ability and the item's difficulty level. Rasch analysis assumes that good data is data that fits the model, not a model that is adjusted to the data (Subagja et al., 2023; Winarti & Mubarak, 2020). Technically, Rasch analysis allows the transformation of ordinal scores into interval scales (logit), allowing direct comparison between respondent ability and item difficulty on a single measuring line. Thus, Rasch provides a reliable and fair way to evaluate individual ability and item quality without relying on sample characteristics or score distributions (Rizbudiani et al., 2021; Priyani & Sugiharto, 2024). In addition, Rasch analysis allows for examination of item characteristics such as statistical fit (infit and outfit MNSQ), unidimensionality, local independence, and Differential Item Functioning (DIF), which are useful for detecting bias between respondent groups (Christensen & Ammentorp, 2024; Prabowo & Rahmadian, 2023). According to Luber et al., (2020), although in the RASCH model, item selection for developing tests is a problem, this is an effort to plan so that the quality of the test meets needs and objective.

This study aims to evaluate the quality of English reading comprehension test in the context of informal institutions through the Rasch model analysis. This study also provides the empirical insights into reliability, the suitability of test items and participants, and the level of difficulty of the questions used. Furthermore, it enables to encourage the test instrument to be better for the informal learning context in the future.

2. Method

3.1. Research Design

This research is quantitative research that uses an approach to evaluate objective theory by testing the relationship between variables that can be measured using research instruments. The data obtained can be analyzed using statistical procedures. This research was designed as survey research, which focuses on collecting information and research samples through questionnaires as the main tool for collecting data and aims to obtain an overview of various aspects of the population group. The modeling used in this research is *item response theory* (IRT). This is a probabilistic model that attempts to explain the relationship between each test taker's response to the test items through the ability variable, which is measured using an instrument.

2.2 Participants

The participants for this research classified in Table 1. were 30 students of class IX of Junior High School from one of the informal education institutions in Bantul. The sampling technique uses *convenience sampling*, namely selecting participants who are willing to be samples in the research (Creswell, 2014). In an effort to minimize bias and errors in responses, there is a briefing before the test was carried out.

Table 1. Data of Participants

Students' Gender	Grade	Total
Female	IX	20
Male	IX	10

The data collection process was carried out using *Google Forms* from 15 to 22 December 2024. Before asking test participants to fill in answers on the test instrument, test participants are first given an understanding of the procedures and confidentiality of the data provided. Furthermore, test participants are also given an explanation regarding the material being tested so that test participants can more easily fill in the answers on the instrument.

2.3 Instrument

This research focuses on developing questions related to reading comprehension as English literacy test. The reading used in this test instrument includes articles, reports, textbooks, etc. (Brown & Abeywickrama, 2018). Based on the objectives of the test instrument, the researchers used descriptive, narrative, greeting, advertisement, recount, and report texts. These types of text were chosen because they fit the context of the English materials that fit. on Merdeka Curriculum standards. The most of the questions used in this instrument were adapted from previous tests to balance the level of differences based on the Bloom Taxonomy. This foundation is used to guide variations at the cognitive level, remembering, understanding, and analyzing text. Details and specifications of questions on the test instrument with a total of 30 multiple choice questions have been presented and sorted by type in Table2.

Table 2. Test Item Specification

Questions Type	Number of Questions	Item Numbers
Identifying information from a text	2	4,25
Understanding information in a text	2	1,7
Understanding the meaning of words or phrases in context	4	2,5,8,15
Understanding the main idea of a text	1	6,21

Questions Type	Number of Questions	Item Numbers
Interpreting information from infographic data	2	26,9
Concluding which readers benefit from the text	3	11, 12, 13
Applying the correct conjunction to complete a gap sentence	1	20
Analyzing the purpose of a text	4	3,22,27,29
Analyzing correct information based on the text	3	10, 19,28
Analyzing the main information from a text	2	16,24
Analyzing the possible outcome of a specific action based on the text	2	17,23
Identifying relationships between ideas and predicting the main idea of a missing paragraph	2	18,30
Predicting actions or responses of the reader	1	14
Total Questions	**30**	

2.4 Data Analysis

Data analysis in this study used the Rasch model approach from the Item Response Theory with the help of the Quest program. This model was chosen because it is able to provide detailed information regarding the quality of the test items and the abilities of the test participants, as well as separating instrument parameters from participant characteristics. There are four main components in this analysis, namely reliability, item and participant suitability, question difficulty level, and identification of unclear items.

First, reliability analysis was conducted to evaluate the consistency of the instrument (item reliability) and the consistency of participants' answers (case reliability). Reliability values were analyzed based on Quest output and interpreted using categories commonly used in Rasch analysis, such as high, medium, and low. Second, the fit of the items and participants to the Rasch model was analyzed using two parameters, namely Infit Mean Square (MNSQ) and Infit t. The range of Infit MNSQ of 0.77 to 1.33 was used as the fit limit, according to the standard of Wright and Mok (2004), which indicates that the participant's response is still within the acceptable variation limits. In addition, the Infit t value with a range of -2 to 2 is used to identify participant responses that deviate statistically from model expectations.

Third, the level of difficulty of each question item is analyzed based on the threshold value in the logit generated from the Quest program. The logit value is then classified into five categories, namely very easy ($b < -2$), easy ($-2 \leq b < -1$), moderate ($-1 \leq b \leq 1$), difficult ($1 < b \leq 2$), and very difficult ($b > 2$). This grouping helps in assessing the balance of the distribution of the level of difficulty of the questions in the instrument.

Finally, identification of problematic questions is done by marking questions that have a perfect score value, namely questions that are answered completely correctly or completely incorrectly by the participants. Questions like this are considered not to provide meaningful information and are categorized as misfits in the Rasch model, so they need to be revised or eliminated. The results of this entire analysis process are then interpreted by referring to related literature to determine the overall quality of the instrument, as well as to provide recommendations for the use, revision, or elimination of certain items in the context of measuring English literacy in informal education.

3.5. Validity and Reliability

The validation process in research is a process that refers to the accuracy or truth of research tools and interpretations taken through research findings (William, 2024). The validation of the instrument was conducted by an English lecturer and a teacher as the expert judgments who inditifiy the content of items in the instrument. Furthermore, the validity and reliability process will be presented based on the result of classical test theory (CTT) and item response theory (IRT). This tool also has important features as a measure of item and person reliability, item suitability to the model, and the level of difficulty of each item (Faradillah & Febriani, 2021). Thus, the validity and reliability of this research instrument can be seen in the results to ensure the quality of the scores on the tests being tested.

3. Findings and Discussion

3.1. Test Reliability

The reliability analysis of the instrument was conducted using the Rasch model with the help of the Quest program. The results of the test reliability analysis on instrument items based on the RASCH model in this study are presented in Table 3.

Table 3. Statistical Summary of Item and Person Estimates

Estimates	Mean	SD	SD (adj)	Reliability	Infit Mean Square		Outfit Mean Square		Infit *t*		Outfit *t*	
					Mean	SD	Mean	SD	Mean	SD	Mean	SD
Item	.00	.93	.67	.52	1.00	.15	1.19	.74	-.01	.65	.20	.87
Case	1.94	.78	.49	.39	1.00	.25	1.19	.77	-.01	.85	.18	.90

Table 4. Item and Person Reliability in Rasch Model

Fit Indicates	Interpretation	Infit MNSQ	Interpretation	Outfit *t*	Interpretation
<0.67	Low	>1.33	Misfit	≤ 2.00	Fit
0.67-0.80	Sufficient	0.77-1.33	Fit	≥ 2.00	Misfit
0.81-0.90	Good	<0.77	Misfit		
0.91-0.94	Very Good				
>0.94	Excellent				

Table 3 presents a summary of the statistics of the item and participant estimates in the Rasch model, focusing on the aspects of reliability and data fit. From the table, it is known that the item reliability value is 0.52, while the participant reliability is 0.39. According to the standards commonly used in measurement theory, this reliability value is relatively low, both for the instrument and the respondents. Low reliability indicates that the instrument is not consistent enough in measuring the participants' abilities, and there is a possibility of large score fluctuations if the test is re-administered under similar conditions. One of the main causes of this low reliability is the limited number of participants and items, namely only 30 students and 30 questions.

However, when reviewed from the Infit Mean Square (MNSQ) and Outfit t values, it was found that all items were within the acceptable range of values. The average Infit MNSQ value for items was 1.00 with a standard deviation of 0.15, and for participants it was also 1.00 with a standard deviation of 0.25. This value is in accordance with the criteria in Table 4, which states that an item is categorized as fit if it has an Infit MNSQ value between 0.77 and 1.33. The Outfit t value on the item also showed supportive results, with an average of 0.20, within the fit range ≤ 2.00. The Infit t for the item was -0.01, this means that the participants were also classified as very close to ideal score.

The suitability of the test items to the Rasch model as seen from the Infit and Outfit values shows that even though the reliability value is low, the instrument is still statistically appropriate to the assumptions of the Rasch model. This reflects the nature of the Rasch model which places more emphasis on the suitability of the response pattern to the mathematical model, rather than relying solely on the consistency of the results (reliability). In other words, the Rasch model can provide useful information about the performance of items and participants even when reliability is low.

However, reliability remains an important aspect in instrument evaluation. Low reliability values may indicate that the instrument has not been able to distinguish participants' abilities sharply. Therefore, increasing the number of questions and participants is highly recommended so that the instrument produces more stable and reliable measurements. That way, both aspects of suitability and consistency can be achieved simultaneously to improve the quality of the test instrument.

3.2. Estimation of Person Fit

After discussing the results of the reliability tests on all items and cases, the next stage is related to the person fitting into the RASCH model. The purpose of these estimates is to provide supporting or contradictory findings.

Table 5. Person Fitting the Rasch Model

Items	Infit MNSQ	Infit *t*	Criterion	Items	Infit MNSQ	Infit *t*	Criterion
1	1.06	.41	Fit	16	.93	-.19	Fit
2	Perfect score		Misfit	17	1.12	.39	Fit
3	1.38	1.06	Fit	18	.97	.14	Fit
4	.90	-.16	Fit	19	Perfect score		Misfit
5	1.04	.22	Fit	20	.75	-1.10	Fit
6	.99	.13	Fit	21	1.43	1.54	Fit
7	.65	-1.76	Fit	22	.72	-.72	Fit
8	.81	-.31	Fit	23	.83	-.55	Fit
9	.71	-.91	Fit	24	.65	-1.32	Fit
10	.77	-.88	Fit	25	1.13	.43	Fit
11	1.32	1.03	Fit	26	1.26	.70	Fit
12	1.18	.54	Fit	27	1.25	.60	Fit
13	.65	-1.32	Fit	28	Perfect score		Misfit
14	1.31	.68	Fit	29	1.29	.96	Fit
15	.97	.05	Fit	30	Perfect score		Misfit

Criterion for fit person: 0. 77 ≤ Infit MNSQ ≤ 1, 33 OR -2 ≤ Infit *t* ≤ 2

Table 5 presents the results of the person fit estimation analysis, which aims to determine whether the answer patterns of test participants are in accordance with the expectations of the Rasch model. In this model, a person is categorized as fit if the Infit Mean Square (MNSQ) value is in the range of 0.77 to 1.33 or the Infit t value is in the range of -2 to 2. The results of the analysis show that out of 30 test participants, 26 participants (86.7%) were in accordance with the fit criteria, while 4 participants (13.3%) were categorized as misfit.

Fit participants showed that their responses to test items were consistent with the predictions of the Rasch model, meaning that they answered easy items correctly and were more likely to answer difficult items incorrectly, as expected. This reflects that most participants followed a rational response pattern appropriate to their ability level.

In contrast, four participants who were classified as misfit had answer patterns that deviated from the model predictions. Three of them obtained a perfect score, that is, answering all questions correctly. Although at first glance it looks positive, in the context of Rasch, a perfect score actually complicates the analysis because it cannot show variations in ability to answer questions with different levels of difficulty. One other participant showed an Infit MNSQ or Infit t value outside the criterion limit, which was likely caused by random answers, unintentional, or lack of concentration during the test.

This issue is essential to observe because although in general the majority of participants are fit; the presence of misfit participants can affect the accuracy of the ability parameter estimation. However, because the number is only 13.3%, its influence on the overall analysis results is relatively small. In the context of instrument development, these results indicate that in general participants can understand and respond to questions consistently, and the instrument is quite good at measuring participants' abilities according to the assumptions of the Rasch model. Overall, this person fit analysis strengthens the validity of the instrument, because the most participants show consistency in their answer forms.

3.3. Estimation of Item Fit

The other part is the estimation of item fit. This aims to show whether the test items on the instrument have functioned well or not based on the person-fit criteria in the previous section. Items fitting are also estimated individually using the results of the QUEST program output. The results of the output show a collection of information related to the suitability and feasibility of each test item for the RASCH model. The results of the items fitting analysis in this research are illustrated in Table 6.

Table 6. Items Fitting the Rach Model

Items	Infit MNSQ	Infit *t*	Criterion	Items	Infit MNSQ	Infit *t*	Criterion
1	.98	.00	Fit	16	.85	-1.00	Fit
2	.87	-.20	Fit	17	.88	.00	Fit
3	1.03	.20	Fit	18	1.11	.40	Fit
4	.85	-.10	Fit	19	1.18	.50	Fit
5	.75	-1.50	Fit	20	.99	.10	Fit
6	.75	-1.50	Fit	21	.82	-1.50	Fit
7	1.29	.90	Fit	22	1.18	.50	Fit
8	1.00	.30	Fit	23	Perfect score		Misfit
9	Perfect score		Misfit	24	1.08	.40	Fit
10	1.18	1.10	Fit	25	Perfect score		Misfit
11	.95	-.30	Fit	26	.83	-1.30	Fit
12	.79	-.40	Fit	27	.82	-.30	Fit
13	.97	.10	Fit	28	1.06	.30	Fit
14	.91	-.10	Fit	29	1.17	.50	Fit
15	1.07	.30	Fit	30	1.10	.40	Fit

Criterion for fit items: 0,77 ≤ Infit MNSQ ≤ 1,33 OR -2 ≤ Infit *t* ≤ 2

Table 6 presents the results of the item fit analysis based on two main indicators in the Rasch model, namely Infit Mean Square (MNSQ) and Infit t. These two indicators are used to assess the extent to which each item functions as it should in measuring the participants' abilities. In the context of the Rasch model, an item is considered fit if the Infit MNSQ value is between 0.77 and 1.33 or the Infit t value is in the range of -2 to 2.

The results of the analysis showed that out of 30 questions, 27 (90%) were included in the fit category, while 3 (10%) were categorized as misfit. The three misfit items were questions number 9, 23, and 25, all of which received perfect scores from the participants. This means that all participants answered the questions correctly, so that there was no variation in responses that could be analyzed. In the Rasch model, this issue is considered uninformative because the questions are unable to differentiate abilities between participants. These types of questions are considered too easy (or in some cases too difficult), so they do not contribute to effective measurement.

Meanwhile, the other 27 items encountered the suitability criteria. The Infit MNSQ values on these items were within the specified tolerance limits, indicating that most of the questions had functioned optimally in measuring participants' reading ability. This indicates that the questions were able to provide valid and representative information regarding the variation in participants' abilities based on the level of difficulty of each item.

These findings generally indicate that the quality of the items in this instrument is quite good, because most of them meet the statistical requirements in the Rasch model. However, the existence of misfit items should be a concern, because it can reduce the effectiveness of the instrument as a whole. Therefore, items that show perfect scores should be revised or replaced with items that have a moderate level of difficulty in order to provide more informative response variations and improve the quality of measurement.

3.4. Item Difficulty Level

To find out the level of difficulty of an item through the QUEST program, you can find out by looking at the data from the item estimate (Threshold) analysis. In Table 7, the data resulting from the analysis is described with the criterion for each item.

Table 7. Items' Difficulty Level

Items	Threshold	Interpretation	Items	Threshold	Interpretation
1	.58	Moderate	16	1.17	Difficult
2	-.28	Moderate	17	-.74	Moderate
3	-.28	Moderate	18	-1.47	Moderate
4	-.74	Moderate	19	-.74	Moderate
5	-.28	Moderate	20	-.24	Moderate
6	1.19	Difficult	21	1.08	Difficult
7	.06	Moderate	22	-.74	Moderate
8	.34	Moderate	23	Missing	Very Easy
9	Missing	Very Easy	24	-1.47	Easy
10	1.19	Difficult	25	Missing	Very Easy
11	1.19	Difficult	26	2.05	Very Difficult
12	-.28	Moderate	27	-.28	Moderate
13	-.28	Moderate	28	-.74	Moderate
14	.06	Moderate	29	-.28	Moderate
15	-.76	Moderate	30	-.28	Moderate
Criterion	b>2 Very difficult		-1≤b≤1 Moderate		-1≤ b ≥-2 Easy
	1<b≤2 Difficult				b<-2 Very Easy

Table 8. Percentage of Items' Difficulty Level and Quality Judgement

Criterion	Quality	Frequency	Item Numbers	Percentage
Very difficult	Very Good	1	26	3.3%
Difficult	Good	5	6,10,11,16,21	16.7%
Moderate	Good	20	1,2,3,4,5,7,8,12,13,14,15,17,18,19,20,22,27,28,29,30	67%
Easy	Not Good	1	24	3.3%
Very Easy	Not Good	3	9,23,25	10%

Based on the results of data analysis in Table 7 and Table 8 regarding the level of difficulty of the questions (item difficulty level), it can be concluded that the majority of questions in the test instrument are in the moderate difficulty level category. As many as 20 out of 30 questions or around 67% are included in the moderate category. This shows that most of the questions are at a level of difficulty that is appropriate for measuring students' general abilities. Questions with a moderate level of difficulty are generally able to provide fairly accurate information about students' abilities, because they are not too easy so that they are not challenging, and not too difficult so that they do not make it excessively difficult for students.

In addition to questions with a moderate level of difficulty, there are 5 questions (16.7%) that are classified as difficult and 1 question (3.3%) that is very difficult. Questions in this category are needed to measure students' abilities with a higher level of understanding, such as the ability to analyze, evaluate, and draw conclusions from reading texts. However, the proportion of these difficult questions still needs to be monitored so that they do not dominate the test, so as not to burden students psychologically or cognitively. On the other hand, there is also 1 question that is classified as easy

(3.3%) and 3 other questions (10%) that are classified as very easy. The existence of these too easy questions does not provide any benefit in differentiating the level of ability between participants, because they tend to be answered correctly by almost all students.

From the analysis, it can also be seen that the distribution of questions in terms of difficulty level is quite ideal because it is dominated by questions with a moderate level of difficulty, while questions that are classified as very difficult and very easy are only a few in number. This is important to ensure that the test instrument can be used effectively to measure students' abilities at various levels of ability. However, the existence of questions that are very easy or very difficult should be reviewed, because these types of questions are likely not to provide sufficient information in the evaluation process.

In addition, based on the overall quality analysis of the questions, 86.7% of the questions were classified as good and only 13.3% were poor. Questions that were included in the poor category generally came from groups of questions that were too easy or too difficult. Therefore, to improve the quality of the test instrument, these questions need to be improved or replaced so that the overall test has better discriminatory ability. The arrangement of the question sequence is also an important concern. It is better to arrange the questions starting from the easiest, then to the medium level, and ending with the difficult ones. This sequence can help participants work on the questions more comfortably and increase their motivation in completing the test. In conclusion, although in general this test instrument is quite good in terms of the distribution of difficulty levels, revisions are still needed on extreme questions so that the test results are more valid and representative in measuring students' abilities as a whole.

3.5. Discussion

The results of this study indicated that the English reading comprehension test instrument used in the context of informal learning had statistical conformity to the Rasch model, although it still shows weaknesses in terms of reliability. The item reliability value of 0.52 and participant reliability of 0.39 are low, indicating that the instrument has not been able to measure participants' abilities consistently. This is in line with the findings of Winarti and Mubarak (2020), and is supported by recent research from Priyani and Sugiharto (2024) which emphasizes that low reliability can be caused by limited number of participants or items that do not match the participants' ability level.

In terms of item fit, 27 out of 30 items (90%) were in the fit category based on the MNSQ Infit and Infit t values, meaning that most of the items functioned as they should in measuring participants' abilities according to the Rasch model's expectations. However, there were three items that showed perfect score from all participants. Although this seems ideal, it is problematic in the Rasch context because it does not provide the response variation needed to differentiate participants' ability levels (Aryadoust et al., 2021). Similar findings were also expressed by Noroozi and Karami (2024) and Morea et al. (2024), who suggested that such items be revised to provide more meaningful information.

The person fit analysis also supports these findings. As many as 86.7% of participants showed a pattern of answers that matched the Rasch model, while the other 13.3% were in the misfit category, most of whom also obtained a perfect score. This is consistent with the view of Rizbudiani et al. (2021), which state that a perfect score can reduce the accuracy of participants ability estimates. Christensen and Ammentorp (2024) even emphasized that Rasch analysis is very effective in detecting anomalies like this, especially in formative and diagnostic assessments that require high precision.

Based on the difficulty level of the questions, most (67%) were in the moderate category, which means they are were in accordance with the general abilities of the participants. This is ideal for learners in informal contexts who tend to have diverse backgrounds and ability levels (Johnson & Majewska, 2022). However, there are also questions that are too easy (10%) and very difficult (3.3%). Questions with extreme levels like this need to be reviewed because they can disrupt the balance of distribution and the overall effectiveness of the instrument (Anggia & Habók, 2023; Nguyen, 2022). Questions that are too easy are unable to distinguish participants' abilities, while questions that are too difficult can cause cognitive and psychological stress (Stenner, 2023).

These findings emphasized the importance of improving the quality of the instrument through item revisions, increasing the number of participants, and compiling more even levels of question difficulty. As suggested by Subagja, Rubini, and Kurniasih (2023), increasing the number of items and respondents can improve the reliability and discrimination of the instrument. Further research can also

consider the use of more complex Rasch models, such as the Many-Facet Rasch Model (Polat, 2022; Toker & Seidel, 2023), to capture broarder variability in the dynamic context of informal learning

In addition, the Rasch model remains a powerful tool despite its low reliability. Research by Dunn (2024) and Prabowo and Rahmadian (2023) showed that Rasch analysis is able to provide an in-depth evaluation of item function, individual response patterns, and potential measurement bias. This is especially relevant in informal education that requires assessment instruments that are adaptive and responsive to the needs of diverse learners (Chong & Reinders, 2022; Tkacová et al., 2022).

Thus, although this instrument has generally shown good conformity with the Rasch model, improvements are still needed on items that are too easy or too difficult, as well as increasing reliability by increasing the number of participants and items. This is important so that the instrument can produce more accurate, fair, and representative measurements of students' literacy skills in the context of informal learning. The application of Rasch analysis in this study has been shown to provide a significant contribution to the scientific and systematic evaluation of the instrument.

4. Conclusion

The English reading comprehension test instrument for students in informal educational institutions has a fairly good level of suitability to the Rasch mathematical model, with 90% of the items included in the fit category. However, the low reliability values for both items and participants indicate that the instrument's ability to measure and differentiate students' abilities still needs to be improved. In addition, the existence of items that are too easy or too difficult, as well as participants with perfect scores, are indicators that revisions to several items need to be made in order to improve the measuring power and fairness of the instrument.

The implications of these findings point to the need for more systematic and sustainable instrument development in the context of informal education. Teachers and managers of educational institutions need to be more careful in designing questions with a balanced level of difficulty and relevant to the abilities of students. The application of the Rasch model has proven effective in diagnosing item quality in depth, so that it can be used as a tool to improve the validity and reliability of the instrument. In practice, the development of questions must be accompanied by an increase in the number of participants and a variety of reading texts that are appropriate to the context of students' lives. Thus, assessment in informal learning is not only an evaluation tool, but also a learning instrument that encourages deep understanding and the development of sustainable literacy competencies.

Acknowledgment

We would like to thank the English Language Education Department, Yogyakarta State University, for providing the facilities and academic support needed for this study. Our sincere thanks go to the English tutoring center and students who participated in this research. Their help and cooperation made this study possible.

Declarations

Author contribution	:	The authors collaboratively carried out all stages of the research process, including the formulation of the research problem, design of the study, data collection, analysis and interpretation of the data, and preparation of the manuscript.
Funding statement	:	No funding is available for this research.
Conflict of interest	:	The authors declare that there are no conflicts of interest related to the content or conduct of this study.
Ethics declaration	:	This study adhered to ethical standards for research involving human participants. Informed consent was obtained from all participants prior to their involvement in the study. Institutional approval was also secured from the relevant academic body.
Additional information	:	No additional information is applicable to this study.

REFERENCES

Anggia, H., & Habók, A. (2023). Textual complexity adjustments to the English reading comprehension test for undergraduate EFL students. *Heliyon*, *9*(1). https://doi.org/10.1016/j.heliyon.2023.e12891

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Brown, H. D., & Abeywickrama, P. (2018). *Language assessment: Principles and Classroom Practices* (3rd ed.). Pearson/Longman.

Butterfuss, R., Kim, J., & Kendeou, P. (2020). Reading Comprehension. In *Oxford Research Encyclopedia of Education*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264093.013.865

Chong, S. W., & Reinders, H. (2022). Autonomy of English language learners: A scoping review of research and practice. *Language Teaching Research*. https://doi.org/10.1177/13621688221075812

Christensen, K. S., & Ammentorp, J. (2024). Rasch analysis of the self-efficacy (SE-12) questionnaire measuring clinical communication skills. *PEC Innovation*, *4*. https://doi.org/10.1016/j.pecinn.2024.100296

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (4th ed.). Sage Publications.

Dunn, K. J. (2024). Random-item Rasch models and explanatory extensions: A worked example using L2 vocabulary test item responses. *Research Methods in Applied Linguistics*, *3*(3). https://doi.org/10.1016/j.rmal.2024.100143

Faradillah, A., & Febriani, L. (2021). Mathematical Trauma Students' Junior High School Based on Grade and Gender. *Infinity Journal*, *10*(1), 53–68. https://doi.org/10.22460/infinity.v10i1.p53-68

Johnson, M., & Majewska, D. (2022). *Formal, non-formal, and informal learning: What are they, and how can we research them?* https://www.cambridge.org/

Jones, I. D., & Brady, G. (2022). Informal Education Pedagogy Transcendence from the 'Academy' to Society in the Current and Post COVID Environment. *Education Sciences*, *12*(1). https://doi.org/10.3390/EDUCSCI12010037

Luber, L., Fögele, J., & Mehren, R. (2020). How Do Students Experience a Deprived Urban Area in Berlin? Empirical Reconstruction of Students' Orientations. *Review of International Geographical Education Online*, *10*(4), 500–532. https://doi.org/10.33403/rigeo.763170

Morea, N., Kasprowicz, R. E., Morrison, A., & Silvestri, C. (2024). Diverse population, homogenous ability: The development of a new receptive vocabulary size test for young language learners in England using Rasch analysis. *Research Methods in Applied Linguistics*, *3*(3). https://doi.org/10.1016/j.rmal.2024.100166

Nguyen, H. T. M. (2022). Reshaping Content Validation: A Case Study of a Reading Achievement Test for Third-year English Majors. *International Journal of Language Testing*, *12*(1), 26-58.

Noroozi, S., & Karami, H. (2024). A Rasch-based validation of the University of Tehran English Proficiency Test (UTEPT). *Language Testing in Asia*, *14*(1). https://doi.org/10.1186/s40468-024-00290-4

Polat, M. (2022). A Rasch Analysis of Rater Behaviour in Speaking Assessment. *International Online Journal of Education and Teaching (IOJET)*, *7*(3), 1126–1141. https://iojet.org/index.php/IOJET/article/view/902

Prabowo, M. Y., & Rahmadian, S. (2023). Equivalency Evidence of the English Competency Test Across Different Modes: A Rasch Analysis. *Teflin Journal*, *34*(2), 301–319. https://doi.org/10.15639/teflinjournal.v34i2/301-319

Priyani, T., & Sugiharto, B. (2024). Analysis of biology midterm exam items using a comparison of the classical theory test and the Rasch model. *JPBI (Jurnal Pendidikan Biologi Indonesia)*, *10*(3), 939–958. https://doi.org/10.22219/jpbi.v10i3.34345

Rafatbakhsh, E., & Ahmadi, A. (2023). Predicting the difficulty of EFL reading comprehension tests based on linguistic indices. *Asian-Pacific Journal of Second and Foreign Language Education*, *8*(1). https://doi.org/10.1186/s40862-023-00214-4

Rizbudiani, A. D., Jaedun, A., Rahim, A., & Nurrahman, A. (2021). Rasch model item response theory (IRT) to analyze the quality of mathematics final semester exam test on system of linear equations in two variables (SLETV). *Al-Jabar: Jurnal Pendidikan Matematika*, *12*(2), 399-412. https://doi.org/10.24042/ajpm.v12i2.9939

Smith, A. (2021). COVID-19 and Informal Education: Considerations for Informal Learning During the Pandemic. *International Journal of Multidisciplinary Perspectives in Higher Education*, *6*(1), 122–127.

Smith, A., & Seal, M. (2021). The contested terrain of critical pedagogy and teaching informal education in higher education. In *Education Sciences 11*(9). MDPI. https://doi.org/10.3390/educsci11090476

Srisang, P., Everatt, J., & Th, A. (2021). Lower and Higher-Level Comprehension Skills of Undergraduate EFL Learners and Their Reading Comprehension. *LEARN Journal: Language Education and Acquisition Research Network*, *14*(1), 427–454. https://so04.tci-thaijo.org/index.php/LEARN/index

Stenner, A. J. (2023). Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement. In *Explanatory Models, Unit Standards, and Personalized Learning in Educational Measurement*. Springer Nature Singapore. https://doi.org/10.1007/978-981-19-3747-7

Subagja, S., Rubini, B., & Kurniasih, S. (2023). Development and Validation of a Test of Science Process Skills for Secondary Students on Cellular Living System Organization Matter: Rasch Model Analysis. *Jurnal Penelitian Pendidikan IPA*, *9*(10), 9056–9062. https://doi.org/10.29303/jppipa.v9i10.5177

Tkacová, H., Králik, R., Tvrdoň, M., Jenisová, Z., & Martin, J. G. (2022). Credibility and Involvement of social media in Education—Recommendations for Mitigating the Negative Effects of the Pandemic among High School Students. *International Journal of Environmental Research and Public Health*, *19*(5). https://doi.org/10.3390/ijerph19052767

Toker, T., & Seıdel, K. (2023). A Mixture Rasch Model Analysis of Data from a Survey of Novice Teacher Core Competencies. *International Journal of Contemporary Educational Research*, *10*(1), 147–156. https://doi.org/10.52380/ijcer.2023.10.1.349

William, F. K. A. (2024). Mastering validity and reliability in academic research: Meaning and significance. *International Journal of Research Publications*, *144*(1), 287-292. https://doi.org/10.47119/ijrp1001441320246160

Winarti, A., & Mubarak, A. (2020). Rasch Modeling: A Multiple-Choice Chemistry Test. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, *2*(1), 1–9. https://doi.org/10.23917/ijolae.v2i1.8985

Wright, B. D., & Mok, M. M. C. (2004). An overview of the family of Rasch measurement models. In E. V. Smith Jr. & R. M. Smith (Eds.). In *Introduction to Rasch measurement theory, models and applications 1*(1), 1-24.