Check for updates



Journal of Education and Learning Research
Journal homepage: https://jelr.greadc.org/index.php/jelr

E-ISSN
3024-9589

Analysis of the Quality of Grade X Indonesian Language Final Semester Assessment Items Based on Classical Test Theory

Ni Wayan Suryanti*, Ida Bagus Putrayasa, I Gede Nurjaya

Indonesian Language and Literature Education Study Program, Faculty of Languages and Arts, Universitas Pendidikan Ganesha, Singaraja, 81116, Indonesia

ARTICLE INFO

Article history:

Received: 06 July 2026
Accepted: 15 Oct 2026
Published online: 10 Nov 2026

Keywords:

Classical Test Theory,
Difficulty Level,
Discrimination Power,
Reliability,
Validity

* *Corresponding author.*

E-mail:*suryanti@student.undiksha.ac.id
ib.putrayasa@undiksha.ac.id
gede.nurjaya@undiksha.ac.id

Article Doi

Doi: https://doi.org/10.62208/jelr.4.2.p.1-8

This is an open access article under the CC BY-SA license.

A B S T R A C T

The quality of evaluation instruments is essential for producing accurate information about student learning outcomes, yet test items are often not empirically analyzed. This study aimed to analyze the quality of Grade X Indonesian Language Final Semester Assessment items at SMK Negeri 1 Bangli based on Classical Test Theory. A quantitative descriptive evaluative approach was employed involving 148 Grade X students and 60 multiple-choice items. Data were collected through documentation of test scripts, answer keys, and student response sheets, and analyzed using Anates version 4.0.9. Results showed that 22 items (36.67%) had sufficient validity, 29 (48.33%) low validity, and 9 (15.00%) very low validity, while test reliability was very high (0.87). Regarding difficulty, 34 items (56.67%) were easy, 20 (33.33%) moderate, and 6 (10.00%) difficult. Discrimination power was good or very good for 20 items (33.33%), sufficient for 23 (38.33%), and poor or not good for 17 (28.33%), while 88 distractors (36.67%) were effective and 152 (63.33%) ineffective. Based on the integrated indicators, 22 items were retained, 29 revised, and 9 replaced. Items with low validity, weak discrimination, and ineffective distractors need to be revised or eliminated to improve the quality of the assessment instrument.

1. Introduction

Learning evaluation is an essential component of the educational process because it determines the extent to which students have achieved the expected competencies and assesses the effectiveness of the learning process. Evaluation results are often used as a basis for decisions concerning learning mastery, instructional improvement, and the development of follow-up programs. Therefore, evaluation instruments must be of high quality to produce accurate and reliable information (Riinwati, 2021). Studies published in the Journal of Education and Learning Research also report that achievement or learning outcome tests were checked by validators or subject specialists before they were used to judge

student learning (Bella & Azhar, 2023; Olude et al., 2025). One way to determine the quality of an evaluation instrument is through item analysis. Item analysis aims to identify the quality of each test item according to specific indicators so that its suitability as a measure of learning outcomes can be established. The analysis includes validity, reliability, difficulty level, discrimination power, and distractor effectiveness (Ramadhan et al., 2024). Through this process, weaknesses in test items can be identified, allowing evaluation instruments to be improved and used more effectively (Qadir et al., 2024).

This study employs the Classical Test Theory (CTT) approach, which views an observed test score as a combination of a true score and measurement error. Classical Test Theory is widely used in educational evaluation because it provides information about instrument quality through analyses of validity, reliability, difficulty level, discrimination power, and distractor effectiveness (Novriyanti et al., 2024). In Indonesian language learning at vocational high schools, the quality of evaluation instruments is particularly important because the subject emphasizes not only mastery of content but also students' literacy and communication skills. In line with the implementation of the Merdeka Curriculum, students are expected to develop integrated reading, writing, speaking, and listening skills (Simbolon, 2023). Therefore, the Final Semester Assessment should be constructed using high-quality instruments capable of measuring student competencies objectively and accurately, and the quality of its items needs to be verified with empirical data rather than assumed.

Several previous studies have analyzed test-item quality using indicators of validity, reliability, difficulty level, discrimination power, and distractor effectiveness. Ashayidatina and Alber (2023) examined Grade X Indonesian language items at SMK Negeri 1 Mempura, Salsabilla et al. (2025) analyzed Indonesian language items for Grade XI accounting students at SMK Negeri 2 Padang, Putri and Retnosari (2024) investigated the quality of end-of-semester items in accounting subjects at SMK Ketintang Surabaya, and Novriyanti et al., (2024) analyzed a general biology midterm test using Classical Test Theory. These studies revealed items with low validity, weak discrimination, disproportionate difficulty levels, and suboptimal distractors. However, they leave a research gap. Among these studies, the two on Indonesian language items were conducted in other vocational schools and, in one case, at a different grade level, while the others concerned accounting and biology tests. Because Classical Test Theory indices depend on the group of examinees, evidence from other schools cannot simply be transferred to the Grade X Indonesian Language Final Semester Assessment at SMK Negeri 1 Bangli. Nevertheless, no previous study has specifically investigated the quality of Grade X Indonesian Language Final Semester Assessment items at SMK Negeri 1 Bangli. Furthermore, empirical information on the quality of the evaluation instrument used at this school has not been available as a basis for item improvement.

This study offers novelty in three respects. Contextually, it provides empirical evidence on the item quality of the Grade X Indonesian Language Final Semester Assessment at SMK Negeri 1 Bangli, drawn from the actual responses of 148 students. Methodologically, it integrates five Classical Test Theory indicators, namely validity, reliability, difficulty level, discrimination power, and distractor effectiveness, and cross-tabulates the item-level results so that each weak item can be traced across indicators instead of being judged by a single coefficient. Practically, the resulting item profile is used to decide which items should be retained, revised, or eliminated and to formulate recommendations for constructing the assessment in the following period. Based on this gap and novelty, this study aims to analyze the quality of Grade X Indonesian Language Final Semester Assessment items at SMK Negeri 1 Bangli using Classical Test Theory indicators comprising validity, reliability, difficulty level, discrimination power, and distractor effectiveness, in order to provide an empirical basis for evaluating and improving assessment instruments used in Indonesian language learning.

2. Methodology

This study employed a quantitative descriptive design with an evaluative approach. It was conducted at SMK Negeri 1 Bangli during the 2025/2026 academic year. The research participants were 148 Grade X students who took the Indonesian Language Final Semester Assessment. The research objects consisted of 60 multiple-choice items from the Grade X Indonesian Language assessment. Research data were collected using documentation techniques. The documents analyzed included the Final Semester Assessment test script, answer key, and student answer sheets. All data were obtained from the Indonesian language teacher at SMK Negeri 1 Bangli. Test-response data were analyzed using Anates version 4.0.9 within the Classical Test Theory framework. The analysis covered five indicators of item quality: validity, reliability, difficulty level, discrimination power, and distractor effectiveness (Qadir et al., 2024; Salsabilla

et al., 2025). The results were interpreted according to the criteria for each indicator to determine item quality and classify items as suitable for use, requiring revision, or unsuitable for use.

The evaluative orientation of the study was selected because the analysis was intended not merely to describe student scores but to judge the functioning of the assessment instrument on the basis of explicit psychometric indicators. The unit of analysis was the individual multiple-choice item, while test-level reliability was also examined to determine the consistency of the complete instrument. The use of 148 student response sets made it possible to calculate empirical indices from the actual administration of the assessment. Because the study used existing assessment documents and student responses, the documentation technique ensured that the analysis reflected authentic school assessment practice rather than responses generated in an artificial testing situation.

The analytical procedure was conducted sequentially. First, the test script, answer key, and student answer sheets were compiled and checked for completeness before the response data were entered into Anates version 4.0.9. Second, item validity statistics were examined and interpreted according to the categories used in the study: very high, high, sufficient, low, and very low. Third, the overall reliability coefficient was interpreted to determine the degree of score consistency. Fourth, item difficulty was classified into very easy, easy, moderate, difficult, and very difficult categories (Qadir et al., 2024). Fifth, discrimination power was categorized as very good, good, sufficient, poor, or not good. Finally, each incorrect option was examined through distractor-effectiveness output and classified as effective when it was chosen by at least 5% of the students and as ineffective when it was chosen by fewer (Qadir et al., 2024). These stages were used to build an integrated profile of item quality rather than to rely on a single coefficient.

Interpretation emphasized the combined meaning of the indicators. An item was not considered satisfactory merely because it met one criterion. For example, an easy item could still contribute to an assessment when it showed acceptable validity and discrimination, whereas an item with weak validity and negative or very poor discrimination would require closer review even if its difficulty appeared acceptable. Similarly, very high test reliability was interpreted at the instrument level and was not treated as proof that every item was individually sound. The final evaluative judgments therefore considered the pattern of evidence across validity, reliability, difficulty, discrimination, and distractor functioning. This approach is consistent with the practical purpose of Classical Test Theory analysis in supporting item retention, revision, and removal decisions.

3. Results and Discussion

Implementation of the study. The study was implemented by analyzing the Final Semester Assessment that had been administered to Grade X students of SMK Negeri 1 Bangli in the 2025/2026 academic year. The assessment consisted of 60 multiple-choice items for the Indonesian Language subject, and each item had one answer key and four distractors, which produced the 240 distractors examined in this study. A total of 148 Grade X students took the test as part of the regular end-of-semester assessment, so the data reflect authentic school examination conditions rather than an artificial testing situation. The researchers obtained the test script, the answer key, and the student answer sheets from the Indonesian language teacher, checked the completeness of the 148 response sets, and entered the responses into Anates version 4.0.9. The program produced the item validity coefficients, the test reliability coefficient, the difficulty and discrimination indices, and the distractor-effectiveness output, which were interpreted according to the criteria of each indicator and then compared across indicators to profile every item. The analysis of 60 Final Semester Assessment items using Anates version 4.0.9 produced findings for validity, reliability, difficulty level, discrimination power, and distractor effectiveness. The results are presented together with their interpretation and discussion.

As shown in Table 1, validity analysis showed that no items were categorized as very high or high, which means that none of the items had a strong relationship with the total score. Twenty-two items (36.67%) were sufficient, 29 items (48.33%) were low, and 9 items (15.00%) were very low. Taken together, 38 items (63.33%) had low or very low validity. The very low category included items 1, 14, 17, 37, 39, 40, 42, 44, and 52, which hardly measured the same competency as the test as a whole and were therefore the first candidates for review. Thus, most items had low validity. The reliability of the test is summarized in Table 2. Reliability analysis yielded a coefficient of 0.87, with an XY correlation of 0.77, a mean score of 40.53, and a standard deviation of 7.87. The coefficient indicates very high reliability.

Table 1. Validity Analysis Results

Validity Criterion	Range	Number of Items	Total	%
Very High	0.91–1.00	–	0	0%
High	0.71–0.90	–	0	0%
Sufficient	0.41–0.70	3, 6, 8, 11, 13, 18, 19, 21, 25, 28, 31, 33, 36, 46, 47, 48, 49, 50, 51, 53, 57, 59	22	36.67%
Low	0.21–0.40	2, 4, 5, 7, 9, 10, 12, 15, 16, 20, 22, 23, 24, 26, 27, 29, 30, 32, 34, 35, 38, 41, 43, 45, 54, 55, 56, 58, 60	29	48.33%
Very Low	Negative–0.20	1, 14, 17, 37, 39, 40, 42, 44, 52	9	15.00%
Total			60	100%

The XY correlation expresses the relationship between the two halves of the test, and the reliability coefficient is the estimate for the whole test. In practical terms, the total score of the instrument was consistent for the group analyzed, although this coefficient says nothing about the quality of individual items, which is examined through the other indicators.

Table 2. Reliability Analysis Results

Indicator	Value
Mean	40.53
Standard Deviation	7.87
XY Correlation	0.77
Test Reliability	0.87

The distribution of item difficulty is presented in Table 3. Difficulty analysis showed that 34 items (56.67%) were easy, 20 items (33.33%) moderate, and 6 items (10.00%) difficult. No items were categorized as very easy or very difficult. Most items were therefore answered correctly by a large share of the students. All of the first eleven items were easy, whereas 12 of the 20 moderate items were located between items 46 and 59, and the six difficult items were items 17, 37, 39, 40, 42, and 44.

Table 4 shows the results of the discrimination power analysis. Discrimination analysis showed 2 items (3.33%) in the very good category, 18 (30.00%) good, 23 (38.33%) sufficient, 11 (18.33%) poor, and 6 (10.00%) not good. Most items were in the sufficient category.

Table 3. Difficulty Level Analysis Results

Difficulty Criterion	Number of Items	Total	Percentage
Very Easy	–	0	0%
Easy	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 15, 16, 20, 21, 23, 24, 25, 26, 27, 28, 29, 31, 33, 34, 35, 36, 38, 41, 43, 45, 50, 58, 60	34	56.67%
Moderate	12, 13, 14, 18, 19, 22, 30, 32, 46, 47, 48, 49, 51, 52, 53, 54, 55, 56, 57, 59	20	33.33%
Difficult	17, 37, 39, 40, 42, 44	6	10.00%
Very Difficult	–	0	0%
Total	60	60	100%

In total, 43 items (71.67%) had sufficient to very good discrimination, whereas 17 items (28.33%) were poor or not good and therefore did not separate higher-ability from lower-ability students adequately.

Table 4. Discrimination Power Analysis Results

Criterion	Number of Items	Total	Percentage
Very Good	48, 53	2	3.33%
Good	12, 13, 18, 19, 21, 22, 25, 28, 30, 31, 46, 47, 49, 50, 51, 55, 57, 59	18	30.00%
Sufficient	2, 6, 7, 9, 11, 15, 16, 20, 24, 26, 29, 32, 33, 34, 35, 36, 38, 41, 45, 52, 54, 56, 58	23	38.33%
Poor	1, 3, 4, 5, 8, 10, 14, 23, 27, 43, 60	11	18.33%
Not Good	17, 37, 39, 40, 42, 44	6	10.00%
Total	60	60	100%

The effectiveness of the distractors is presented in Table 5. Of 240 distractors analyzed, 88 (36.67%) functioned effectively because they were chosen by at least 5% of the 148 students, while 152 (63.33%) were ineffective because they were chosen by fewer than 5% of the students. Most distractors therefore did not work as intended, and on average fewer than two of the four distractors in each item attracted students. This predominance of nonfunctioning distractors is in line with the predominance of easy items in Table 3.

Table 5. Distractor Effectiveness Analysis Results

Distractor Criterion	Total	Percentage
Effective (chosen by ≥ 5%)	88	36.67%
Ineffective (chosen by < 5%)	152	63.33%
Total	240	100%

Recapitulation of the Final Decisions on Test Items

The results of the five indicators were integrated to determine the follow-up decision for each item, as summarized in Table 6. Of the 60 items, 22 items (36.67%) were retained because all indicators met the criteria; 29 items (48.33%) were revised because one or more indicators had not yet met the criteria, such as nonfunctioning distractors or a difficulty level that needed adjustment; and 9 items (15.00%) were replaced or not reused because their validity was very low and their discrimination power was not good or negative, so the items could not function optimally in measuring student ability.

Table 6. Recapitulation of Final Item Decisions

Decision	Item Numbers	Total	%
Retained	3, 6, 8, 11, 13, 18, 19, 21, 25, 28, 31, 33, 36, 46, 47, 48, 49, 50, 51, 53, 57, 59	22	36.67%
Revised	2, 4, 5, 7, 9, 10, 12, 15, 16, 20, 22, 23, 24, 26, 27, 29, 30, 32, 34, 35, 38, 41, 43, 45, 54, 55, 56, 58, 60	29	48.33%
Replaced or not reused	1, 14, 17, 37, 39, 40, 42, 44, 52	9	15.00%
Total		60	100%

The findings show that most items had low or very low validity, indicating that many items were not yet able to measure student competencies accurately. Validity reflects an item's capacity to measure the competency it is intended to assess; therefore, items with low validity require revision in terms of content, construction, and language (Hamimi et al., 2020). This finding is consistent with Ketaren et al., (2024), who also reported that most items did not meet validity criteria and required revision before reuse. In contrast, Salsabilla et al., (2025) reported that 70% (Package A) and 80% (Package B) of the items they analyzed were valid, although they classified items only as valid or invalid by comparing the calculated and tabled correlation coefficients. The proportion of items with at least sufficient validity in the present study (36.67%) is therefore considerably lower. The distribution of validity across the test explains part of this pattern. Nine of the 22 items with sufficient validity (items 46, 47, 48, 49, 50, 51, 53, 57, and 59) were located in the last part of the test, whereas all six difficult items (items 17, 37, 39, 40, 42, and 44) fell into the very low validity category together with items 1, 14, and 52. An item that is answered correctly by few students, including students with high total scores, hardly follows the total score, so its validity coefficient falls toward zero or becomes negative. The very low validity of these items is therefore more likely to stem from construction problems, such as a questionable answer key, ambiguous wording, or content that had not been taught, than from genuine difficulty, and their keys and wording should be checked first.

The reliability coefficient of 0.87 indicates that the instrument had very high consistency and could produce relatively stable results under comparable conditions. Reliability refers to the consistency of an instrument in generating measurement scores (Yusuf, 2024). This result is consistent with Ketaren et al., (2024), who obtained a reliability coefficient of 0.98, and with Yusuf (2024), who likewise reported high reliability. It is also higher than the coefficient of 0.80 reported by Olude et al. (2025) for a 20-item multiple-choice achievement test, although that coefficient was estimated with a test-retest procedure and the two tests differ in length. It is also higher than the coefficients reported by Salsabilla et al. (2025) for their two test packages (0.345 and 0.743), the first of which was classified as low. Nevertheless, high reliability does not necessarily indicate that every individual item is of good quality. This is evident in the present study because, despite very high overall test reliability, 38 items had low or very low validity and therefore still require revision before being reused. Reliability can be influenced by test length and by the covariance among items, so a test of 60 items may obtain a high coefficient even when several items need improvement. The school should therefore preserve the consistency of the instrument while improving the weaker items, and the test should be re-analyzed after revision because removing or rewriting items can change the reliability coefficient.

The difficulty analysis showed that most items were categorized as easy. This indicates that the majority of students were able to answer the items correctly, resulting in a generally low level of difficulty. Difficulty level is an important indicator of instrument quality because, together with discrimination power, it helps determine how accurately an item measures student ability (Qadir et al., 2024). This finding is consistent with Salsabilla et al. (2025)for one of the two test packages they analyzed (Package B, 60% easy items), whereas the other package was dominated by moderate items. In contrast, Ashayidatina and Alber (2023) found a predominance of difficult items, while

Novriyanti et al. 2024) reported mostly moderate to difficult items. Such differences may be influenced by student characteristics, content coverage, and the complexity of the items constructed. The cross-tabulation of the present results shows why difficulty matters for item quality: 15 of the 20 moderate items (75.00%) had good or very good discrimination, whereas only 5 of the 34 easy items (14.71%) did, and 10 of the 11 items with poor discrimination were easy items. Because all of the first eleven items were easy, the opening part of the test added little to the separation of students with different ability levels. Easy items are not inherently defective, but when they dominate a test the scores provide limited information about differences among students at moderate and high achievement levels (Qadir et al., 2024).

The findings indicate that most items had sufficient discrimination power, although some were categorized as poor or not good. This suggests that several items could distinguish between higher- and lower-ability students, but this function was not optimal across all items. Hamimi et al. (2020) state that good discrimination power reflects an item's ability to differentiate students who have mastered the material from those who have not. The present results are better than those reported by Novriyanti et al. (2024) and Putri and Retnosari (2024), whose studies were dominated by items with low discrimination power. However, they are less favorable than the findings of Salsabilla et al. (2025), who reported that most items had good to very good discrimination. The two very good items (items 48 and 53) were both moderate in difficulty and had sufficient validity, which illustrates the profile of a well-functioning item. Weak discrimination, in contrast, had two different sources. Ten of the 11 poor items were easy, so almost all students answered them correctly regardless of ability, whereas the six not-good items (items 17, 37, 39, 40, 42, and 44) were the difficult items with very low validity, which points to a problem with the answer key, with more than one defensible option, or with content that students had not been taught. Items with weak discrimination should therefore be checked against the answer key and the test blueprint and then revised to improve their ability to differentiate student performance (Qadir et al., 2024).

The results show that 88 distractors (36.67%) functioned effectively, whereas 152 distractors (63.33%) did not function optimally because each was chosen by fewer than 5% of the 148 students. Effective distractors attract students who have not mastered the material toward plausible incorrect alternatives and can therefore improve item quality. Distractors are response alternatives presented in a test that are not the correct answer key (Qadir et al., 2024). A nonfunctioning distractor may be implausible, grammatically inconsistent with the stem, or obviously different in length or style, so students can eliminate it without demonstrating the intended knowledge or skill. Qadir et al. (2024) regard a distractor as effective when it is chosen by at least 5% of test takers, whereas a distractor that no student selects cannot perform its function. The present result is below that of Salsabilla et al. (2025), who reported distractor effectiveness of 90–100%, and similar to Novriyanti et al. (2024), who found that most distractors did not function optimally. The predominance of nonfunctioning distractors is in line with the predominance of easy items, because when most students can identify the key, few of them choose the incorrect options. Distractors should be made as similar to the key as possible so that they tempt students who have not mastered the material, and distractors that prove ineffective should be replaced (Qadir et al., 2024). In Indonesian language assessment, plausible distractors can be developed from alternative interpretations of a text, common linguistic errors, or incomplete applications of relevant concepts, provided that only one answer remains clearly best. Nonfunctioning alternatives should therefore be revised to improve the overall quality of the assessment instrument.

The recapitulation in Table 6 shows how combining the five indicators produces a more useful picture than any single coefficient. Twenty-two items (36.67%; items 3, 6, 8, 11, 13, 18, 19, 21, 25, 28, 31, 33, 36, 46, 47, 48, 49, 50, 51, 53, 57, and 59) were retained because all indicators were judged to meet the criteria and can be kept in the item bank. Twenty-nine items (48.33%) were revised because one or more indicators had not yet met the criteria, such as nonfunctioning distractors or a difficulty level that needs adjustment, and these items can be improved by revising the stem or the distractors rather than replacing the whole item. At the other extreme, 9 items (15.00%; items 1, 14, 17, 37, 39, 40, 42, 44, and 52) combined very low validity with discrimination that was not good or negative, so they could not function optimally in measuring student ability and should be replaced or not reused. This decision logic is more useful for teachers than a simple pass-fail classification because it directs revision effort first to the items whose weaknesses on several indicators raise the strongest concerns about measurement quality.

The findings have several implications for constructing the Final Semester Assessment for Grade X Indonesian Language in the following period at SMK Negeri 1 Bangli. First, the 22 retained items can form the nucleus of an item bank in which the

empirical indices are recorded after each administration, the 29 items identified for revision should be improved, and the 9 items identified for replacement should not be reused before they are rewritten. Second, because easy items dominate the test, the next test blueprint should plan a more balanced distribution of difficulty with more moderate items, since moderate items produced the best discrimination in this study; difficult items should be kept only after their keys and wording have been verified. Third, distractors should be rewritten from realistic student errors and misconceptions so that more of the four alternatives function. Fourth, statistical analysis should be integrated with qualitative review by subject teachers so that psychometric evidence is considered alongside curriculum alignment and language appropriateness, and the revised test should be re-analyzed after it has been administered. These steps can strengthen the evidential basis of semester assessment and reduce dependence on untested items (Qadir et al., 2024; Salsabilla et al., 2025).

This study also has limitations that should inform interpretation and future research. Classical Test Theory statistics are influenced by the characteristics of the examinee group, meaning that the indices observed for these 148 Grade X students may differ if the same items are administered to another population. The study also focuses on one school, one subject, one grade level, and one assessment administration. In addition, the quantitative analysis identifies patterns of item functioning but does not independently determine the specific linguistic or content reason for every weak item. Future research could therefore combine Classical Test Theory with expert judgment, cognitive interviews, or modern measurement approaches such as Rasch analysis. Replication across cohorts would also help determine whether problematic item characteristics persist over time.

4. Conclusion

This study achieved its aim of describing the quality of the Grade X Indonesian Language Final Semester Assessment items at SMK Negeri 1 Bangli based on Classical Test Theory, and the analysis answers the research question. The instrument showed very high overall score consistency but still requires substantial improvement at the item level. Most items did not yet measure the intended competencies adequately, and the distribution of item difficulty was unbalanced because easy items dominated the test. Although a considerable proportion of items could distinguish between students with different ability levels, several items showed weak discrimination and should be revised or eliminated. Most distractors did not function as intended, indicating that many incorrect answer options were not plausible enough to attract students who had not mastered the material. Accordingly, the research succeeded in producing an empirical basis for deciding which items to retain, revise, or replace. Future test development should include systematic empirical item analysis before items are reused, with particular attention to content alignment, item construction, language clarity, discrimination quality, and distractor plausibility. Teachers are also encouraged to maintain an item bank that records analysis results so that high-quality items can be retained, weaker items can be improved, and unsuitable items can be removed. Further research may compare Classical Test Theory results with modern measurement approaches, such as Rasch analysis, to obtain a more comprehensive evaluation of item quality.

Acknowledgement

The authors express their gratitude to the academic supervisors for their guidance, direction, and constructive feedback throughout the research and article-writing process. Appreciation is also extended to the Principal of SMK Negeri 1 Bangli, the Indonesian language teacher, and all Grade X students who assisted and participated in the study. The authors further thank all parties who provided support, assistance, and encouragement that contributed to the successful completion of this research.

References

Ashayidatina, L., & Alber, A. (2023). Analisis Butir Soal Ujian Semester Ganjil Bahasa Indonesia Kelas X SMK Negeri 1 Mempura. *JP-BSI (Jurnal Pendidikan Bahasa dan Sastra Indonesia)*, *8*(2), 131. https://doi.org/10.26737/jp-bsi.v8i2.5039

Bella, S., & Azhar, A. (2023). Development of Think-Pair-Share (TPS) Model-Based Learning Tools for Global Warming Materials. *Journal of Education and Learning Research*, *1*(1), 20–29. https://doi.org/10.62208/jelr.1.1.p.20-29

Hamimi, L., Zamharirah, R., & Rusydy, R. (2020). Analisis Butir Soal Ujian Matematika Kelas VII Semester Ganjil Tahun Pelajaran 2017/2018. *Mathema: Jurnal Pendidikan Matematika*, *2*(1), 57. https://doi.org/10.33365/jm.v2i1.459

Ketaren, M. A., Nainggolan, J. R. B., Sitepu, H. E. B., & Simanjorang, D. (2024). Validitas dan Tingkat Reliabilitas Soal Buatan Pilihan Ganda dengan Tes Sumatif Siswa Kelas V UPT SD Negeri 065013 Medan. *Jurnal Pengabdian Masyarakat Bangsa*, *1*(11), 2884–2889. https://doi.org/10.59837/jpmba.v1i11.620

Novriyanti, E., Sarifah, I., & Riyadi, R. (2024). Analisis Kualitas Butir Soal Ujian Tengah Semester Biologi Umum Menggunakan Teori Tes Klasik. *Jurnal Riset dan Inovasi Pembelajaran*, *4*(3), 1844–1860. https://doi.org/10.51574/jrip.v4i3.2245

Olude, S. A., Oludipe, O. S., Saibu, S. O., Ademola, I. A., Owolabi, T., & Akintoye, H. (2025). Incorporating Local Craft in Chemical Processes to Enhance Students' Comprehension of Concepts and Interest in Chemistry. *Journal of Education and Learning Research*, *2*(2), 105–117. https://doi.org/10.62208/jelr.2.2.p.105-117

Putri, D., & Retnosari, D. S. (2024). Analisis Kualitas Soal Penilaian Akhir Semester Mata Pelajaran Dasar Program Keahlian Akuntansi dan Keuangan Lembaga di SMK Ketintang Surabaya. *Liabilities (Jurnal Pendidikan Akuntansi)*, *7*(1), 8–17. https://jurnal.umsu.ac.id/index.php/LIAB/article/view/17186

Qadir, A., Huda, N., & Hermina, D. (2024). Analisis Butir Tes: Tingkat Kesukaran, Daya Pembeda dan Efektivitas Pengecoh. *Al-Furqan: Jurnal Agama, Sosial, dan Budaya*, *3*(3), 1450–1467. https://publisherqu.com/index.php/Al-Furqan/article/view/1069

Ramadhan, M. F., Siroj, R. A., & Afgani, M. W. (2024). Validitas dan Reliabilitas. *Journal on Education*, *6*(2), 10967–10975. https://doi.org/10.31004/joe.v6i2.4885

Riinwati. (2021). *Pengantar Evaluasi Pendidikan ISI*. Thema Publishing.

Salsabilla, N., Khairani, Z., & Marisya, S. (2025). Analisis Butir Soal Penilaian Akhir Semester (PAS) Ganjil Mata Pelajaran Bahasa Indonesia Kelas XI Akuntansi SMK Negeri 2 Padang. *Jurnal Armada Pendidikan*, *3*(3), 339–349. https://doi.org/10.60041/jap/v3i3.391

Simbolon, J. (2023). Transformasi Pembelajaran Bahasa Indonesia melalui Penerapan Literasi di Sekolah. *JBSI: Jurnal Bahasa dan Sastra Indonesia*, *3*(01), 162–171. https://doi.org/10.47709/jbsi.v3i01.2941

Yusuf, F. W. (2024). Analisis Butir Soal Asesmen Sumatif Biologi Materi Perubahan Lingkungan dengan Menggunakan Anates pada Kelas X SMA. *LEARNING: Jurnal Inovasi Penelitian Pendidikan dan Pembelajaran*, *4*(1), 126–135. https://doi.org/10.51878/learning.v4i1.2777

.

How to cite this article:

Suryanti, N. W., Putrayasa, I. B., & Nurjaya, I. G. (2026). Analysis of the Quality of Grade X Indonesian Language Final Semester Assessment Items Based on Classical Test Theory. *Journal of Education and Learning Research*, *4(2)*, 1-8.