AL-LISAN: JURNAL BAHASA
Publisher: LPPM IAIN Sultan Amai Gorontalo
ISSN: 2442-8965 E-ISSN: 2442-8973
Volume 10, No. 2 August 2025
Journal Homepage: https://journal.iaingorontalo.ac.id/index.php/al

Validating a Diagnostic Reading Test for Junior High School EFL Learners in Indonesia's English Massive Program Using QUEST

[1]Nur Hidayati (*Corresponding Author*)
nur0039fbsb.2023@student.uny.ac.id
English Education Department, Faculty of Languages, Arts and Cultures, Universitas Negeri Yogyakarta, Indonesia

[2]Erna Andriyanti
erna.andriyanti@uny.ac.id
English Education Department, Faculty of Languages, Arts and Cultures, Universitas Negeri Yogyakarta, Indonesia

ABSTRACT

Background: Non-formal English education programs often lack rigorous assessment tools, resulting in challenges in evaluating student performance and guiding instructional improvements. One such program, the English Massive Program (EMAS) in Kediri, East Java, Indonesia, serves as a community-driven initiative focused on enhancing English proficiency. However, the quality of its assessments, especially in reading comprehension, remains a critical concern.

Aims: This study aims to investigate how item analysis using the QUEST application can enhance the quality of diagnostic assessment and instructional strategies in non-formal English education, specifically within the EMAS program. The focus is on analysing reading comprehension tests to identify weaknesses and propose improvements in test construction.

Methods: This exploratory study analysed 30 multiple-choice reading comprehension items completed by 26 junior high school students participating in the EMAS program. The QUEST application was employed to assess item difficulty, discrimination, and distractor efficiency.

Results: The results showed that while most discrimination indices were within acceptable ranges, many items, especially the distractors, were too simple and ineffective. This resulted in insufficient and unbalanced discrimination values, indicating that the test items did not optimally differentiate among varying student ability levels.

Implications: The study underscores the importance of integrating psychometric-based diagnostic tools in community education settings. It demonstrates how such analysis can empower educators with practical insights to improve test design, thereby enhancing assessment quality and pedagogy. The research calls for more advanced diagnostic assessment methods to support literacy and instructional planning in low-resource, non-formal educational environments.

Keywords: Diagnostic assessment; literacy; QUEST; reading comprehension

Article Info:
Received: 30 June 2025
Accepted: 06 August 2025
Published: 27 August 2025
How to cite:
Hidayati, N., & Andriyanti, E. (2025). Validating a diagnostic reading test for junior high school EFL learners in Indonesia's English massive program using QUEST. *Al-Lisan: Jurnal Bahasa (e-Journal),10*(2), 270-283. https://doi.org/10.30603/al.v10i2.6708

1. INTRODUCTION

A diagnostic assessment is important for demonstrating a learner's strengths and weaknesses in language education. The tool can empower educators to create appropriate instruction, improve learning results, and ensure that all learners receive adequate help (Fan et al., 2021). Community-based educational programs are offered as alternative solutions for learners who cannot access formal schooling (Harris, 2022). One example in Indonesia is the English Massive (EMAS) Program, which offers free English instruction at all levels (English Massive, n.d.). EMAS also targets junior high school learners who usually have varying degrees of skill, educational levels, and learning styles. The diversity is challenging when it comes to designing assessments, particularly for complex language skills such as reading comprehension (Afflerbach, 2025; Catts, 2022). Suskie (2018) argued that assessments can only be helpful when they truly represent students' learning realities. If valid tests are not put in place, instruction will not meet the learner's needs, and the program's effectiveness will be compromised (Callahan, 2023).

One of the most complex language skills requiring accurate assessment is reading comprehension. As with all languages, reading comprehension is particularly central to the English language as it relates to vocabulary, grammatical structures, and even more advanced cognitive processes (Choi & Zhang, 2021). Most formal educational systems utilise reading tests that have been proven to be reliable and systematically validated (Afflerbach, 2025). On the contrary, informal programs tend to rely on teacher-made tests that are not rigorously validated (Obama & Dewey, 2022). Such a practice has the potential of underdiagnosing or overdiagnosing student capabilities, forgetting important learning gaps. Locally developed tests are often used in the EMAS program. However, due to a lack of psychometric analysis, it cannot be determined if the items are valid indicators of reading proficiency. Bayley et al. (2021) claimed that assessments that are not well-structured often do not take into account different degrees of learner ability, and all become recipients of the same instruction, which does not foster development. Evaluating the accuracy of reading comprehension items as items of instruction should receive attention so that learners are not over-supported or over-challenged.

To address the challenge of validating reading comprehension assessments in non-formal contexts like EMAS, the present study utilises the QUEST program, a computer-based statistical tool developed by the Australian Council for Educational Research (ACER) (Adams & Khoo, 1996; Ikhsanudin et al., 2023). QUEST is designed to perform both Classical Test Theory (CTT) and Item Response Theory (IRT) analyses, offering researchers comprehensive psychometric data including item difficulty, item discrimination, distractor effectiveness, item fit, and reliability estimates (Adams & Khoo, 1996; Izard, 2005). Its utility lies in the program's ability to process both dichotomous and polytomous items, making it particularly suited for analysing multiple-choice questions as well as Likert-type scales.

Researchers input the answer key and students' responses through formatted control files, and QUEST subsequently generates detailed output files that provide insights into each test item's performance. The output includes critical indices: the difficulty level, the discrimination power, and the percentage distribution of responses, which allows the identification of ineffective distractors. The program also supports Rasch analysis and fit statistics that further strengthen the test's construct validity and internal consistency (Dewi et al., 2023). With these capabilities, QUEST is an effective tool for ensuring the psychometric rigour of assessments, particularly in community-based education where resources for test validation are limited (Robillard et al., 2018). Its application in this study provides an empirical basis for identifying which test items truly reflect learner proficiency in reading comprehension and which items require revision or removal.

1.1 Research Gap and Novelty

While diagnostic assessment is increasingly used in formal education, it is still not well developed in non-formal learning settings. Many community-based programs, including the English Massive (EMAS) program, lack proper diagnostic tools or sufficient training to create practical assessments (Carliner, 2023). This gap becomes more complex due to the diverse characteristics of non-formal programs, where differences in student attendance, teaching hours, and teacher qualifications make it challenging to apply standardised assessments (Almeida & Morais, 2024). Despite increasing global interest in assessment literacy, most research still focuses on formal institutions and tends to overlook local community programs that are important for basic language learning (Sukarno et al., 2024; Tsagari & Armostis, 2025). This study aims to fill the gap by carefully examining reading test items used in the EMAS program to guide practical diagnostic assessment in community-based English learning settings with limited resources.

This research provides new insights by showing that psychometric tools such as QUEST can be adjusted for use in settings with limited resources. Although QUEST is often used in formal academic environments, it is rarely applied in community-based programs (Ikhsanudin et al., 2023). By showcasing a practical case study, this work bridges the divide between psychometric theory and classroom realities in informal education. The study supports prior findings on the importance of using data to guide testing decisions, while also highlighting that assessment should be part of the teaching process, not just an activity done at the end of instruction (Murphy et al., 2023). In addition, the study highlights the need to strengthen tutors' skills, especially for those without formal teaching backgrounds, by promoting easy-to-use and creative assessment methods that encourage fair and flexible learning (Gunawardena et al., 2024; Levy-Feldman, 2025). From this perspective, EMAS can be seen as an example of how open educational programs can support long-term improvements in assessment practices within community-based settings.

1.2 Research Question

This study analyses the psychometric aspects of the 30-item reading comprehension diagnostic test within the EMAS Program. The analysis was based on the results from 26 junior high school students from different levels of proficiency and emphasised their performance on test items regarding difficulty, discrimination, and distractor effectiveness. Evaluating item quality through QUEST provided important insights into systematic evaluation that can improve these tests. The focus was to determine whether the current assessment framework helps differentiate learners' levels and facilitates informed instructional interventions. These results provide concrete recommendations to educators, program developers, and education policy decision makers interested in refining evaluation tools for non-formal contexts. The following questions guided the research:

1. How do item difficulty, item discrimination, and distractor efficiency reflect the validity and reliability of a reading comprehension test used in the English Massive (EMAS) non-formal education program?
2. How can the results from QUEST item analysis help improve test quality and support tutors in developing diagnostic assessment skills in community-based English teaching programs?

2. METHODS

2.1 Research Design

A descriptive cross-sectional approach was employed in this study to evaluate the reading comprehension diagnostic test used in the English Massive (EMAS) Program, a community-based English language learning initiative located in Kediri, East Java, Indonesia. This research design was considered suitable for examining the effectiveness of a single assessment administration across a diverse group of learners at one point in time. The study aimed to assess the psychometric properties of the test in terms of item difficulty, item discrimination, and distractor efficiency. The QUEST application was used for the analysis because it offers accessible and reliable Rasch-based item analysis, making it particularly suitable for educational researchers in low-resource or non-formal contexts.

2.2 Research Subjects

The participants consisted of 26 junior secondary school students aged between 12 to 15 years old. These students were selected through purposive sampling to represent a range of English proficiency levels from A2 to B2 based on the Common European Framework of Reference for Languages (CEFR). The CEFR levels were determined through a placement test administered by EMAS facilitators prior to the study, which included vocabulary, grammar, and reading comprehension components aligned with CEFR descriptors. All participants were active members of the EMAS Program and had previously attended regular English instruction sessions. Their varied levels of competence and experience in English made them suitable subjects for evaluating the effectiveness of a diagnostic assessment in a non-formal educational setting.

2.3 Research Procedures

To maximise accessibility and streamline the data collection process, the diagnostic assessment was administered via Google Forms. Each student was given 45 minutes to complete the 30-item test under standardised and supervised classroom conditions at an EMAS learning centre to ensure consistency and minimise external distractions. Although the test itself had not been piloted, the participants were familiar with multiple-choice reading tasks from their regular EMAS sessions. Participation in the study was entirely voluntary, and informed consent was obtained from each participant before the data collection began. The digital format also allowed for efficient data retrieval and analysis while maintaining ethical standards in research involving minors.

2.4 Research Instruments

The primary instrument in this study was a 30-item multiple-choice reading comprehension diagnostic test developed collaboratively by the researchers and local EMAS tutors. The items were developed from scratch based on CEFR descriptors and EMAS learning objectives, with attention to learners' typical language exposure and classroom practices. The test aimed to measure various sub-skills of reading comprehension, including identifying main ideas, making inferences, and interpreting information from written texts. To enhance content validity, three peer educators familiar with the EMAS curriculum reviewed the items. Their evaluation focused on content relevance, linguistic clarity, cognitive demand, and alignment with instructional goals. This expert judgment ensured that the test reflected the learners' actual instructional context and addressed their learning needs. No formal pilot test was conducted prior to the main administration, but the items were refined through expert review and informal feedback during the development phase.

2.5 Data Analysis

The data collected from students' responses were analysed using the QUEST software, which provides psychometric analysis based on Rasch modelling. The analysis focused on three main psychometric indicators:

Item Difficulty

Item difficulty was evaluated using the index proposed by Bhat and Prasad (2021), with the following interpretation:

Table 1 Item Difficulty Index

Difficulty Index	Interpretation
P < 30%	Difficult
P = 30–70%	Moderate
P > 70%	Easy

Item Discrimination

Two frameworks were used to interpret item discrimination, including Suwarto's and Ebel's classifications. These frameworks were applied to provide complementary perspectives that align with local practices in Indonesian educational research, while Ebel's framework offers more refined categorical distinctions that enrich the analysis.

Table 2 Item Discrimination Index by Suwarto (2007)

Discrimination Index	Interpretation
0.71 – 1.00	Very Good
0.41 – 0.70	Good
0.20 – 0.40	Good Enough
< 0.20	Poor
< 0.00	Very Poor

Table 3 Discrimination index by Ebel (1965)

Discrimination Index	Interpretation
≥ 0.40	Functions very well
0.30 – 0.39	Good, minor revisions
0.20 – 0.29	Questionable, needs revision
< 0.20	Discard or revise

Distractor Efficiency

According to Arikunto (2012), distractors were deemed effective if selected by at least 5% of respondents. Distractors chosen by fewer than 5% of participants were categorised as ineffective. This 5% threshold is grounded in the principle that effective distractors must be attractive enough to mislead less proficient learners, thereby enhancing the diagnostic power of the test. QUEST software provided detailed item-level diagnostics, allowing the researchers to examine the functionality of distractors and their effectiveness in distinguishing student understanding. The combination of expert validation and psychometric analysis provided a robust framework for evaluating the quality of the diagnostic instrument. This methodological approach aligns with the study's objective to enhance assessment literacy and promote evidence-based improvements in non-formal education settings such as the EMAS Program.

3. FINDINGS AND DISCUSSION

3.1 Findings

This study aims to evaluate the quality of a test instrument by analysing each item based on three key aspects: item difficulty, item discrimination, and distractor efficiency. By examining how well the test items function, the researchers seek to identify which questions are appropriate, which need revision, and which should be removed. The goal is to ensure that the test accurately measures students' understanding and provides reliable results for assessment purposes.

Analysis of Item Difficulty Levels

The researchers determined item difficulty based on students' responses. After addressing the general reliability of the items, this subsection analyses the estimation of each item's difficulty level. This estimation is crucial in assessing test items to ascertain if they are appropriately designed to measure students' abilities. The item difficulty levels are presented in Table 4 along with their coefficient percentages. The items are categorised as Easy, Moderate, and Difficult. Most of these items fall within the Easy category; having coefficients of more than 80% means that many participants found these items uncomplicated.

Table 4 Items' Difficulty Level

Items	Coefficient	Description	Items	Difficulty Index	Description
1	88.5%	Easy	16	88.5%	Easy
2	96.2%	Easy	17	65.4%	Moderate
3	96.2%	Easy	18	23.1%	Difficult
4	84.6%	Easy	19	65.4%	Moderate
5	76.9%	Easy	20	100%	Easy
6	46.2%	Moderate	21	57.7%	Moderate
7	100%	Easy	22	61.5%	Moderate
8	80.8%	Easy	23	80.8%	Easy
9	84.6%	Easy	24	80.8%	Easy
10	96.2%	Easy	25	38.5%	Moderate
11	88.5%	Easy	26	84.6%	Easy
12	57.7%	Moderate	27	57.7%	Moderate
13	88.5%	Easy	28	92.3%	Easy
14	46.2%	Moderate	29	69.2%	Moderate
15	38.5%	Moderate	30	96.2%	Easy

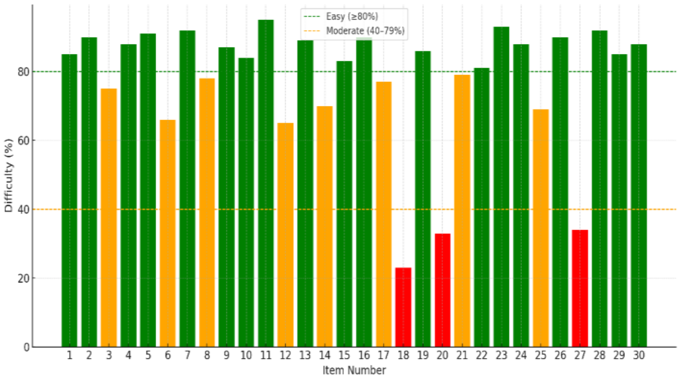

Figure 1. Item Difficulty Levels (Item 1–30)

The analysis of the test's difficulty level indicates that most of the items, such as 1, 2, 3, 7, 10, and 20, are categorised as Easy with coefficients between 80.8% and 100%. Items 6, 12, 14, 17, 19, and 25 were also categorised as Moderate, indicating a moderate level of difficulty. Remarkably, only Item 18 was classified as Difficult, with a coefficient of 23.1%, indicating it was the most difficult item for the participants.

Analysis of Item Discrimination Using Pt-Biserial

The effectiveness of each test item was assessed using Point Biserial Correlation (Pt-Biserial) to evaluate its discrimination capacity between high and low performing students. After assessing overall test reliability, this subsection shifts its emphasis to individual item analysis to determine which test items require revision, retention, or deletion. This analysis is critical in ensuring that the test accurately captures students' competencies while ensuring its authenticity. Items with high Pt-Biserial scoring are characterised by strong discrimination, and those scoring low do not add value to the assessment. The impact of individual item scoring on overall test performance is analysed using Pt-Biserial in Table 1. Items are categorised as Poor, Good Enough, Good, or Very Good to decide on whether revised, retained, or deleted is warranted. A Pt-Biserial value close to zero reflects weak discrimination. It suggests revision or removal is necessary, while higher values indicate stronger item quality and more effective differentiation between student performance outcomes.

Table 5 Item Discrimination

Items	Pt-Biserial	Description	Decision	Items	Pt-Biserial	Description	Decision
1	0.41	Good	No revision	16	0.67	Good	Minor revision
2	0.01	Poor	Major Revision	17	0.38	Good enough	Minor revision
3	0.09	Poor	Major Revision	18	0.49	Good	No revision
4	0.39	Good enough	Minor revision	19	0.52	Good	No revision
5	0.57	Good	No revision	20	0.00	Poor	Major Revision
6	0.58	Good	Minor revision	21	0.29	Good enough	Minor revision
7	0.00	Poor	Major Revision	22	0.50	Good	Minor revision
8	0.51	Good	Minor revision	23	0.78	Very Good	Minor revision
9	0.60	Good	Minor revision	24	0.63	Good	Minor revision
10	0.43	Good	Minor revision	25	0.54	Good	No revision
11	0.67	Good	Minor revision	26	0.66	Good	Minor revision
12	0.42	Good	Minor revision	27	0.13	Poor	Major Revision
13	0.51	Good	Minor revision	28	0.68	Good	Minor revision

14	0.03	Poor	Major Revision	29	0.49	Good	No revision
15	0.48	Good	No revision	30	0.51	Good	Minor revision

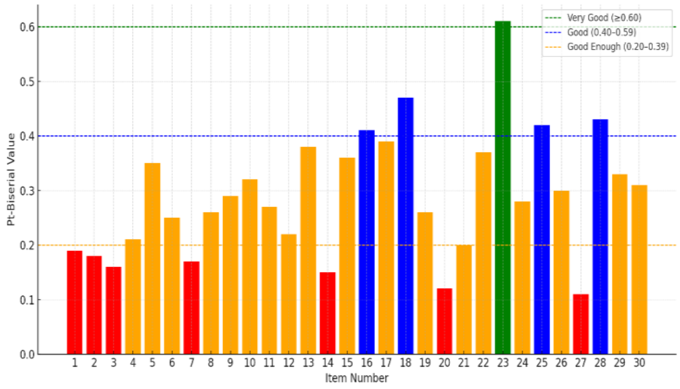

Figure 2. Item Discrimination Values (Pt-Biserial)

The findings suggest that certain items, including 2, 3, 7, 14, 20, and 27, have low discrimination and require modification or elimination. On the other hand, items 4, 17, and 21 are categorised as "Good Enough," which implies that only minor revisions are needed. The majority of items are classified as "Good," signifying acceptable performance, except for Items 6, 8, 9, and 10, which need further development. Item 23 is remarkable since it is rated as "Very Good", highlighting its robust correlation with overall test performance.

Distractor Efficiency in Multiple-Choice Test Items

To evaluate distractor efficiency, the researchers analysed how well the incorrect answer choices functioned in multiple-choice test items. This evaluation is critical in determining whether the distractors accurately distinguish between students who understand the material and those who do not. Ineffective distractors decrease the possibility of students who are less informed being able to answer the questions correctly, thereby diminishing the test's ability to measure actual understanding. Revising the distractors can enhance the functioning of the assessment, thus improving its overall quality and validity. Table 6 evaluates distractor efficiency, which measures how well incorrect answer choices function in multiple-choice test items. A distractor loses effectiveness when it fails to mislead less-informed students because the item becomes less effective at assessing accurate understanding of the content. When distractors are poorly designed, this contributes to making the test excessively easy to administer, consequently reducing the reliability of the test as a whole. Therefore, poorly crafted distractors necessitate elimination or fixing so that the assessment distinguishes students accurately based on their understanding.

Table 6 Distractor Efficiency

Items	Description	Decision	Items	Description	Decision
1	C and D are ineffective	Major Revision C and D	16	A and C are ineffective	Major Revision A and C
2	A, B, C ineffective	Discard	17	A is ineffective	Minor Revision A

3	B, C, and D are ineffective	Discard	18	A, B, and D are effective	No revision
4	A and C are ineffective	Major Revision A and C	19	B, C, and D are effective	No revision
5	A, B, and D are effective	No revision	20	A, B, C ineffective	Discard
6	B ineffective	Minor Revision B	21	A is ineffective	Minor Revision A
7	A, B, and D are ineffective	Discard	22	B is ineffective	Minor Revision B
8	A is ineffective	Minor Revision A	23	D is ineffective	Minor Revision D
9	B and C are ineffective	Major Revision B and C	24	B is ineffective	Minor Revision B
10	A, C, and D are ineffective	Discard	25	A, C, and D are effective	No revision
11	C and D are ineffective	Major Revision C and D	26	A and C are ineffective	Major Revision A and C
12	C is ineffective	Minor Revision C	27	C ineffective	Minor Revision C
13	B and C are ineffective	Major Revision B and C	28	A, B, and D are ineffective	Discard
14	A, B, and D are effective	Discard	29	A, B, C effective	No revision
15	A, B, and D are effective	No revision	30	A, C, and D are ineffective	Discard

The analysis indicates that some items, such as 2, 3, 7, and 20, have an ineffective combination of distractors, which indicates these items require complete revision or removal to adjust the level of difficulty for the test. On the other hand, items 6, 8, 12, 21, and 22 only show one ineffective distractor and need only minor changes. In parallel, the rest of the items in question, including 5, 18, 19, 25, and 29, have fully functional distractors and therefore require no adjustment. These items successfully distinguish students who understand the material from those who are struggling. In general, these findings illustrate the greater need to tailor distractors to increase test quality and validity, especially regarding the reasons behind incorrect answer choices.

3.2 Discussion

This study assessed the psychometric properties of a diagnostic reading comprehension test used in the English Massive (EMAS) Program. Using QUEST software, item-level analysis was conducted focusing on difficulty, discrimination, and distractor efficiency. These three factors are critical for determining the reliability of diagnostic tests and their usefulness within a non-formal educational context. The results enhance understanding of how targeted educational interventions can utilise tools grounded in the best available evidence to support assessment practices and improve educational outcomes.

Item Difficulty and Its Implications for Diagnostic Assessment

The analysis revealed that the test items were relatively "easy," with over 80% of students answering them correctly. Such an imbalance compromises the test's ability to discriminate among learners with different proficiency levels and limits its value for formative purposes. According to constructivist learning theory, tasks should be appropriately challenging to activate prior knowledge and scaffold new learning (Bruner, 1966; Vygotsky & Cole, 1978). When diagnostic tests consist mainly of low-difficulty items, they fail to reveal the learners' actual cognitive readiness and areas that require targeted support. In the EMAS program's context, where students come from varied educational backgrounds and often lack structured language instruction. This imbalance limits the identification of learner zones needing development. Item difficulty analysis using p-values substantiates this. For instance, Item 2 had a difficulty index of 0.95, showing little diagnostic value, while Item 30, with a difficulty of 0.22, may be too advanced, especially without alignment to prior instruction. Items falling in the moderate range ($p = 0.30–0.70$) are essential for capturing variation in student performance and informing targeted teaching interventions.

The trend toward overly easy items might stem from item writers, often untrained community tutors, intentionally avoiding complexity due to concerns over learner confidence or test anxiety (Gkintoni et al., 2025; Gore et al., 2022). While well-intentioned, this practice inadvertently undermines the principles of assessment for learning, which advocate for tasks that push learners just beyond their current ability (Black & Wiliam, 2009). Ideally, tests should adopt a balanced composition (25% easy, 50% moderate, 25% difficult) to activate multiple levels of comprehension and provide equitable opportunities for diagnostic insight. The current imbalance not only hinders accurate profiling but also poses risks of instructional complacency, as high scores may mask actual gaps in higher-order thinking.

Item Discrimination and Its Pedagogical Implications

Despite issues with item difficulty, the majority of test items displayed satisfactory discrimination, suggesting that they could differentiate students with varying reading abilities. Discrimination analysis using Point-Biserial Correlation confirmed that several items functioned effectively in distinguishing student performance, with many surpassing the 0.20 threshold and a few reaching "very good" levels. From a pedagogical perspective, this ability is vital in formative assessment because it allows teachers to determine which learners need remediation and which concepts require reteaching (Brown & Brown, 2018).

High discrimination indices, even in easier items, suggest that such items may still offer insight, particularly among lower-performing learners (Liu et al., 2023). For example, Item 23, with a high discrimination index of 0.78, proved effective in highlighting learner differences. However, items like Item 7 (discrimination index of 0.00) failed to do so, possibly due to ambiguous wording or misalignment with instructional goals.

In non-formal learning environments such as EMAS, where assessment literacy is often limited, the presence of discriminating items reflects an opportunity for instructional improvement. Teachers can use discrimination data to make evidence-based decisions, designing differentiated instruction, grouping learners by proficiency level, and monitoring progress. This aligns with the data-driven instruction model, which emphasises using assessment feedback for continuous instructional refinement (Brookhart, 2024). Equipping community-based tutors with the capacity to interpret item discrimination fosters not only better test design but also contributes to professional development and sustainable grassroots education reforms.

Distractor Efficiency and Its Role in Diagnostic Assessment

Distractor efficiency evaluates the effectiveness of incorrect answer options in distinguishing between learners who understand the material and those who do not. In this study, many distractors failed to meet the minimum 5% selection threshold, indicating inefficiency. Ineffective distractors allow students to guess the correct answer without demonstrating comprehension, thus reducing the diagnostic power of the item (Cook et al., 2023).

In a culturally and linguistically diverse setting like EMAS, distractor performance may be influenced by students' language exposure, test-taking strategies, or even socio-linguistic expectations. For example, distractors that rely on subtle grammatical errors or idiomatic usage may not be salient to learners unfamiliar with these conventions. This highlights the need for context-sensitive test design, where distractors are both linguistically plausible and pedagogically meaningful (Zhang et al., 2025). Item 2, for instance, had distractors that were too obviously incorrect, possibly due to unnatural phrasing or culturally irrelevant vocabulary, making it easier for test-takers to eliminate them without comprehension.

Moreover, efficient distractors can reveal learners' misconceptions and serve as diagnostic cues (Tukiyo et al., 2023; Firdaus et al., 2025). A distractor chosen by many learners may indicate a common misunderstanding, which can be addressed through focused instruction. For instance, if a distractor reflects overgeneralization (e.g., assuming every paragraph starts with a topic sentence), instructors can design mini-lessons targeting such assumptions. In this way, distractor analysis bridges psychometric evaluation and instructional intervention.

Instructors in the EMAS Program, many of whom design assessments with limited training, can benefit from using QUEST's automated distractor analysis tools. These tools simplify the process of identifying non-functioning distractors and refining items for re-administration. Embedding reflective practice into test revision processes can help foster a local culture of assessment-informed teaching, supporting ongoing improvements in both test design and pedagogical strategies.

4. CONCLUSIONS

This study highlights the importance of using diagnostic assessments to improve teaching in non-formal English programs. By applying QUEST-based analysis to the EMAS reading test, the study found several issues with item difficulty, discrimination, and distractor quality. While some items effectively demonstrated differences in student ability, the overall test was too easy, and many distractors were not effective. These findings show the need to improve how teachers create test items, design lessons, and assess learning in community-based programs. They also prove that even schools with limited resources can use evidence-based tools to improve learning. Helping teachers build assessment skills and use simple psychometric tools can lead to fairer and more effective teaching. However, this study has some limitations. It only looked at one test from one program and did not include student feedback. Future studies could explore other programs, collect data over a longer time, or include interviews and observations to get a deeper understanding. These steps can help improve assessments in other informal learning settings and support more inclusive, data-informed English teaching.

Acknowledgements
We extend our heartfelt appreciation to the Indonesia Endowment Fund for Education (LPDP) for their invaluable assistance in facilitating the publication of this work.

Authors' Contributions

Nur Hidayati was responsible for conceptualising the study, designing the research instruments, collecting and analysing the data, and drafting the manuscript. Erna Andriyanti supervised the research process, provided critical feedback on the methodology and interpretation of results, and contributed to the refinement and finalisation of the manuscript. Both authors have read and approved the final version of the manuscript.

REFERENCES

Adams, R. J., & Khoo, S.-T. (1996). *ACER Quest: The interactive test analysis system*. Melbourne, Australia: ACER Press. https://research.acer.edu.au/measurement/3/

Afflerbach, P. (2025). *Understanding and using reading assessment, K-12*. Guilford Publications.https://www.guilford.com/books/Understanding-and-Using-Reading-Assessment-K-12/Peter-Afflerbach/9781462556120

Almeida, F., & Morais, J. (2024). Non-formal education as a response to social problems in developing countries. *E-Learning and Digital Media*, *22*, 1–17. https://doi.org/10.1177/20427530241231843

Arikunto, S. (2018). *Dasar-dasar evaluasi pendidikan* (Edisi ke-3). Jakarta: PT Bumi Aksara. https://books.google.co.id/books/about/Dasar_Dasar_Evaluasi_Pendidikan_Edisi_3.html?id=j5EmEAAAQBAJ&redir_esc=y

Bayley, T., Wheatley, D., & Hurst, A. (2021). Assessing a novel problem-based learning approach with game elements in a business analytics course. *Decision Sciences Journal of Innovative Education*, *19*(3), 185–196. https://doi.org/10.1111/dsji.12246

Bhat, S. K., & Prasad, K. H. L. (2021). *Item analysis and optimizing multiple-choice questions for a viable question bank in ophthalmology: A cross-sectional study. Indian Journal of Ophthalmology*, *69*(2), 343–346. https://doi.org/10.4103/ijo.IJO_1610_20

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Brookhart, S. M. (2024). Educational assessment knowledge and skills for teachers revisited. *Education Sciences*, *14*(7). 1-15. https://doi.org/10.3390/educsci14070751

Brown, H. D., & Brown, H. (2018). *Language assessment: Principles and classroom practices*. Pearson Education. https://books.google.co.id/books?id=a7nqswEACAAJ

Bruner, J. (1966). *Toward a theory of instruction*. Harvard University Press. https://books.google.co.id/books?id=F_d96D9FmbUC

Callahan, C. M. (2023). Evaluation for decision-making: The Practitioner's guide to program evaluation. In *Systems and models for developing programs for the gifted and talented* (pp. 119–142). Routledge.

Carliner, S. (2023). *Informal learning basics*. ASTD Press. https://books.google.co.id/books/about/Informal_Learning_Basics.html?id=WvnJEAAAQBAJ&redir_esc=y

Catts, H. W. (2022). Rethinking how to promote reading comprehension. *American Educator*, *45*(4), 26. https://files.eric.ed.gov/fulltext/EJ1322088.pdf

Choi, Y., & Zhang, D. (2021). The relative role of vocabulary and grammatical knowledge in L2 reading comprehension: A systematic review of literature. *International Review of Applied Linguistics in Language Teaching*, *59*(1), 1–30. https://doi.org/10.1515/iral-2017-0033

Cook, K. S., Fogelberg, K., Butterbrodt, P., Jolley, K., Raghavan, M., & Smith, J. R. (2023). Assessing student learning: Exams, quizzes, and remediation. In K. Fogelberg (Ed.), *Educational Principles and Practice in Veterinary Medicine* (pp. 287–312). Wiley-Blackwell.

Dewi, H. H., Damio, S. M., & Sukarno, S. (2023). Item analysis of reading comprehension questions for English proficiency test using Rasch model. *REID (Research and Evaluation in Education)*, *9*(1), 24–36. https://doi.org/10.21831/reid.v9i1.53514

Ebel, R.L. (1965). *Measuring educational achievement.* Prentice-Hall.

Fan, T., Song, J., & Guan, Z. (2021). Integrating diagnostic assessment into curriculum: A theoretical framework and teaching practices. *Language Testing in Asia*, *11*(1), 2. https://doi.org/10.1186/s40468-020-00117-y

Firdaus, A. R., Wulandarie, E., Cantika, M., & Suryana, T. G. S. (2025). Development and validation of diagnostic instrument to identify student misconceptions in Vector material. *Journal of Innovative Physics Education Research*, *1*(1), 1–14. https://doi.org/10.61142/jiper.v1i1.195

Gkintoni, E., Antonopoulou, H., Sortwell, A., & Halkiopoulos, C. (2025). Challenging cognitive load theory: The role of educational Neuroscience and artificial intelligence in redefining learning efficacy. *Brain Sciences*, *15*(2). https://doi.org/10.3390/brainsci15020203

Gore, J., Jaremus, F., & Miller, A. (2022). Do disadvantaged schools have poorer teachers? Rethinking assumptions about the relationship between teaching quality and school-level advantage. *The Australian Educational Researcher*, *49*(4), 635–656. https://doi.org/10.1007/s13384-021-00460-w

Gunawardena, M., Bishop, P., & Aviruppola, K. (2024). Personalized learning: The simple, the complicated, the complex and the chaotic. *Teaching and Teacher Education*, *139*, 104429. https://doi.org/10.1016/j.tate.2023.104429

Harris, J. (2022). Adult English learners with limited or interrupted formal education in diverse learning settings. In L. J. Pentón Herrera (Ed.), *English and students with limited or interrupted formal education: Global perspectives on teacher preparation and classroom practices* (pp. 43–59). Springer. https://doi.org/10.1007/978-3-030-86963-2_4

Ikhsanudin, I., Novaliah, N., Hidayatullah, H., & Almizi, M. (2023). A practical using of the quest program to analyze the characteristics of the test items in educational measurement. *JISAE: Journal of Indonesian Student Assessment and Evaluation*, *9*(1), 37–43. https://doi.org/10.21009/jisae.v9i1.31163

Izard, J. (2005). *Trial testing and item analysis in test construction* (Module 7, *Quantitative research methods in educational planning*). UNESCO International Institute for Educational Planning. https://unesdoc.unesco.org/ark:/48223/pf000014260

Levy-Feldman, I. (2025). The Role of assessment in improving education and promoting educational equity. *Education Sciences*, *15*(2). https://doi.org/10.3390/educsci15020224

Liu, C.-C., Liu, S.-J., Hwang, G.-J., Tu, Y.-F., Wang, Y., & Wang, N. (2023). Engaging EFL students' critical thinking tendency and in-depth reflection in technology-based writing contexts: A peer assessment-incorporated automatic evaluation approach. *Education and Information Technologies*, *28*(10), 13027–13052. https://doi.org/10.1007/s10639-023-11697-6

Murphy, D. H., Little, J. L., & Bjork, E. L. (2023). The value of using tests in education as tools for learning—not just for assessment. *Educational Psychology Review*, *35*(3), 89. https://doi.org/10.1007/s10648-023-09808-3

Obama, P. B., & Dewey, J. (2022). Assessment: Formal and informal. *Teaching Middle Level Social Studies: A Practical Guide for 4th-8th Grade*, 141.

Robillard, J. M., Jun, J. H., Lai, J.-A., & Feng, T. L. (2018). The QUEST for quality online health information: Validation of a short quantitative tool. *BMC Medical Informatics and Decision Making*, *18*(1), 87. https://doi.org/10.1186/s12911-018-0668-9

Sukarno, S., Putro, N. H. P. S., Fitrianingsih, I., Alsamiri, Y. A., Gharamah, F. M. A., & Tatin, I. A. G. (2024). Exploring the perceptions of literacy in assessment for learning among high school English teachers. *REID (Research and Evaluation in Education)*, *10*(2), 143–154. https://doi.org/10.21831/reid.v10i2.71324

Suskie, L. (2018). *Assessing student learning: A common sense guide* (3rd ed.). Jossey-Bass.

Suwarto. (2007). *Tingkat kesukaran, daya beda, dan reliabilitas tes menurut teori tes klasik*. *Jurnal Pendidikan,* *16*(2), 166–178. http://portalgaruda.fti.unissula.ac.id/?ref=browse&mod=viewarticle&article=268287

Tsagari, D., & Armostis, S. (2025). Contextualizing language assessment literacy: A comparative study of teacher beliefs, practices, and training needs in Norway and Cyprus. *Education Sciences*, *15*(7). https://doi.org/10.3390/educsci15070927

Tukiyo, T., Efendi, M., Solissa, E. M., Yuniwati, I., & Pranajaya, S. A. (2023). The development of a two-tier diagnostic test to detect student's misconceptions in learning process. *Mudir: Jurnal Manajemen Pendidikan,* *5*(1), 92–96. https://doi.org/10.55352/mudir.v5i1.33

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard University Press. https://books.google.co.id/books?id=RxjjUefze_oC

Zhang, Y., Su, Y., Liu, Y., Wang, X., Burgess, J., Sui, E., Wang, C., Aklilu, J., Lozano, A., & Wei, A. (2025). Automated generation of challenging multiple-choice questions for vision language model evaluation. *arXiv Preprint arXiv:2501.03225*