

Item Analysis of the Developed Reading Preparation Materials for the English Proficiency Test

^{1*} Issatirrodiah, ² Ahmad Zamzam, ³ Ahmad Junaidi

^{1,2,3} English Education Program, Language and Arts Education Department, Faculty of Teacher Training and Education, University of Mataram, Indonesia

*Corresponding Author e-mail: issatirrodiah257@gmail.com

Received: September 2026; Revised: September 2026; Published: October 2026

Abstract

This study is an integral part of research and development (R&D) on digitalized reading materials for the English proficiency test. It focused on item analysis of the fifty reading comprehension questions embedded in a newly developed Digitalized English Proficiency Test (EPT) Reading Preparation Materials platform. Although previous studies have examined item quality in teacher-made tests, little is known about the statistical quality of reading items developed specifically for a digital EPT preparation platform at the university level; this study addresses that gap and offers novelty by integrating item facility, item discrimination, and distractor efficiency into an item-level decision matrix for platform improvement. Because a fellow researcher independently constructed the items and they had never been empirically evaluated, this study aimed to analyze item facility, item discrimination, and distractor efficiency. It used a quantitative descriptive design and analyzed response data from 38 English Education students who completed all 50 items, following Arikunto's (2018) classical test theory procedures. The results showed a mean item facility index of .64, with 50% of items falling in the medium/acceptable range. The mean item discrimination index was .34, and 109 of 150 distractors (72.7%) were functional. Based on the combined criteria, 6 items (12%) were recommended for retention, 41 items (82%) required revision, and 3 items (6%) were recommended for removal. The main limitation is that reliability was not calculated and the pilot sample was relatively homogeneous in English proficiency, which likely restricted discrimination estimates; therefore, findings should be interpreted as initial, sample-dependent evidence. These findings provide an empirical basis for iteratively improving the platform's item bank before wider implementation.

Keywords: Reading Proficiency Test; Item Analysis; Item Facility; Item Discrimination; Distractor Efficiency.

How to Cite: Issatirrodiah, Zamzam, A., & Junaidi, A. (2026). Item Analysis of the Developed Reading Preparation Materials for the English Proficiency Test. *Journal of Authentic Research*, 5(4), 7247-7258. <https://doi.org/10.36312/jar.v5i4.7016>



<https://doi.org/10.36312/jar.v5i4.7016>

Copyright©2026, Issatirrodiah et al.

This is an open-access article under the [CC-BY-SA](#) License.



INTRODUCTION

Assessment plays a central role in language education because it determines whether students have reached the English proficiency level required for academic advancement (Brown, 2004; Hughes, 2003). At the University of Mataram, students in the English Education Department must pass an English Proficiency Test (EPT) to graduate (Narayana & Soepriyanti, 2023). As digital technology continues to reshape language education, there has been a growing shift from traditional paper-based assessment toward web-based digital learning platforms (Warschauer & Ware, 2006). In line with this trend, a fellow researcher in the English Education Department is currently developing a Digitalized English Proficiency Test Reading Preparation Materials platform, a web-based

application intended to help students practice reading comprehension in preparation for the EPT. As part of this development, the same researcher independently constructed 50 new multiple-choice reading comprehension items based on the EPT syllabus to serve as the platform's main test instrument.

For a digital EPT platform to genuinely support student preparation, the items it contains must be of demonstrably good quality (Brown, 2004; Fulcher, 2010). This requirement is closely tied to the idea of autonomous learning, that is, learners' capacity to manage their own learning process (Little, 2007). Little (2007) argues that autonomous learning is meaningful only when learners have access to assessment instruments that provide accurate feedback about their actual proficiency; if items are too easy, too difficult, or fail to discriminate between stronger and weaker learners, the feedback they generate becomes misleading. This concern is reinforced by Brown (2004) and Hughes (2003), who both argue that valid and reliable test instruments are basic prerequisites for meaningful language assessment, and that tests lacking these qualities risk undermining learners' motivation and confidence.

The central problem this study addresses is that the statistical quality of the 50 reading comprehension items has never been examined. Unlike items produced by established testing agencies, these items were constructed by an undergraduate researcher as part of a research and development (R&D) thesis project, and although they were developed with reference to the EPT syllabus, they had not undergone any empirical evaluation prior to this study. Consequently, three basic qualities of the items remained unknown: their item facility (P), which shows whether an item is too easy, too difficult, or appropriately challenging; their item discrimination (D), which shows whether an item is able to distinguish between higher- and lower-proficiency test-takers; and their distractor efficiency (DE), which shows whether the incorrect answer options are functioning as intended by attracting responses from test-takers who have not mastered the tested concept (Arikunto, 2018; Rezigalla et al., 2024).

Previous research in the Indonesian EFL context has highlighted various challenges in locally developed English tests and reading comprehension materials (Azizah et al., 2021). Several studies have documented recurring issues such as unbalanced difficulty levels, poor discrimination power, and non-functional distractors in teacher-made tests across different educational levels (Santosa et al., 2025; Maharani & Putro, 2020; Sugianto, 2020). Other studies in university contexts have also examined reading comprehension challenges, cognitive level distributions, test coverage, and test-taking strategies (Anjani et al., 2025; Narayana & Soepriyanti, 2023; Sari et al., 2023). However, much of this existing research has focused on traditional paper-based tests, general reading difficulties, or learner strategies. None of these studies has empirically examined the statistical quality of reading comprehension items developed specifically for a digital EPT preparation platform at the university level (Hidayat et al., 2025). This is the gap that the present study addresses. The novelty of this study lies in applying classical test theory (CTT) item analysis to a digital EPT preparation platform and in integrating item facility, item discrimination, and distractor efficiency into a single decision-making matrix that yields item-level recommendations to retain, revise, or discard each item.

Building on this gap, and following the methodological approach demonstrated by Danuwijaya (2018), who showed that even items developed from established frameworks may perform differently once tested on a specific population, this study applied the item analysis procedures established by Arikunto (2018) to evaluate the 50 reading comprehension items embedded in the Digitalized English Proficiency Test Reading Preparation Materials. Specifically, this study aimed to (1) analyze the item facility index, (2) analyze the item discrimination index, and (3) evaluate the distractor efficiency of the 50 reading comprehension items, in order to provide the platform developer with an empirical basis for deciding which items should be retained, revised, or discarded before the platform is used more widely.

METHOD

This study employed a quantitative descriptive research design to evaluate the statistical quality of the 50 reading comprehension items embedded in the Digitalized English Proficiency Test Reading Preparation Materials. A descriptive design was appropriate because the study aimed to describe the items' current quality in terms of facility, discrimination, and distractor efficiency, rather than test a hypothesis or establish a cause-and-effect relationship (Creswell, 2014; Cohen et al., 2018). The quantitative approach was applied through the calculation of three indices, item facility (P), item discrimination (D), and distractor efficiency (DE), for each of the 50 items, based on student response data collected through the platform (Arikunto, 2018).

The participants in this study were students from the English Education Department at the University of Mataram who were members of SAE2P (Student Association of English Education Program) and actively engaged in EPT preparation activities. The criteria for selecting participants were: (1) active members of SAE2P, (2) had used the Digitalized English Proficiency Test Reading Preparation Materials platform, and (3) had completed all 50 reading comprehension items in full. Purposive sampling was used because the study required participants who had direct experience with the platform and whose responses were complete, so that the item statistics would reflect actual interaction with the developed instrument (Cohen et al., 2018). This group was selected because SAE2P membership includes students from multiple academic semesters, which produces a range of English proficiency levels among its members. This variation was methodologically important because an accurate calculation of the item discrimination index requires a sufficiently wide spread of ability levels, so that the upper 27% and lower 27% groups represent meaningfully different levels of proficiency (Arikunto, 2018). Following Brown's (2004) principle of norm-referenced evaluation, participants' total scores were ranked from highest to lowest, and the top 27% were designated as the upper group while the bottom 27% were designated as the lower group.

Data were collected from 38 participants on 24 July 2026. Following Arikunto's (2018) guidance that a minimum of 30 to 40 participants is sufficient for a reliable classical test theory item analysis, these 38 participants were divided into an upper group and a lower group of 10 participants each, with the remaining 18 participants forming the middle group.

This study used three complementary instruments. The first and primary instrument was the set of 50 reading comprehension multiple-choice items, each consisting of a written passage followed by a question with four options (one key and three distractors), covering sub-skills such as identifying main ideas, locating specific details, understanding vocabulary in context, and making inferences. The second instrument was a Digital Student Response Sheet (Matrix), an automatically generated spreadsheet that recorded each participant's selected option and binary score (1 for correct, 0 for incorrect) for all 50 items and allowed sorting into upper, middle, and lower groups. The third instrument was an Item Analysis Recapitulation Sheet, a structured worksheet used to tabulate the number of correct responses from the upper and lower groups, compute the facility and discrimination indices, and calculate the percentage of participants selecting each distractor option for the distractor efficiency analysis.

Data were collected by administering the 50-item reading comprehension test through the web-based platform. All participants completed the 50 items in a single session using their personal devices. The test link was distributed through a WhatsApp group coordinated with the platform developer. Supervision was conducted online through group coordination and platform log monitoring, and participants were asked to work independently without external assistance. The data collection took place on 24 July 2026 through the Digitalized EPT Reading Preparation Materials platform. After all participants had submitted their responses, the response data were extracted from the platform's database and organized into the Digital Student Response Sheet. The responses were then processed in two ways: for the item facility and discrimination analysis, each response was converted into a binary score (1 for correct, 0 for incorrect); for the distractor efficiency analysis, the specific option selected for each incorrect response was recorded and tabulated as a percentage of the total number of participants.

All calculations were carried out in Microsoft Excel following the procedures established by Arikunto (2018). The item facility index (P) was calculated by dividing the number of participants who answered an item correctly by the total number of participants. Each item was then classified as difficult, medium/acceptable, or easy, following the criteria in Table 1.

Table 1. Classification of Item Facility Index (P)

Index of Item Facility (P)	Classification
.00 - .30	Difficult
.31 - .70	Medium / Acceptable
.71 - 1.00	Easy

Source: Arikunto (2018)

The item discrimination index (D) was calculated by subtracting the proportion of correct responses in the lower group from the proportion of correct responses in the upper group ($D = BA/JA - BB/JB$, where BA and JA are the number of correct responses and the total number of students in the upper group, and BB and JB are the corresponding figures for the lower group). Each item was then classified as poor, satisfactory, good, or excellent; items with a negative D value were flagged as functioning in the wrong direction, following the criteria in Table 2.

Table 2. Classification of Item Discrimination Index (D)

Index of Item Discrimination (D)	Classification
.00 – .20	Poor
.21 – .40	Satisfactory
.41 – .70	Good
.71 – 1.00	Excellent
Negative value (--)	Very poor

Distractor efficiency (DE) was calculated for each of the three distractors in every item as the percentage of participants who selected that option, using a formula adapted from Berbiru et al. (2024). A distractor was classified as functional if it was selected by at least 5% of participants and non-functional if selected by fewer than 5% (Arikunto, 2018).

Table 3. Classification of Distractor Efficiency (DE)

Percentage of Selected Distractor (DE)	Classification
≥ 5%	Functional
< 5%	Non-Functional

At the item level, the number of non-functional distractors (NFDs) in each item was used to classify the item's overall distractor efficiency, following the criteria adapted from Rezigalla et al. (2024), as shown in Table 4.

Table 4. Classification of Item-Level Distractor Efficiency

Number of NFDs	Interpretation	Score
3	Poor	0%
2	Moderate	33,3%
1	Good	66,6%
0	Excellent	100%

Note: NFD = Non-Functional Distractor, adopted from Rezigalla et al. (2024)

Finally, the results of the three indices were combined into a single decision for each item, retain, revise, or discard, based on the integrated criteria synthesized from Arikunto (2018) and Rezigalla et al. (2024), presented in Table 5.

Table 5. Criteria of the Decision-Making Matrix for Test Items

Final Decision	Integrated Criteria (Facility + Discrimination + DE)	Action Required
RETAIN	All three parameters meet the ideal standards: a) facility is medium (0.31–0.70); b) discrimination is good/excellent (0.41–1.00); c) DE is excellent (0 NFD).	No action needed; the item is kept in the test bank.
REVISE	One or two parameters fall below the ideal standards but the item is not fundamentally flawed: a) facility is too easy/difficult; b) discrimination is poor/satisfactory (0.00–0.40); c) DE has 1 or 2 NFDs.	The item's stem or its non-functional distractor(s) must be revised.
DISCARD	At least one parameter indicates a fatal flaw: a) discrimination index is negative; b) DE is completely poor (3 NFDs).	The item is removed from the platform or rewritten entirely.

Note: synthesized from Arikunto (2018) and Rezigalla et al. (2024)

Content validity was established through a formal expert judgment process conducted as part of the broader R&D study that produced the platform. Three expert validators were involved: two lecturers in English language teaching and language assessment and one EPT coordinator. The qualifications required were a doctoral or master's degree in ELT or language assessment and at least five years of experience in language testing or EPT preparation. The validators reviewed the items against the EPT syllabus indicators for content appropriateness, cognitive depth, and structural accuracy. Because the 50 items were newly constructed and this study represents an initial evaluation within the "Develop" phase of the R&D framework, overall test reliability was not calculated at this stage. This decision is acknowledged as a limitation because reliability estimates are needed before the instrument is used for high-stakes or broad-scale decisions (Brown, 2004; Fulcher, 2010). Reliability will be examined in a subsequent, larger-scale field trial.

RESULT AND DISCUSSION

Item Facility of the Reading Comprehension Items

Of the 50 items analyzed, 4 items (8%) were classified as difficult, 25 items (50%) as medium/acceptable, and 21 items (42%) as easy, with a mean item facility index of 0.64 (Table 6). This mean value indicates that, overall, the test was moderately easy for the participants.

Table 6. Frequency Distribution of Item Facility Index (P)

Index of Item Facility (P)	Number of Items (f)	Percentage (%)
0.00 – 0.30 (Difficult)	4	8%
0.31 – 0.70 (Medium/ Acceptable)	25	50%
0.71 – 1.00 (Easy)	21	42%
Total	50	100%

Exactly half of the items fell within the medium range, which is the most appropriate range for EPT preparation because it provides the most informative data on differences in learners' reading proficiency (Arikunto, 2018). However, the fact that 42% of the items were classified as easy suggests that the instrument may not sufficiently challenge higher-proficiency test-takers. Azizah et al. (2023) reported a similar pattern of unbalanced item difficulty in a teacher-made English test, where a substantial proportion of items were either too easy or too difficult. This suggests that difficulty-related weaknesses are common in locally developed tests in this context, whether items are newly constructed, as in the present study, or adapted from existing sources. A related pattern also appears in Sari et al. (2023), who found that most reading comprehension questions in a vocational school textbook targeted lower-order rather than higher-order thinking skills; although that study examined cognitive level rather than statistical difficulty, both findings point to a tendency for locally produced reading materials to lean toward the easier end of the spectrum.

This tendency is worth attention because, as Little (2007) argues, autonomous learners depend on accurate feedback from their practice materials; if a platform's items are disproportionately easy, students preparing independently for the EPT may develop an overly positive impression of their reading ability, which could reduce the accuracy of the feedback the platform is meant to provide. At the same time, this

statistical ease should not be taken to mean the test is not challenging for every student. Narayana and Soepriyanti (2023) found that postgraduate students at the same university reported considerable difficulty with reading comprehension due to unfamiliar vocabulary and time pressure, showing that a statistically easy item is not always experienced as easy by every test-taker. Taken together, these findings suggest that the facility index alone does not fully capture all of the factors that influence how difficult a reading item feels to an individual test-taker.

Item Discrimination of the Reading Comprehension Items

Of the 50 items, 15 (30%) were classified as poor, 20 (40%) as satisfactory, 12 (24%) as good, and 2 (4%) as excellent, while 1 item (2%) produced a negative discrimination value, meaning that the lower group answered that item correctly more often than the upper group. The mean discrimination index across all 50 items was 0.34 (Table 7). When combined, only 14 items (28%) showed good or excellent discrimination, while 35 items (70%) were classified as poor or satisfactory.

Table 7. Frequency Distribution of Item Discrimination Index (D)

Index of Item Discrimination (D)	Number of Items (f)	Percentage (%)
.00 – .20 (Poor)	15	30%
.21 – .40 (Satisfactory)	20	40%
.41 – .70 (Good)	12	24%
.71 – 1.00 (Excellent)	2	4%
Negative value	1	2%
Total	50	100%

This pattern broadly aligns with Azizah et al. (2023), who also found that a considerable number of items in a teacher-made test had poor discrimination power (15 items, or 30%, along with one item with a negative value). The similarity between the two studies suggests that weak discrimination power may not depend solely on who develops the test items, but may instead be a common outcome of using newly written items that have not been piloted or refined beforehand.

An important explanation for this pattern lies in the nature of Classical Test Theory (CTT) itself, which is inherently sample-dependent. As Alderson (2000) explains, reading assessment is a highly interactive process in which test-taker variables, such as background knowledge, vocabulary mastery, and general English proficiency, strongly influence how test-takers respond to items. Within CTT, item statistics, especially the discrimination index, are therefore not fixed properties of the item itself but reflections of how the item performs within a particular group of test-takers. If the group taking the test is relatively homogeneous in ability, the upper and lower groups will not differ as sharply, and the test will mathematically struggle to produce a high discrimination index, even for a reasonably well-constructed item. Because the participants in this study were members of SAE2P, a student association whose members share a broadly similar level of exposure to EPT preparation, the restricted variability (homogeneity) in English proficiency among these pilot subjects offers a plausible explanation for the relatively high number of items with weak discrimination power, and, by extension, for why weak discrimination was the most frequently cited reason behind the 82% of items recommended for revision. This interpretation is supported by Brown (2004) and Hughes (2003), who both stress that

a test's overall validity and reliability depend heavily on the extent to which its items can distinguish between test-takers of differing ability, an ability that is itself contingent on the range of proficiency present in the sample used to pilot the instrument. Viewed this way, the weak discrimination power identified in this study does not necessarily indicate that the items were poorly constructed; rather, it demonstrates that the instrument's statistical performance is closely linked to the specific characteristics, in this case the relative homogeneity, of the initial pilot sample. A broader field trial involving a more heterogeneous population would be needed to determine whether these discrimination values change under more varied conditions.

Distractor Efficiency of the Reading Comprehension Items

Of the 150 distractor options examined (three per item across 50 items), 109 (72.7%) were classified as functional, meaning that each was selected by at least 5% of participants, while the remaining 41 (27.3%) were classified as non-functional (Table 8). At the item level, the mean number of non-functional distractors per item was 0.82, with a standard deviation of 0.82, a minimum of 0, and a maximum of 3.

Table 8. Frequency Distribution of Distractor Functionality (Option Level)

Classification of Distractors (DE)	Number of Distractors (f)	Percentage (%)
Functional ($\geq 5\%$)	109	72.7%
Non-Functional ($< 5\%$)	41	27.3%
Total	150	100%

At the item level, 20 items (40%) contained no non-functional distractors and were classified as excellent, 21 items (42%) contained one non-functional distractor and were classified as good, 7 items (14%) contained two non-functional distractors and were classified as moderate, and 2 items (4%) contained three non-functional distractors and were classified as poor (Table 9). Combined, 41 items (82%) had at most one non-functional distractor; that is, they were classified as either good or excellent.

Table 9. Item-Level Distractor Efficiency Classification

Number of NFDs	Interpretation	Number of Items (f)	Percentage (%)
0	Excellent	20	40%
1	Good	21	42%
2	Moderate	7	14%
3	Poor	2	4%
Total		50	100%

These results indicate that, in general, the distractor options in the 50 items were reasonably well constructed, since almost three-quarters of all distractors succeeded in attracting responses from lower-proficiency test-takers, and non-functional distractors were concentrated in a relatively small number of items rather than spread evenly across the whole item set. This finding relates to Azizah et al. (2023), who also identified non-functional distractors as a major weakness in a teacher-made test, confirming that distractor quality is a common challenge in test construction generally, not a problem unique to this platform. At the same time, the fact that 41 of the 150 distractors analyzed here (27.3%) were still non-functional shows that although the item writer worked without prior piloting, most, though not all, of

the plausible misunderstandings that lower-proficiency readers might have were anticipated, which is a reasonably strong result for a first version of the instrument.

Overall Item Quality and Decision-Making

When the three indices were combined using the integrated decision-making criteria in Table 5, 6 items (12%), specifically items 5, 15, 30, 33, 35, and 44, met the ideal standard on all three indices and were recommended for retention without modification. Forty-one items (82%) fell short on one or two of the indices, most often due to weak discrimination power or the presence of one non-functional distractor, and were recommended for revision. The remaining 3 items (6%) were recommended for removal (Table 10).

Table 10. Recapitulation of Final Decision-Making for Test Items

Final Decision	Number of Items (f)	Percentage (%)
RETAIN	6	12%
REVISE	41	82%
DISCARD	3	6%
Total	50	100%

Among the items recommended for removal, items 2 and 10 shared a similar profile: they were answered correctly by nearly all participants, and none of their distractors attracted any responses, meaning these two items provided almost no information about differences in reading ability among the participants. Item 12 presented a different problem: although all three of its distractors functioned well, its discrimination index was negative, meaning that more participants in the lower group answered it correctly than participants in the upper group. Because the distractors functioned as intended, this reversed pattern more likely stems from the item stem or the answer key; for example, ambiguous wording that leads higher-proficiency readers to overanalyze an otherwise simple question, or an answer key that identifies an option as correct when it is not the most appropriate choice. Item 12 was therefore flagged for a review of its stem and its answer key rather than its distractors.

Taken as a whole, the item quality profile indicates that the item bank is not yet a fully polished instrument because only a small proportion of items met all three ideal standards at once. However, the fact that only 6% of items showed a fatal flaw, compared with 82% requiring comparatively minor revision, suggests that the current version of the instrument represents a solid starting point rather than a fundamentally flawed one. Most of the necessary improvements are targeted and manageable rather than requiring items to be rewritten from scratch. These findings also address a gap identified in earlier University of Mataram-based studies (Anjani et al., 2025; Azizah et al., 2023; Narayana & Soepriyanti, 2023; Sari et al., 2023), none of which examined the statistical quality of reading items developed specifically for a digital EPT preparation platform at the university level. By applying classical test theory item analysis to this new context, the present study extends this line of research and provides the platform developer with concrete, item-by-item evidence to guide the next phase of development.

CONCLUSION

This study analyzed the statistical quality of 50 reading comprehension items embedded in the Digitalized English Proficiency Test Reading Preparation Materials at the University of Mataram, using classical test theory procedures to examine item facility, item discrimination, and distractor efficiency. Regarding item facility, half of the items (50%) fell within the medium/acceptable range, with a mean facility index of 0.64, indicating that the instrument was moderately easy overall. Regarding item discrimination, only 14 items (28%) achieved good or excellent discrimination power, while 15 items (30%) were classified as poor and 1 item (2%) produced a negative value; the mean discrimination index was 0.34. Regarding distractor efficiency, 109 of 150 distractors (72.7%) were functional, and 41 items (82%) had at most one non-functional distractor.

Taken together, these findings suggest that the instrument shows solid initial construction in terms of item facility and distractor efficiency but weaker performance in item discrimination. This weakness, together with the fact that 82% of items required some form of revision, appears to be strongly influenced by external, sample-related factors rather than by fundamental flaws in the items themselves. Because the pilot participants were relatively homogeneous in English proficiency, the classical test theory statistics used in this study, particularly the discrimination index, were mathematically limited in their ability to differentiate between higher- and lower-proficiency test-takers. Based on the combined criteria, 6 items (12%) are recommended for retention, 41 items (82%) for revision, and 3 items (6%) for removal. Although a large proportion of items require revision, these results should be viewed as actionable, item-level guidance for targeted improvement rather than evidence of a fundamentally flawed instrument.

The theoretical implication of this study is that item statistics in classical test theory should be interpreted as sample-dependent rather than as fixed properties of items. The practical implication is that the platform developer can use the item-level decision matrix to prioritize revisions, especially for items with weak discrimination and non-functional distractors. The methodological implication is that future evaluations should include reliability estimation and a broader, more heterogeneous sample to obtain more stable item statistics. This evaluation represents an important step within the “Develop” phase of the platform’s research and development process, providing the empirical basis needed to move the Digitalized English Proficiency Test Reading Preparation Materials toward a more valid, reliable, and standardized reading assessment tool for English Education students preparing for the EPT.

RECOMMENDATIONS

Based on the findings, several recommendations are offered: (1) the platform developer should use the item quality profile to refine the item bank, prioritizing the removal or rewriting of items 2, 10, and 12, followed by targeted revision of the 15 items with poor discrimination power and the items with two or more non-functional distractors; (2) lecturers and EPT coordinators can use this profile to identify reading sub-skills that need targeted instruction and should adopt periodic item analysis as a routine procedure; (3) students should not rely too heavily on high scores on the platform, given the instrument’s relative ease, and should continue practicing more

demanding skills such as making inferences and understanding academic vocabulary in context; and (4) future researchers are suggested to extend this framework to other EPT sections, conduct broader field trials with heterogeneous populations, calculate reliability estimates, and integrate Item Response Theory (IRT) for a more comprehensive analysis.

REFERENCES

- Alderson, J. C. (2000). *Assessing reading* (Vol. 10). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511732935>
- Anjani, T. K., Farmasari, S., & Amin, M. (2025). Exploring Toefl Taking Strategies of High Achiever Students Of English Education Department: Case Study at University of Mataram. *Pendas: Jurnal Ilmiah Pendidikan Dasar*, 10(04), 111-120. <https://doi.org/10.23969/jp.v10i04.35816>
- Arikunto, S. (2018). *Dasar-dasar evaluasi pendidikan* (R. Damayanti (ed.). PT. Bumi Aksara.
- Santosa, Y. T., Maulida, D. W., Ferdianto, J., Sutarni, S., & Hidayati, Y. M. (2025). Evaluating the Quality of Mid-Semester Mathematics Summative Assessment in Secondary School: A Psychometric Analysis of Test Items. *IJECA (International Journal of Education & Curriculum Application)*, 8(3), 431-449. <https://doi.org/10.31764/ijeca.v8i3.36276>
- Azizah, N., Inderawati, R., & Vianty, M. (2021). Developing descriptive reading materials in EFL classes by utilizing the local culture. *Studies in English Language and Education*, 8(2), 596-621. <https://doi.org/10.24815/siele.v8i2.18562>
- Berbiru, S., Tomaso, T. F., & Ubra, F. W. (2024). Analisis Kualitas Butir Soal Tes Multiple Choice Mata Pelajaran Ekonomi pada Siswa Kelas XI IPS SMA Kristen Aprian Damer. *Jurnal Pendidikan Ekonomi*, 4(2), 78-86. <https://doi.org/10.30598/jpe.v4.i2.p78-86>
- Brown, H. D. (2004). *Language Assessment: Principles and Classroom Practices*. Pearson/Longman.
- Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). Routledge.
- Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (4th ed.). SAGE Publications.
- Danuwijaya, A. A. (2018). Item Analysis of Reading Comprehension Test for Post-Graduate Students. *English Review: Journal of English Education*, 7(1), 29-40. <https://doi.org/10.25134/ERJEE.V7I1.1493>
- Fulcher, G. (2010). *Practical language testing*. Hodder Education.
- Hidayat, H., Waloyo, E., & Priajana, N. (2025). Online Reciprocal Teaching and Its Impact on Islamic University Students' Reading Comprehension in Standardized Testing. *Register Journal*, 18(1), 153-172. <https://doi.org/10.18326/register.v18i1.153-172>
- Hughes, A. (2003). *Testing for Language Teachers* (2nd ed.). Cambridge University Press.
- Little, D. (2007). Language Learner Autonomy: Some Fundamental Considerations Revisited. *Innovation in Language Learning and Teaching*, 1(1), 14-29. <https://doi.org/10.2167/ILLT040.0>

- Maharani, A. V., & Putro, N. H. P. S. (2020). Item Analysis of English Final Semester Test. *Indonesian Journal of EFL and Linguistics*, 5(2), 491. <https://doi.org/10.21462/IJEFL.V5I2.302>
- Narayana, I. G. P. P., & Soepriyanti, H. (2023). Students' Difficulties in Coping with TOEFL Test: A Case Study in a Postgraduate English Program of a State University in Mataram. *International Journal of Multicultural and Multireligious Understanding*, 10(6), 321–325. <https://doi.org/10.18415/IJMMU.V10I6.4780>
- Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaifi, J., ALGhamdi, M. A., Al Ameer, A. Y., Yahia, A. I. O., Mohammed, O. A., & Adam, M. I. E. (2024). Item Analysis: The Impact of Distractor Efficiency on the Difficulty Index and Discrimination Power of Multiple-Choice Items. *BMC Medical Education*, 24(1). <https://doi.org/10.1186/s12909-024-05433-y>
- Sari, I. K., & Lestari, Y. B. (2023, December). The Analysis of Reading Comprehension Question Levels in English Textbook for Vocational High School Based on Revised Bloom's Taxonomy. In *Journal of English Education Forum (JEEF)* (Vol. 3, No. 2, pp. 25-29). <https://doi.org/10.29303/jeef.v3i2.566>
- Sugianto, A. (2020). Item analysis of English summative test: EFL teacher-made test. *Indonesian EFL Research and Practices*, 1(1), 35-54.
- Warschauer, M., & Ware, P. (2006). Automated Writing Evaluation: Defining the Classroom Research Agenda. *Language Teaching Research*, 10(2), 157–180. <https://doi.org/10.1191/1362168806lr190oa>