IJASE
Indonesian Journal of Applied Science and Engineering
Vol. 5, No. 1, 2027, pp. 1-15
ISSN 3030-9174
https://e-journal.ptti.info/index.php/ijase

A Classical Test Theory Analysis of an English Achievement Test

Rezkilaturahmi [a,1,*], Edi Istiyono [a,2], Widihastuti [a,3], Muhammad Istiqlal [a,4], Oscar Ndayizeye [b,5], John Rey Osben Pelila [c], Muthmainnah [d]

[a] Department of Educational Research and Evaluation, Universitas Negeri Yogyakarta, Indonesia
[b] Department of English Language Education, Hebei Foreign Studies University, Shijiazhuang, China
[c] Department of English, College of Arts and Humanities at Benguet State University, Benguet, Philippines
[d] Department of Islamic Religious Education, Faculty of Islamic Religious Education, Kendari, Indonesia
[1] rezkilaturahmi98@gmail.com; [2] edi_istiyono@uny.ac.id; [3] widihastuti@uny.ac.id;
[4] muhammad.istiqlal2023@student.uny.ac.id; [5] oscarndayiz@gmail.com
* Corresponding Author

ARTICLE INFO

Article history
Received August 06, 2026
Revised September 10, 2026
Accepted October 25, 2026

Keywords
Classical Test Theory;
English Test;
Assessment;
Reliability;
Item Difficulty;
Item Discrimination

©2025 The Author.
This is an open-access article under the CC–BY license.

ABSTRACT

The quality of assessment items is an important aspect of ensuring that English language tests provide reliable and meaningful information about students' achievement. This study aimed to analyze the quality of a 15 item English language achievement test using Classical Test Theory (CTT). The analysis focused on test reliability, item difficulty, item discrimination, and the effect of item deletion on internal consistency. A quantitative approach was employed using responses from 96 students from two public senior high schools in Muna Regency, Southeast Sulawesi, Indonesia. The data were analyzed using descriptive item analysis and Cronbach's alpha to evaluate the internal consistency of the assessment. The results indicated that the assessment demonstrated satisfactory internal consistency. The item difficulty analysis showed variation across the items, with items ranging from very difficult to moderate. The discrimination analysis revealed that several items demonstrated good to very good discrimination, while several items showed poor discrimination and require further review. The alpha if deleted analysis indicated that removing individual items did not improve the overall internal consistency of the assessment. These findings suggest that the assessment has an acceptable foundation but still requires item-level refinement. Items with satisfactory discrimination can be retained, whereas poorly discriminating items should be reviewed in terms of item wording, answer keys, distractors, and alignment with the intended learning objectives. The study highlights the usefulness of CTT as a practical approach for identifying problematic items and supporting the improvement of English language assessment quality.

1. Introduction

In the 21st century, assessment has become an essential component of educational practice because it provides systematic information about students' knowledge, skills, competencies, and learning progress. Assessment is not merely a process of assigning scores to students but also an important mechanism for determining whether the intended learning objectives have been achieved [1][2][3]. Through assessment, teachers can identify students' strengths and weaknesses, evaluate the effectiveness of instructional strategies, and make informed decisions about subsequent learning activities. [4][5][6] emphasize that assessment related skills are essential for both teachers and students because assessment provides important information about students' abilities and learning

https://doi.org/10.59247/ijase.v5i1.181

development. Similarly, [7][8][9][10] argue that assessment provides feedback that can be used to improve teaching and learning processes and to ensure that instructional practices remain aligned with educational objectives. In other words, the role of assessment has become increasingly important as contemporary education emphasizes not only the acquisition of factual knowledge but also the development of higher-order thinking, problem-solving, communication, and the ability to apply knowledge in meaningful situations.

Consequently, assessment needs to provide valid and reliable evidence of students' actual competencies. [11][12][13][14] explain that assessment is closely related to evaluating the effectiveness of the learning process and determining the competencies acquired by students after participating in instruction. Assessment also enables teachers to collect information from different sources and through various approaches. In addition, [15][16][17] emphasize the importance of using appropriate assessment approaches to obtain information about students' learning outcomes. Thus, assessment should be understood as a continuous and systematic process of collecting, interpreting, and using evidence about student learning rather than simply as an activity conducted at the end of instruction.

One important function of assessment is assessment for learning, which places assessment as an integral part of the learning process. Assessment for learning focuses on obtaining information about students' knowledge, skills, understanding, and learning difficulties so that teachers can use the information to improve subsequent instruction. [18] notes that assessment for learning provides information that can help teachers adjust instructional strategies according to students' learning needs. In this regard, assessment can be integrated throughout the curriculum through classroom activities, assignments, quizzes, formative tests, and other forms of evidence. It can also be implemented at important points in the learning process, including examinations conducted at the end of a learning period. [8][9][19] further emphasize that the purpose of assessment extends beyond measuring students' knowledge because assessment can provide feedback for improving instructional planning and learning outcomes. Therefore, high-quality assessment should enable teachers not only to determine what students know but also to understand how well they can apply their knowledge.

Tests constitute one of the most frequently used forms of assessment in educational settings. A test can be conceptualized as a systematic set of questions or tasks designed to obtain information about an individual's knowledge, skills, or proficiency within a particular domain. [3][20][21] describe a test as a measurement tool used to determine an individual's proficiency according to predetermined criteria. Similarly, [22] explain that tests are used to assess students' abilities through a series of questions administered under specified conditions and within a particular period. Tests therefore provide structured evidence that can be used to determine whether students have achieved predetermined learning competencies. [23] further indicates that tests play an important role in identifying students' competencies and monitoring their progress following a learning period or course. Thus, although tests are widely used in education, the quality of the information obtained from a test depends strongly on the quality of its items. A test consisting of poorly constructed items may produce scores that do not accurately represent students' actual abilities. Therefore, teachers and test developers need to ensure that test items are appropriately constructed, relevant to the intended learning outcomes, and capable of distinguishing differences in students' levels of achievement.

Furthermore, [3][24][25] emphasize the importance of constructing effective assessments because test results are frequently used to support educational decisions. However, several criteria that should be considered when developing language tests, including validity, reliability, practicality, and the appropriateness of the assessment procedures. The development of high-quality test items, however, remains a considerable challenge for educators. [14] notes that constructing tests for educational evaluation is a longstanding challenge because test developers must ensure that questions appropriately represent the intended content and competencies. [26] further point out that developing effective and fair questions requires careful consideration of the characteristics of the items and the target population. An item should neither be unnecessarily difficult nor excessively easy, and it should be capable of differentiating students according to their level of mastery. Consequently, empirical

item analysis is required after an assessment has been administered to determine whether individual items function as intended.

Two important indicators in item analysis are item difficulty and item discrimination [27]. Item difficulty provides information about the proportion of students who answer an item correctly and therefore indicates the relative level of challenge presented by an item. An assessment containing only extremely easy or extremely difficult items may provide limited information about differences in students' achievement. A balanced distribution of item difficulty is therefore important for developing an assessment that can adequately represent different levels of student performance [28][29]. In addition to difficulty, item discrimination is an important indicator of item quality because it shows the extent to which an item can differentiate between students who have higher levels of overall performance and those with lower levels of performance. Items with appropriate discrimination contribute more effectively to the measurement of individual differences, whereas items with very low or negative discrimination may indicate problems in item construction, content alignment, distractor functioning, or interpretation by students.

The need for systematic item analysis is particularly important in language assessment. English language assessment involves multiple dimensions of language ability, including vocabulary, grammar, reading comprehension, interpretation, and the ability to understand and use language in context [5][30]. In educational contexts, English tests are often used to determine students' achievement of learning competencies and to provide information for instructional decision-making. However, the usefulness of an English test depends not only on whether the test covers the intended curriculum but also on whether its individual items function adequately. An item that does not discriminate effectively or that is substantially more difficult or easier than intended may influence the accuracy of the resulting test scores [4][31]. Consequently, empirical analysis of English test items is necessary to provide evidence regarding the quality and functioning of the assessment.

Classical Test Theory (CTT) is one of the established approaches used to evaluate the quality of educational tests. CTT provides relatively straightforward procedures for examining characteristics such as test reliability, item difficulty, item discrimination, and item-total relationships [32]. One of the main advantages of CTT is its practical applicability in educational settings because its procedures can be implemented using relatively accessible statistical techniques. In addition, [33][34] through CTT, teachers and researchers can identify items that perform well and items that require revision, replacement, or removal. This makes CTT particularly useful during the development and refinement of classroom-based and educational assessments. Despite the development of more sophisticated psychometric approaches, CTT remains relevant for preliminary item analysis because it provides direct information about how test items perform within a particular sample. [35][36] The analysis of Cronbach's alpha, item difficulty, and discrimination can provide complementary evidence regarding the internal consistency and quality of individual items. Reliability analysis indicates the consistency of scores across items, while item-level statistics provide information about whether individual questions contribute appropriately to the overall assessment. Therefore, combining test-level and item-level analyses can provide a more comprehensive evaluation of an assessment instrument.

In the Indonesian educational context, the quality of classroom and school-based assessments remains an important issue because assessment results are frequently used to make decisions concerning students' achievement and instructional improvement. Teachers are expected to develop assessments that are aligned with learning objectives and capable of providing meaningful information about student performance [37][38]. However, the development of test items is not always followed by systematic empirical analysis of item functioning. As a consequence, items that are too easy, too difficult, or unable to discriminate effectively may remain in an assessment without adequate evidence regarding their psychometric quality. This condition highlights the importance of conducting empirical item analysis as part of the assessment development and quality assurance process. For English language assessment, this issue is particularly relevant because students' responses may be influenced by linguistic complexity, vocabulary familiarity, contextual interpretation, and the cognitive demands of the questions. An item may appear appropriate from a content perspective but perform poorly statistically when administered to students. Therefore, expert judgment during item

development should be complemented by empirical evidence obtained from student responses. The integration of theoretical and empirical evaluation can help ensure that test items are both substantively appropriate and psychometrically functional.

Although empirical item analysis has been widely recognized as an important component of test evaluation, the quality of a test cannot be determined solely from the number of items or from the content coverage established during test construction. In particular, a short English assessment consisting of only 15 dichotomously scored items requires careful empirical examination because each individual item contributes relatively substantially to the overall test score and internal consistency. A poorly functioning item may therefore have a noticeable effect on the reliability and interpretability of the assessment. In addition, an item that appears appropriate from a content perspective may not necessarily function adequately when administered to students. Therefore, analyzing the performance of each of the 15 items is necessary to determine whether the items provide consistent and meaningful measurement evidence.

However, an important gap remains in understanding how the individual items in a short dichotomously scored English assessment contribute to the overall quality of the test. In particular, examining item difficulty and discrimination together with the effect of item deletion on internal consistency provides information not only about whether individual items function appropriately, but also about whether retaining or removing a particular item affects the consistency of the assessment as a whole. This item level and test level perspective is important because an item may demonstrate acceptable characteristics in isolation but still contribute differently to overall test reliability. Thus, the novelty of this study lies in providing an integrated empirical evaluation of the 15 item English assessment by examining item difficulty, item discrimination, overall reliability, and changes in internal consistency associated with item deletion. This approach provides more specific evidence for identifying which items should be retained, revised, or considered for removal.

Based on this background, the present study focuses on the analysis of a dichotomously scored English language assessment using Classical Test Theory. The study examines the overall reliability of the assessment and evaluates individual items in terms of item difficulty, discrimination index, and the effect of item deletion on internal consistency. More broadly, the contribution of this research is to provide empirical evidence on the quality of a dichotomously scored English language assessment using Classical Test Theory (CTT). Specifically, the study contributes a systematic evaluation of test reliability, item difficulty, item discrimination, and the effect of item deletion on internal consistency. The findings provide practical information for identifying items that function adequately and items that require revision, thereby supporting the development of more reliable and effective English language assessments.

2. Method

This study employed a quantitative approach to analyze the quality of an English achievement test using Classical Test Theory (CTT). The study was conducted in two public senior high schools in Muna Regency, Southeast Sulawesi, Indonesia. The schools were purposively selected because they used an English achievement test developed by the English Teachers Working Group (Musyawarah Guru Mata Pelajaran/MGMP) in Muna Regency. Thus, the selection of the schools was based on their use of the MGMP developed test, which was the focus of the present psychometric analysis. The participants consisted of 96 students from the two participating schools who completed the multiple-choice test used in the study.

The test items were developed based on three achievement indicators namely text structure, social function, and linguistic elements. In addition, these indicators represented the learning competencies targeted by the English test used by the participating schools. The present study focused on the multiple-choice items included in the test trial. A total of 15 multiple choice items were analyzed. The number of items was determined by the multiple-choice items available in the test used for the trial and was not arbitrarily selected by the researchers after data collection. The items were therefore analyzed as an empirical set of test items used to obtain evidence regarding their psychometric

functioning. The data analysis was conducted using the *R* statistical programming environment. The CTT analysis examined the reliability and item level characteristics of the English achievement test. Test reliability was evaluated using Cronbach's alpha to determine the internal consistency of the 15 items. At the item level, item difficulty was calculated based on the proportion of students who answered each item correctly, while item discrimination was examined using the corrected item-total correlation to determine the ability of each item to differentiate between students with higher and lower overall test performance. In addition, alpha if item deleted was calculated to identify whether the removal of a particular item would increase the overall reliability of the test.

The analysis also included descriptive statistics of the total test scores, including the mean and standard deviation. The results of the item analysis were presented in tables and graphical visualizations generated using *R* to facilitate the interpretation of item functioning and identify items that were satisfactory, required review, or needed revision. The 15 multiple-choice items were distributed across three achievement indicators. Eight items assessed text structure (items 1–8), two items assessed social function (items 9–10), and five items assessed linguistic elements (items 11–15). The item specifications are presented in Table 1.

Table 1. Item Specification

Questions Type	Number of Question	Item Number
Text Structure	8	1, 2, 3, 4, 5, 6, 7 and 8
Social Function	2	9 and 10
Linguistic Elements	5	11, 12, 13, 14 and 15

The use of *R* enabled the analysis to be conducted systematically and reproducibly, while the combination of reliability and item-level analyses provided comprehensive evidence of the quality of the English achievement test.

3. Results and Discussion

3.1. Cronbach's Alpha

The internal consistency of the English language assessment was examined using Cronbach's alpha. The analysis was conducted on responses from 96 students who completed 15 dichotomously scored items. The resulting Cronbach's alpha coefficient was 0.627 (Fig. 1), indicating a satisfactory level of internal consistency among the test items. This result suggests that the items generally measure a relatively consistent underlying construct and that the test can provide reasonably reliable scores for the group of students involved in the study. The obtained reliability coefficient is important because the consistency of test scores is a fundamental consideration in evaluating the quality of an educational assessment. A reliability coefficient close to 0.70 indicates that the observed variation in students' scores is sufficiently stable to support the interpretation of the assessment results. In this study, the value of 0.627 demonstrates that the 15 items work together adequately as a test set. Although the coefficient does not indicate perfect reliability, it provides evidence that the assessment has an acceptable level of internal consistency for an initial evaluation of item quality.

The reliability result also provides an overall picture before examining the functioning of individual items. Therefore, the reliability coefficient of 0.627 indicates that the assessment has a sufficiently consistent measurement structure. This finding supports the use of the test for further item-level evaluation and provides an empirical basis for examining which items should be retained, revised, or reconsidered in subsequent test development.

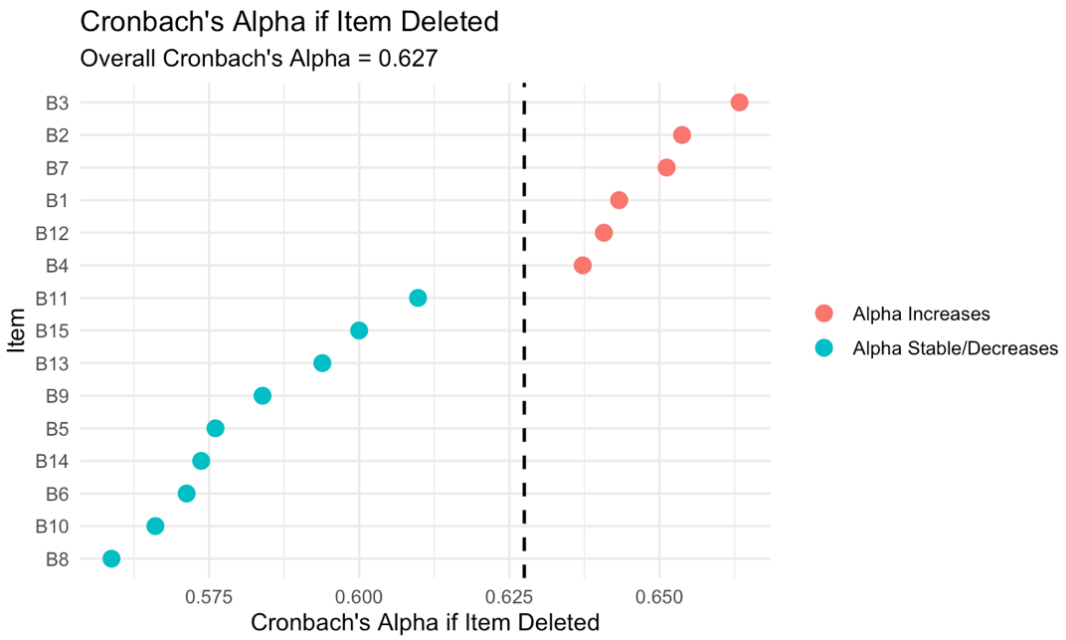


Fig. 1. Cronbach's Alpha of the test

3.2. Item Difficulty

Item difficulty was examined to determine the proportion of students who successfully answered each of the 15 dichotomously scored items (Fig. 2). For dichotomous items, the item difficulty index was calculated based on the proportion of students obtaining a correct response. A higher difficulty index indicates that a larger proportion of students answered the item correctly, meaning that the item is relatively easier. Conversely, a lower index indicates that fewer students answered correctly and that the item is relatively more difficult. The results demonstrate variation in item difficulty across the assessment. This variation is important because a good achievement test should not consist entirely of items that are either extremely easy or extremely difficult. A range of difficulty levels allows the assessment to capture differences in students' levels of achievement more effectively. Items with different levels of difficulty can provide information about students across a broader range of ability levels.

From a test-development perspective, item difficulty should not be interpreted independently of the purpose of the assessment. An item with a relatively high proportion of correct responses is not automatically a poor item. Easy items may be appropriate when they represent essential learning outcomes or foundational competencies. Similarly, a difficult item may be valuable when it is intended to measure higher-order understanding. Therefore, decisions regarding item retention should consider both statistical characteristics and the substantive relevance of each item to the intended learning outcomes. The variation in item difficulty also contributes to the overall score distribution. When items have different difficulty levels, students are more likely to obtain differentiated total scores rather than being concentrated at either the maximum or minimum score. Consequently, the item difficulty analysis provides evidence that should be considered together with the total score distribution when evaluating the overall functioning of the assessment.

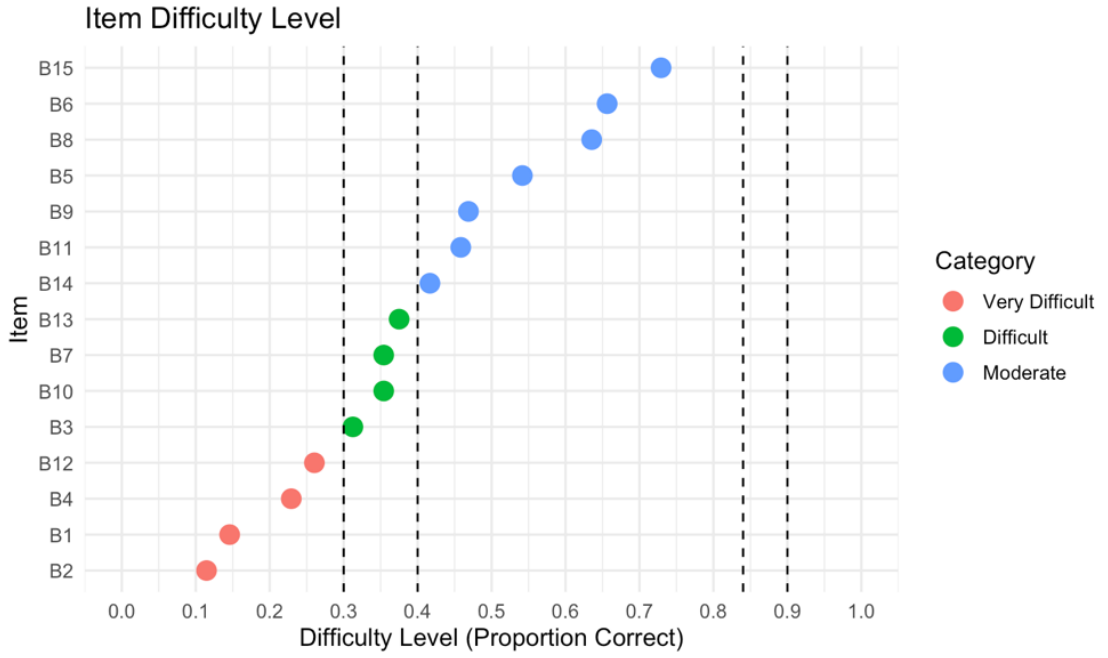


Fig. 2. Item difficulty

3.3. Item Discrimination

Item discrimination was analyzed to determine the extent to which each item was able to distinguish between students with relatively higher overall test performance and those with relatively lower overall performance (Fig. 3). In this analysis, the corrected item-total correlation was used as an indicator of item discrimination. The correction was applied by correlating each individual item with the total test score excluding that item. This approach prevents the item from contributing directly to the total score against which it is correlated. The results indicate that the discrimination indices varied considerably across the 15 items. Several items demonstrated relatively strong positive relationships with the overall test score, suggesting that students who performed well on the test were also more likely to answer these items correctly. Such items can be considered to function effectively because they contribute to the differentiation of students according to their overall level of achievement.

Items with stronger discrimination are particularly valuable in an achievement assessment because they provide greater information for distinguishing students who have mastered the assessed competencies from those who have not. A positive item-total relationship indicates that the item is generally functioning in the expected direction. In practical terms, students with higher total test scores tend to have a greater probability of answering these items correctly. The analysis also identified several items with relatively weak discrimination. Weak discrimination may indicate that an item does not sufficiently differentiate between higher- and lower-performing students. This situation may occur for several reasons. The item may be too easy or too difficult, the distractors may not function effectively, the wording may be ambiguous, or the content of the item may not be sufficiently aligned with the construct measured by the remaining items. Therefore, items with weak discrimination should not immediately be discarded solely based on their statistical index. Instead, they should undergo substantive and qualitative review.

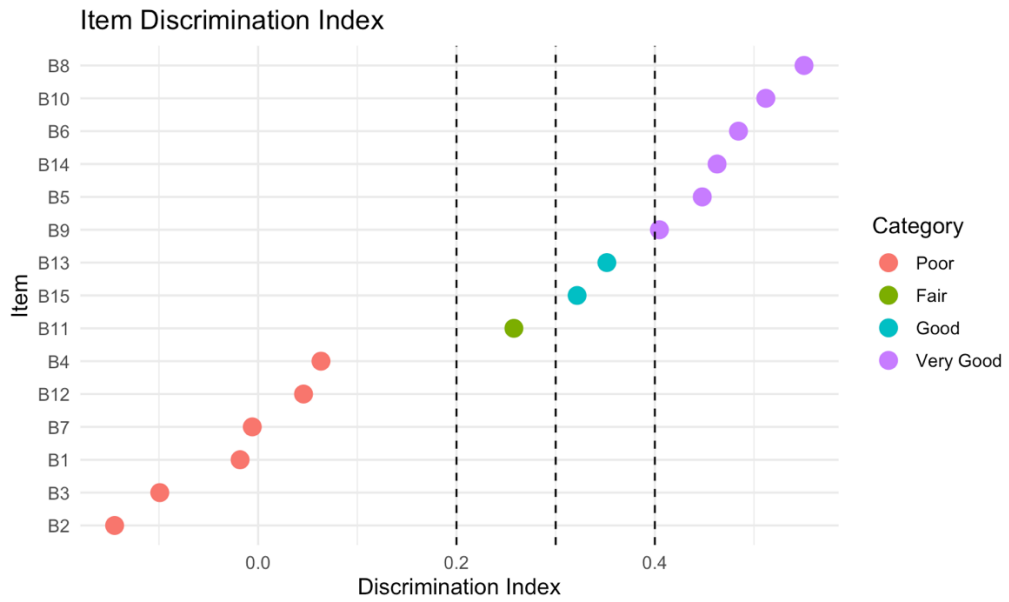


Fig. 3. Item discrimination test

3.4. Total Scores

Table 2 presents the item-level characteristics of the English test based on Classical Test Theory. The results show that the items vary in their level of difficulty and ability to discriminate between students with different levels of achievement. Overall, the test contains a mixture of challenging and moderately difficult items, which provides variation in the level of demand across the assessment. The discrimination results show that several items function effectively, indicating that they are able to distinguish students according to their overall test performance. However, several items demonstrate weak discrimination and therefore require further examination. The presence of poor discrimination, particularly among some of the more difficult items, suggests that difficulty alone does not necessarily indicate good item quality.

Table 2. Item Categories of English Test

Item	Item Difficulty Index	Item Difficulty	Item Discrimination	Discrimination Category	Alpha if Delected
B1	0.146	Very Difficult	-0.018	Poor	0.643
B2	0.115	Very Difficult	-0.145	Poor	0.654
B3	0.312	Difficult	-0.099	Poor	0.663
B4	0.229	Very Difficult	0.063	Poor	0.637
B5	0.542	Moderate	0.448	Very Good	0.576
B6	0.656	Moderate	0.484	Very Good	0.571
B7	0.354	Difficult	-0.006	Poor	0.651
B8	0.635	Moderate	0.550	Very Good	0.559
B9	0.469	Moderate	0.405	Very Good	0.584
B10	0.354	Difficult	0.512	Very Good	0.566
B11	0.458	Moderate	0.258	Fair	0.610
B12	0.260	Very Difficult	0.046	Poor	0.641
B13	0.375	Difficult	0.352	Good	0.594
B14	0.417	Moderate	0.463	Very Good	0.574
B15	0.729	Moderate	0.322	Good	0.600

The alpha-if-deleted results further indicate that removing individual items would not improve the overall internal consistency of the test. Therefore, item revision should not be based solely on reliability statistics. Instead, items with weaker performance should be reviewed in terms of wording, answer options, distractors, answer keys, and alignment with the intended learning objectives.

Therefore, Table 2 provides evidence that the assessment has a reasonable foundation but would benefit from targeted item refinement before further use.

The distribution of total scores was examined to describe the overall performance of the 96 students on the 15-item assessment (Fig. 4). Because each item was scored dichotomously, students could obtain total scores ranging from 0 to 15**.** The total score represents the number of items answered correctly and therefore provides an overall descriptive indicator of students' achievement on the assessment. The analysis showed a variation in students' total scores, indicating that the assessment was able to produce differentiated levels of observed performance. The distribution of scores is important because an assessment should ideally provide sufficient variation to distinguish among students rather than producing highly concentrated scores at one end of the scale. A highly concentrated distribution, particularly near the maximum score, may indicate that the assessment is too easy for the target group. Conversely, concentration near the minimum score may indicate that the assessment is excessively difficult. The observed total score distribution should also be considered in relation to the item difficulty results. The presence of items with different difficulty levels contributes to the ability of the test to generate variation in students' total scores. Students who possess stronger mastery of the assessed English competencies are expected to answer a greater number of items correctly, resulting in higher total scores. Students with lower levels of mastery are expected to obtain fewer correct responses.

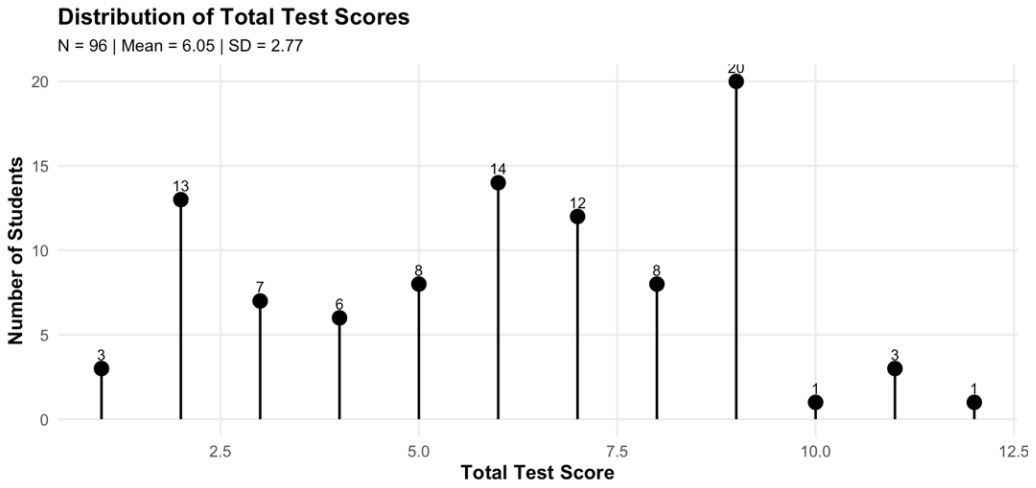


Fig. 4. Distribution of total scores

The results show that the test had a satisfactory reliability coefficient (Cronbach's alpha = 0.797). The item difficulty analysis revealed that four items were very difficult (B1, B2, B4, and B12), three were difficult (B3, B7, and B10), and eight were moderate (B5, B6, B8, B9, B11, B13, B14, and B15). This variation indicates that the test includes different levels of difficulty, allowing it to assess students with varying levels of English proficiency. Regarding item discrimination, six items (B5, B6, B8, B9, B10, and B14) demonstrated very good discrimination. B13 and B15 were classified as good, while B11 showed fair discrimination. In contrast, B1, B2, B3, B4, B7, and B12 showed poor discrimination. These findings suggest that most items with good discrimination can be retained, whereas items with poor discrimination need further review, particularly in terms of wording, answer keys, and distractor quality. The alpha results was Cronbach's alpha of 0.627. This indicates that removing any single item would not improve the test's internal consistency. Therefore, decisions about item revision should consider both discrimination results and content relevance rather than relying solely on reliability statistics. Therefore, the findings indicate that the assessment has satisfactory internal consistency, with several items demonstrating good psychometric characteristics. However, items with poor discrimination require further evaluation to improve the overall quality of the test.

The findings indicate that the English language assessment has generally acceptable quality, although the performance of individual items varies. The satisfactory internal consistency suggests

that the items are sufficiently related to measure students' English achievement [39][40]. However, the item level analysis provides more detailed information about the strengths and weaknesses of the assessment. This finding supports the importance of examining individual items rather than relying only on the overall reliability of a test [41][42]. The difficulty analysis showed that the items covered different levels of difficulty. This variation is useful because an assessment should not consist only of easy or difficult questions. A combination of difficulty levels allows the test to provide information about students with different levels of achievement. However, difficult items should be examined carefully when they are accompanied by poor discrimination. An item may be difficult because it measures a challenging competency, but it may also be difficult because of unclear wording, inappropriate distractors, unfamiliar content, or an incorrect answer key.

The discrimination results provide an important explanation of the quality of the individual items. Several items demonstrated good and very good discrimination, indicating that these items were able to distinguish students with higher overall achievement from those with lower achievement. These items can be considered appropriate candidates for retention in future test administrations. Their characteristics may also provide useful examples for developing or revising other items. In contrast, several items showed poor discrimination, including items with negative discrimination. Such findings require particular attention because a well-functioning achievement item is generally expected to be answered correctly more often by students with higher overall achievement. Poor or negative discrimination may indicate problems with the item construction or content. Therefore, these items should be reviewed in terms of the clarity of the question, suitability of the answer options, accuracy of the answer key, and alignment between the item and the intended learning objective [43][44][45]. The relationship between difficulty and discrimination is also important. The results show that a difficult item does not necessarily indicate a poor quality item. Some relatively difficult items were able to discriminate well between students. This suggests that difficulty alone should not be used as the basis for removing an item. Instead, item difficulty should be interpreted together with discrimination and the substantive purpose of the item. An item that is challenging but discriminates well may be valuable for identifying students who have achieved a higher level of competence.

The alpha if deleted analysis provides further support for retaining a comprehensive approach to item evaluation. The removal of individual items did not produce an improvement in the overall internal consistency of the test. Therefore, items should not be removed simply because their individual discrimination is lower than that of other items. Instead, statistical evidence should be combined with qualitative evaluation [45]. This is particularly important because an item may represent an important learning objective even when its statistical performance requires improvement. These findings are consistent with the principles of classical test theory, in which item difficulty and discrimination are commonly used to evaluate the functioning of test items [46][47]. Previous measurement literature emphasizes that item statistics should support, rather than replace, expert judgment and content analysis. Consequently, the statistical results obtained in this study should be followed by a review of the content and construction of the problematic items.

From an educational perspective, the findings also highlight the importance of systematic test development. Teachers and test developers need to examine not only whether students answer an item correctly but also whether the item contributes meaningfully to the measurement of achievement. Items should represent the intended learning outcomes and provide an appropriate level of challenge for the target students. Careful construction of stems, distractors, and answer keys is therefore essential to ensure that item performance reflects students' actual English competence rather than weaknesses in item construction [5][48][49]. In other words, the study demonstrates that CTT provides practical information for improving English language assessment. The analysis makes it possible to identify items that function well and items that require revision. Rather than eliminating problematic items immediately, test developers should first investigate the possible causes of poor performance and revise the items accordingly. This process can contribute to the development of a more balanced, reliable, and informative English language assessment.

4. Conclusion

This study evaluated the psychometric quality of a 15 item English language assessment using Classical Test Theory. The findings indicate that the quality of the assessment varies across individual items. Several items demonstrated acceptable levels of difficulty and discrimination, whereas six items showed poor discrimination, including items with negative discrimination values. These findings indicate that the test should not be considered uniformly acceptable in its current form. Instead, the quality of the assessment needs to be considered at the item level, with particular attention to items that do not adequately differentiate students according to their overall achievement. The item level findings provide a basis for decisions regarding item retention and revision. Items demonstrating appropriate difficulty and positive discrimination can be retained for further use, whereas the six items with poor discrimination should be reviewed and revised before being used in subsequent assessments. The revision process should examine the clarity of item wording, the appropriateness of the distractors, the accuracy of the answer keys, and the alignment between each item and the intended learning indicators. Therefore, item analysis should be used together with content review and expert judgment rather than relying solely on the overall reliability coefficient.

This study has several limitations. First, the analysis was based on a relatively small sample of 96 students from only two public senior high schools in Muna Regency, which may limit the generalizability of the findings to other student populations and school contexts. Second, the assessment consisted of only 15 multiple choice items, which limits the breadth of content and competencies that can be represented and may influence the stability of the reliability estimate. Third, the analysis was conducted using Classical Test Theory, whose item statistics are dependent on the characteristics of the sample used in the analysis. Consequently, the findings should be interpreted within the context of the participating students and schools.

Future research should replicate the analysis using a larger and more diverse sample of students and schools to obtain more stable evidence of item functioning. Further studies should also consider more advanced psychometric approaches, particularly Item Response Theory (IRT) and the Rasch model, to examine item characteristics from a model-based measurement perspective. Combining CTT with IRT/Rasch analysis may provide stronger evidence for evaluating item difficulty, discrimination, item fit, and the overall measurement quality of the English assessment. Such evidence can support more rigorous decisions regarding the retention, revision, or replacement of items in future English language assessments.

Supplementary Materials: No supplementary materials are available for this study.

Author Contribution: Conceptualization, methodology, data collection, data analysis, writing, and revision were carried out by the authors. All authors have read and approved the final version of the manuscript.

Funding: This research received no external funding. The APC was not funded by any external funding source.

Acknowledgment: The authors would like to express their sincere gratitude to the principals of the two participating schools in Muna Regency for granting permission and supporting the data collection process. The authors also thank the teachers and students who willingly assisted with and participated in the research. Special appreciation is extended to the lecturers of the Measurement and Evaluation Theory course for their valuable guidance, feedback, and support throughout the process of developing and completing this article. Their contributions were invaluable to the completion of this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Y. Liu *et al.*, "International Psychogeriatrics Harmonized Cognitive Assessment Protocol in the English Longitudinal Study of Ageing: Contrasting approaches to evaluation of factor structure," *Int. Psychogeriatrics*, p. 100042, 2025, https://doi.org/10.1016/j.inpsyc.2025.100042.

[2] A. Müller, A. Brenner, and M. Barghamadi, "Benchmarking English proficiency standards in International Nursing Education: A cross-sectional analysis of registration and university entry requirements," *J. Prof. Nurs.*, vol. 67, pp. 256–266, 2026, https://doi.org/10.1016/j.profnurs.2026.08.019.

[3] W. Yao, "Multimodal Collaborative Assessment of English Speech Based on LLM and Fuzzy Logic," *Syst. Soft Comput.*, vol. 9, p. 200620, 2026, https://doi.org/10.1016/j.sasc.2026.200620.

[4] P. A. Chala-bejarano, M. Rodríguez-uribe, and S. Patricia, "Pre-service english teachers' perceptions of language assessment in a colombian language teacher education program," *International Journal of Educational Research Open*, vol. 8, 2025, https://doi.org/10.1016/j.ijedro.2024.100405.

[5] L. Stoffelsma, I. N. Mwinlaaru, and W. Spooren, "Testing university students' morphological awareness of English as a second-language," *System*, vol. 130, p. 103626, 2025, https://doi.org/10.1016/j.system.2025.103626.

[6] S. Park, S. B. Piasta, and P. M. Sayer, "Early Childhood Research Quarterly An initial yet rigorous test of multisensory alphabet instruction for english monolingual and emergent bilingual children," *Early Child. Res. Q.*, vol. 72, pp. 328–338, 2025, https://doi.org/10.1016/j.ecresq.2025.04.008.

[7] C. Mallart *et al.*, "Research Methods in Applied Linguistics Assessing the validity of syntactic alternations as criterial features of proficiency in L2 writings in English," *Res. Methods Appl. Linguist.*, vol. 4, no. 3, p. 100238, 2025, https://doi.org/10.1016/j.rmal.2025.100238.

[8] B. Naismith, Y. Attali, and G. T. Laflair, "The impact of task duration on the scoring of independent writing responses of adult L2-English writers," *Assess. Writ.*, vol. 62, p. 100895, 2024, https://doi.org/10.1016/j.asw.2024.100895.

[9] L. Chih-chung and Y. Fu-yun, "Social Sciences & Humanities Open Impact of scaffold structures on student-generated questions in English grammar learning," *Soc. Sci. Humanit. Open*, vol. 12, no. 1, p. 102126, 2025, https://doi.org/10.1016/j.ssaho.2025.102126.

[10] K. Barkaoui, J. Holmes, and S. Van Viegen, "Journal of English for Academic Purposes Pre-admission EAP course completion as evidence of English language proficiency for university admission: A longitudinal evaluation study," *J. English Acad. Purp.*, vol. 77, p. 101555, 2025, https://doi.org/10.1016/j.jeap.2025.101555.

[11] S. Nabhan and A. Habok, "Development and validation of the Teacher Multiliteracies Scale (TMS) for English language teacher education," *System*, vol. 131, 2025, https://doi.org/10.1016/j.system.2025.103673.

[12] E. Yisihak, A. Damtew, and K. Terms, "Social Sciences & Humanities Open Primary school English teachers' practice of teaching early grade reading," *Soc. Sci. Humanit. Open*, vol. 9, p. 100840, 2024, https://doi.org/10.1016/j.ssaho.2024.100840.

[13] A. Benati and G. Benati, "The online primary and secondary transfer-of-training effects of structured input on the acquisition of English past tense and third person singular present tense markers: A self-paced reading study measuring accuracy and response time," *Ampersand*, vol. 15, p. 100229, 2025, https://doi.org/10.1016/j.amper.2025.100229.

[14] P. Heisterkamp, A. Schüppert, V. Baaijen, and M. Michel, "Learning in English : A research synthesis on European countries with advanced English skills," *Journal of English for Academic Purposes*, vol. 77, 2025, https://doi.org/10.1016/j.jeap.2025.101564.

[15] E. C. M. Simatupang, "Social Sciences & Humanities Open Task-based learning and translanguaging: A dual approach to empowering hotel staff in real-world English communication," *Soc. Sci. Humanit. Open*, vol. 12, p. 102005, 2025, https://doi.org/10.1016/j.ssaho.2025.102005.

[16] R. Zhang, D. Zou, H. Xie, and F. Lee, "Computers and Education: Artificial Intelligence Engagements with GPT responses and learner prompts in ChatGPT-based learning of English argumentative writing logic and their impacts," *Comput. Educ. Artif. Intell.*, vol. 9, p. 100489, 2025, https://doi.org/10.1016/j.caeai.2025.100489.

[17] A. Montakantiwong and N. Funada, “Measuring the Global Englishes mindset: Development and validation of the GEM-Q for EFL learners,” *System*, vol. 134, p. 103847, 2025, https://doi.org/10.1016/j.system.2025.103847.

[18] H. Tran, P. Crosthwaite, Q. Huynh, and P. Pham, “Journal of English for Academic Purposes Students’ self-determination in using machine translation and generative AI tools for English for academic purposes,” *J. English Acad. Purp.*, vol. 78, p. 101578, 2025, https://doi.org/10.1016/j.jeap.2025.101578.

[19] A. Mizumoto, N. Shintani, M. Sasaki, and M. Feng, “Research Methods in Applied Linguistics Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment,” *Res. Methods Appl. Linguist.*, vol. 3, no. 2, p. 100116, 2024, https://doi.org/10.1016/j.rmal.2024.100116.

[20] M. Sánchez-Torrón, E. Ipek, and V. E. Raído, "Creating terminological correspondence recognition tests with GPT-4: a case study in English-to-Turkish translations in the engineering domain," *International Journal of Artificial Intelligence in Education*, vol. 35, no. 4, pp. 2185-2201, 2025, https://doi.org/10.1007/s40593-025-00465-x.

[21] M. P. Saadi *et al.*, “Assessment of Interrater Reliability in Point-of-Care Ultrasound for Assessing Congestion in Cardiovascular Intensive Care,” *ABC Imagem Cardiovascular*, vol. 38, no. 2, 2025, https://doi.org/10.36660/abcimg.20250022i.

[22] M. Ibnu, B. Indriyani, H. Inayatullah, and Y. Guntara, “Aplikasi Rasch Model: Pengembangan Instrumen Tes Untuk Mengukur Miskonsepsi Mahasiswa Pada Materi Mekanika,” *Prosiding Seminar Nasional Pendidikan FKIP*, vol. 2, no. 1, pp. 205–210, 2019, https://jurnal.untirta.ac.id/index.php/psnp/article/view/5669.

[23] P. Levrai and P. García, “Journal of English for Academic Purposes Understanding the collaborative assessment identity of English for academic purposes practitioners in higher education,” *J. English Acad. Purp.*, vol. 79, p. 101621, 2026, https://doi.org/10.1016/j.jeap.2025.101621.

[24] Y. Hwang and H. Lee, “Computers & Education: X Reality An exploratory study on the development of pre-service English teachers’ VR-contextualized testing prototype,” *Comput. Educ. X Real.*, vol. 8, p. 100126, 2026, https://doi.org/10.1016/j.cexr.2025.100126.

[25] Y. Kuo, S. I. C. Wang, E. Zhi, and F. Liu, “Peer assessment using one-to-one tablets to improve the English oral presentation of fourth-grade students in Taiwan,” *Comput. Educ. Open*, vol. 8, p. 100262, 2025, https://doi.org/10.1016/j.caeo.2025.100262.

[26] Z. Erdil-Moody and A. S. Thompson, “Exploring motivational strategies in higher education: Student and instructor perceptions,” *Eurasian J. Appl. Linguist.*, vol. 6, no. 3, pp. 387–413, 2020, https://doi.org/10.32601/ejal.834670.

[27] Y. Altın, B. Ozkurt, and H. F. Küçükibis, “Examination of the psychometric properties of physical activity enjoyment scale (PACES) based on classical test theory and item response theory,” *Acta Psychologica*, vol. 261, 2025, https://doi.org/10.1016/j.actpsy.2025.105855.

[28] F. Lalot, J. Räikkönen, and S. Ahvenharju, "An item response theory approach to measurement in environmental psychology: A practical example with environmental risk perception," *Journal of Environmental Psychology*, vol. 101, 102520, 2025, https://doi.org/10.1016/j.jenvp.2025.102520.

[29] N. Trajtenberg and P. Ezquerra, “An item response theory approach to punitive attitudes,” *J. Crim. Justice*, vol. 98, p. 102402, 2025, https://doi.org/10.1016/j.jcrimjus.2025.102402.

[30] N. Azizah, M. Suseno, and B. Hayat, “Item analysis of the rasch model items in the final semester exam indonesian language lesson,” *World J. English Lang.*, vol. 12, no. 1, pp. 15–26, Mar. 2022, https://doi.org/10.5430/wjel.v12n1p15.

[31] S. Park, S. B. Piasta, and P. M. Sayer, “Early Childhood Research Quarterly An initial yet rigorous test of multisensory alphabet instruction for english monolingual and emergent bilingual children,” *Early Child. Res. Q.*, vol. 72, pp. 328–338, 2025, https://doi.org/10.1016/j.ecresq.2025.04.008.

[32] L. Y. Tay, C. P. Lim, S. Y. Lye, K. J. Ng, and S. K. Lim, “Open-source learning management system and Web 2.0 online social software applications as learning platforms for an elementary school in Singapore,”

Learn. Media Technol., vol. 36, no. 4, pp. 349–365, 2011, https://doi.org/10.1080/17439884.2011.615322.

[33] R. A. N. Grande *et al.*, "Psychometric testing of the nurse professional competence scale Arabic version: An exploratory and confirmatory factor analyses," *Nurse Educ. Pract.*, vol. 70, 2023, https://doi.org/10.1016/j.nepr.2023.103679.

[34] X. Tang, B. D. Schalet, J. D. Peipert, and D. Cella, "Does Scoring Method Impact Estimation of Significant Individual Changes Assessed by Patient-Reported Outcome Measures? Comparing Classical Test Theory Versus Item Response Theory," *Value Heal.*, vol. 26, no. 10, pp. 1518–1524, 2023, https://doi.org/10.1016/j.jval.2023.06.002.

[35] O. G. Yildirim and N. Ozdener, "Development and Validation of the Programming Anxiety Scale," *Int. J. Comput. Sci. Educ. Sch.*, vol. 5, no. 3, pp. 17–34, May 2022, https://doi.org/10.21585/ijcses.v5i3.140.

[36] A. P. Bal, R. Yilmaz, and V. Atas, "Development of self-efficacy scale of differentiated instruction for teachers," *JRAMathEdu (Journal Res. Adv. Math. Educ.)*, vol. 7, no. 2, pp. 93–104, Apr. 2022, https://doi.org/10.23917/jramathedu.v7i2.16204.

[37] H. Alfiana, H. Karyono, and W. Gunawan, "Analisis Butir Tes Keterampilan Berpikir Kritis dan Pengetahuan Prosedural Grammar Bahasa Inggris," *JINOTEP (Jurnal Inov. dan Teknol. Pembelajaran) Kaji. dan Ris. Dalam Teknol. Pembelajaran*, vol. 8, no. 1, pp. 20–29, 2021, https://doi.org/10.17977/um031v8i12021p020.

[38] M. Muchlisin, D. Mardapi, and F. A. Setiawati, "An analysis of Javanese language test characteristic using the Rasch model in R program," *Res. Eval. Educ.*, vol. 5, no. 1, pp. 61–74, Jun. 2019, https://doi.org/10.21831/reid.v5i1.23773.

[39] S. Menggo and H. C. Darong, "LLT Journal: A Journal on Language and Language Learning http://e-journal.usd.ac.id/index.php/LLT Sanata Dharma University, Yogyakarta, Indonesia," *LLT J. A J. Lang. Lang. Learn.*, vol. 25, no. 1, pp. 132–148, 2022, https://doi.org/10.24071/llt.v25i1.4159.

[40] T. Bahrani, "Technology as an Assessment Tool in Language Learning," *Int. J. English Linguist.*, vol. 1, no. 2, pp. 295–298, 2011, https://doi.org/10.5539/ijel.v1n2p295.

[41] A. S. Konca, Ö. Baltaci, and Ö. F. Akbulut, "Problematic Technology Use Scale for Young Children (PTUS-YC): Validity and Reliability Study," *Int. J. Assess. Tools Educ.*, vol. 9, no. 2, pp. 267–289, Jun. 2022, https://doi.org/10.21449/ijate.888936.

[42] D. Pratama, "Pengembangan Skala Thurstone Metode Equal Appearing Interval untuk Mengukur Sikap Moderasi Beragama Siswa Sekolah Menengah Atas," *J. Psikol. Teor. dan Terap.*, vol. 11, no. 1, p. 71, 2020, https://doi.org/10.26740/jptt.v11n1.p71-82.

[43] J. R. Rico-juan and M. Arevalillo-herráez, "D-CTT: A deterministic and scalable heuristic for real-time item calibration in adaptive learning systems," *Expert Syst. Appl.*, vol. 333, p. 134180, 2027, https://doi.org/10.1016/j.eswa.2026.134180.

[44] A. Muhson, B. Lestari, S. Supriyanto, and K. Baroroh, "The development of practical item analysis program for indonesian teachers," *Int. J. Instr.*, vol. 10, no. 2, pp. 199–210, 2017, https://doi.org/10.12973/iji.2017.10213a.

[45] M. Molaei, P. Fadakar, and M. Shafieetabar, "Psychometric properties and factor structure of the cognitive distortions questionnaire (CD Quest)," *Psychiatry Res. Commun.*, vol. 5, no. 4, p. 100241, 2025, https://doi.org/10.1016/j.psycom.2025.100241.

[46] L. Huang *et al.*, "The Journal of nutrition, health and aging Development validation of a ultra-brief screening tool for sarcopenia in Chinese community-dwelling older adults," *J. Nutr. Heal. aging*, vol. 30, no. 9, p. 100928, 2026, https://doi.org/10.1016/j.jnha.2026.100928.

[47] Z. J. Kunicki, M. R. Schick, N. S. Spillane, and L. L. Harlow, "Addictive Behaviors Reports Creation and validation of the barriers to alcohol reduction (BAR) scale using classical test theory and item response theory," *Addict. Behav. Reports*, vol. 7, pp. 47–52, 2018, https://doi.org/10.1016/j.abrep.2018.01.004.

[48] W. Y. Hwang, T. K. Shih, Z. H. Ma, R. Shadiev, and S. Y. Chen, "Evaluating listening and speaking skills in a mobile game-based learning environment with situational contexts," *Comput. Assist. Lang. Learn.*, vol. 29, no. 4, pp. 639–657, 2016, https://doi.org/10.1080/09588221.2015.1016438.

[49] H. Stadthagen-González, L. López, M. C. Parafita Couto, and C. A. Párraga, "Using two-alternative forced choice tasks and Thurstone's law of comparative judgments for code-switching research," *Linguist. Approaches to Biling.*, vol. 8, no. 1, pp. 67–97, 2018, https://doi.org/10.1075/lab.16030.sta.