AN ANALYSIS OF TEACHER-MADE ENGLISH SUMMATIVE TEST ITEMS AT SENIOR HIGH SCHOOL

Sahrul Rizki[1*], Taufiqulloh[2*], Ihda Rosdiana[3*]

[123]*English Education Department, Universitas Pancasakti Tegal (INDONESIA)*

[*]*Corresponding author: sahrulrizki628@gmail.com*

Abstract

The implementation of Kurikulum Merdeka has changed English teaching and assessment in Indonesian senior high schools by emphasizing communicative and functional language use. In this context, teacher-made summative tests are expected to measure students' actual achievement accurately and fairly. This study evaluated the psychometric quality of a teacher-made English summative test under Kurikulum Merdeka. Using a quantitative descriptive evaluative design, the research analyzed 35 multiple-choice items administered to 108 Grade XI students from three classes at SMA Negeri 2 Slawi. Item validity was examined using Point-Biserial correlation, reliability using KR-20, and difficulty using item difficulty index. The pooled analysis showed that all items were valid, with 25.7% highly valid, 62.9% moderately valid, and 11.4% low valid. However, cross-class verification revealed instability, with 1 invalid item in XI-1, 3 in XI-2, and 10 in XI-5, indicating sample dependency bias. The test also showed very high reliability (KR-20 = 0.922). In terms of difficulty, 57.1% of items were easy, 42.9% moderate, and none difficult. These results suggest that although the test was reliable and generally valid, its item functioning was not fully stable across classes and its difficulty level was unbalanced.

Keywords: Teacher-made test; Item Analysis; Classical Test Theory; Reliability Paradox; Sample Dependency;

1 INTRODUCTION

The implementation of English teaching and assessment at the senior high school level in Indonesia has undergone substantial changes following the implementation of Kurikulum Merdeka. The current curriculum emphasizes the meaningful use of language in communicative, analytical, and functional contexts rather than the memorization of linguistic forms. Consequently, assessment practices are expected to measure students' ability to apply language knowledge in authentic situations and demonstrate higher-order thinking skills. In this context, English summative tests are no longer merely administrative instruments used to determine report card grades but have become essential tools for evaluating students' learning achievement and curriculum attainment (Pratiwi & Inayati, 2023). The need for high-quality assessment instruments has become increasingly important because students continue to experience learning challenges in the post-pandemic era. Learning loss has affected students' English proficiency development, making accurate assessment necessary for identifying actual learning outcomes and instructional needs (Firmansyah et al., 2023). Therefore, poorly designed tests may generate inaccurate information and lead to inappropriate pedagogical decisions (Semiun et al., 2022).

The importance of teacher-made assessments has increased significantly since the abolition of the National Examination (*Ujian Nasional*) through the Ministry of Education Circular Letter Number 1 of 2021. Under the School-Based Assessment system, schools and teachers are now fully responsible for evaluating student achievement and determining graduation outcomes. Although this policy reduces examination pressure and promotes instructional flexibility, it simultaneously places greater responsibility on teachers to develop valid and reliable assessment instruments (Tamsiyati et al., 2025). Consequently, the quality of teacher-made summative tests has become a critical factor in ensuring fairness and accountability within educational evaluation.

According to (Brown, 2004), a test can only function effectively when it consistently measures the intended construct. Similarly, Arikunto (2009) emphasizes that a quality educational test should fulfill several psychometric requirements, particularly validity, reliability, and an appropriate level of difficulty. Within the framework of Classical Test Theory (CTT), validity refers to the extent to which test items accurately measure the intended construct, reliability indicates the consistency of test scores, and difficulty level reflects the proportion of students who answer an item correctly (Allen & Yen, 1979;

Arikunto, 2009). These indicators are commonly employed to evaluate the quality of teacher-made assessments and ensure that test scores accurately represent students' actual abilities.

Several previous studies have investigated the psychometric quality of teacher-made English summative tests. Sugianto (2017) reported that 64% of test items were valid and that the instrument achieved a high reliability coefficient of 0.907. Rohmah (2019) found acceptable reliability despite weaknesses in item discrimination. Ningsih & Widowati (2021) concluded that test quality was only fair because several items exhibited ineffective distractors and weak discrimination power. Sengkaton et al. (2020) reported that most test items fell within moderate to poor quality categories. Meanwhile, Noviasmy & Risma (2024) identified a psychometric paradox in which a large proportion of valid items did not guarantee overall test reliability. These studies demonstrate the importance of conducting item analysis to evaluate the quality of teacher-made assessments.

Despite their contributions, previous studies share a common methodological limitation. Most analyses were conducted using a single aggregated sample and relied exclusively on overall test statistics. Such an approach is vulnerable to sample dependency bias, a well-known limitation of Classical Test Theory. An item may appear valid when analyzed using pooled data but may function differently across student groups with varying ability levels. Consequently, the reported validity of an item may represent an illusion of validity rather than genuine measurement stability. Furthermore, none of the previous studies conducted cross-class verification to examine whether item validity remained consistent across different classroom contexts.

To address this gap, the present study evaluated a teacher-made English summative test administered to eleventh-grade students at SMA Negeri 2 Slawi. The study employed a larger sample consisting of 108 students from three parallel classes and incorporated cross-class validity verification to examine potential sample dependency bias. Specifically, the study aimed to analyze the validity status of each test item and its consistency across three different class groups, determine the reliability coefficient of the test using KR-20, and classify the test items according to their difficulty levels within the framework of Classical Test Theory.

2 METHODOLOGY

This study employed a quantitative approach with a descriptive evaluative design to examine the psychometric quality of a teacher-made English summative test. Quantitative item analysis was selected because the study aimed to evaluate existing test items through statistical procedures rather than investigate causal relationships among variables. The evaluation focused on three psychometric indicators derived from Classical Test Theory (CTT), namely item validity, test reliability, and item difficulty (Arikunto, 2009; Creswell, 2014).

The population consisted of all Grade XI students at SMA Negeri 2 Slawi during the 2025/2026 academic year. Since all classes followed the same curriculum, syllabus, learning objectives, and assessment procedures, the population was considered homogeneous. Cluster random sampling was employed to select three representative classes from the nine available Grade XI classes. As a result, classes XI-1, XI-2, and XI-5 were selected as the research sample. The final sample consisted of 108 students who had completed the same English End-of-Semester Summative Test.

The research instrument was a teacher-made English summative test consisting of 35 multiple-choice items. The instrument was designed by the English teacher and administered as part of the official school assessment program. Students' answer sheets and the official answer key served as the primary sources of data for the analysis.

Data were collected through documentation techniques. Student responses were coded dichotomously, with correct answers scored as 1 and incorrect answers scored as 0. The data were then processed using SPSS 22. Item validity was analyzed using the Point-Biserial correlation, with items considered valid when the significance value was lower than 0.05. Test reliability was estimated using the Kuder-Richardson Formula 20 (KR-20), while item difficulty was calculated by dividing the number of correct responses by the total number of test takers. To reduce sample dependency bias, a cross-class validity verification procedure was conducted by recalculating item validity separately for classes XI-1, XI-2, and XI-5.

3 RESULTS

This section presents the empirical analysis of the research instrument, which consisted of 35 test items. The analysis was conducted based on response data obtained from 108 students who participated in the instrument try-out. The purpose of this examination was to evaluate the statistical quality of each item as well as the overall feasibility of the instrument before it was used in the main data collection. To provide a systematic overview, the results are presented in stages. The following section sequentially describes the results of the item validity test, the estimation of the instrument reliability coefficient, and the item difficulty analysis for each test item.

3.1 Validity Analysis

The validity analysis was conducted using the Point-Biserial correlation based on the responses of 108 students. The results revealed that all 35 test items obtained significance values below 0.05, indicating that every item met the statistical criterion for validity.

Table 1. *General Valifity of the Test.*

Validity Status	Number of Items
Valid	35
Invalid	0

As shown in Table 1, all 35 items met the validity criterion, indicating that the test successfully measured the intended construct. The absence of invalid items suggests that the instrument performed well when analyzed using the pooled dataset.

Table 2. *Summary of Validity Classification.*

Category	Number of Items	Percentage
High Validity	9	25.7%
Moderate Validity	22	62.9%
Low Validity	4	11.4%
Invalid	0	0%

As shown in Table 2, the validity profile was dominated by moderate items (22 items or 62.9%), followed by high-validity items (9 items or 25.7%), while only 4 items (11.4%) were classified as low validity and none were invalid. This pattern indicates that the test was generally acceptable in measuring the intended construct, although several items still need revision to improve their statistical contribution to the test.

To verify whether this perfect validity rate remained stable across student groups, a cross-class verification procedure was conducted.

Table 3. *Cross-Class Validity Verification.*

Class	Valid Items	Invalid Items
XI-1	34	1
XI-2	32	3
XI-5	25	10

As displayed in Table 3, the number of valid items decreased when the analysis was conducted separately for each class. The greatest discrepancy occurred in Class XI-5, where ten items failed to satisfy the validity criterion. Item 25 was particularly noteworthy because it was valid in the pooled analysis but became invalid in both Class XI-2 and Class XI-5. These findings indicate that item validity was not entirely stable across groups and suggest the presence of sample dependency bias.

The findings support the sample dependency characteristic of Classical Test Theory. Although the pooled analysis suggested excellent validity, the cross-class verification demonstrated that item performance varied according to the characteristics of each student group. Allen & Yen (1979) explain that item statistics are inherently dependent on the sample from which they are derived. Consequently, an item that appears valid in one population may not function similarly in another. The current findings reinforce the importance of cross-group verification when evaluating teacher-made tests because reliance solely on aggregated data may create an illusion of validity. Similar findings were reported by Sugianto (2017), who

found a high proportion of valid items in teacher-made assessments, although the stability of item performance across groups still required careful interpretation.

3.2 Reliability Analysis

The reliability analysis was conducted using the KR-20 formula through the Alpha model in SPSS to determine the internal consistency of the test items.

Table 4. *Reliability Statistics.*

Coefficient	Value	Category
KR-20	0.922	Very High

Table 3 shows that the English summative test obtained a KR-20 coefficient of 0.922, which falls within the very high category. This result indicates that the instrument had strong internal consistency and that students who answered some items correctly were also likely to answer other items correctly, reflecting a consistent response pattern throughout the test.

Brown (2004) emphasizes that reliability reflects the consistency of measurement and the reduction of random error. The relatively large number of items in the test may also have contributed to the high coefficient because longer tests generally provide more stable estimates of student ability. Similar to the validity findings, the reliability result suggests that the test functioned consistently at the score level and was capable of producing stable measurement outcomes.

However, reliability should not be interpreted as direct evidence of overall test quality. Although the reliability coefficient was exceptionally high, the cross-class validity analysis revealed instability in item functioning across groups. This pattern resembles the psychometric paradox identified by Noviasmy & Risma (2024), who reported that strong item statistics do not necessarily guarantee the overall quality of a teacher-made assessment. Therefore, the high KR-20 coefficient should be interpreted as evidence of score consistency rather than proof that all items function equally well across different classroom contexts.

3.3 Difficulty analysis

The item difficulty analysis revealed a relatively imbalanced distribution of difficulty levels across the test items.

Table 4. *Distribution of Item Difficulty Levels.*

Class	Valid Items	Invalid Items
Easy	20	57.1%
Moderate	15	42.9%
Difficult	0	0%

As presented in Table 4, easy items dominated the test, accounting for more than half of all items. Moderate items represented 42.9% of the test, while no item was classified as difficult. The easiest item was Item 5, with a difficulty index of 0.90, indicating that 90% of the students answered the question correctly. Conversely, Item 22 demonstrated the most ideal difficulty level, with an index of 0.47, indicating a nearly balanced distribution of correct and incorrect responses.

The dominance of easy items indicates that the test primarily measured lower-level learning outcomes. According to Arikunto (2009), a well-constructed test should contain a balanced combination of easy, moderate, and difficult items to maximize discrimination among students. Similarly, Brown (2004) argues that moderately difficult items provide the most useful information because they allow meaningful differences in student ability to emerge. The complete absence of difficult items therefore limits the capacity of the test to distinguish higher-achieving students from average performers.

The difficulty analysis suggests that the English summative test was more effective in measuring basic achievement than advanced proficiency. While the predominance of easy items may contribute to student confidence and high average scores, it simultaneously reduces the instrument's ability to identify students with superior mastery of the material. Taken together with the validity and reliability findings, the test can be considered statistically acceptable in general, but it still exhibits important psychometric weaknesses

related to validity stability and difficulty balance. These findings emphasize the importance of routine item analysis and cross-class verification in the development of teacher-made assessments.

4 CONCLUSIONS

This study evaluated the psychometric quality of a teacher-made English summative test administered to Grade XI students at SMA Negeri 2 Slawi. The pooled validity analysis indicated that all 35 items were statistically valid. However, cross-class verification revealed substantial variation in item performance, with one invalid item identified in Class XI-1, three invalid items in Class XI-2, and ten invalid items in Class XI-5. These findings indicate the presence of sample dependency bias and demonstrate that item validity was not fully stable across groups.

The reliability analysis produced a KR-20 coefficient of 0.922, indicating very high internal consistency. Nevertheless, the difficulty analysis showed an unbalanced distribution, consisting of 57.1% easy items, 42.9% moderate items, and no difficult items. Therefore, although the test demonstrated strong statistical consistency, its psychometric quality was not uniformly robust across classroom contexts. The findings highlight the importance of conducting routine item analysis and cross-class verification to ensure that teacher-made assessments provide fair, objective, and stable measurements of student achievement.

REFERENCES

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Brooks/Cole Publishing Company.

Arikunto, S. (2009). *Dasar-Dasar Evaluasi Pendidikan* (9th ed.). Bumi Aksara.

Brown, H. D. (2004). *Language assessment: Principles and Classroom Practices* (7th ed.). Pearson Education.

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

Firmansyah, B., Hamamah, H., & Emaliana, I. (2023). Recent Students' Motivation Toward Learning English After the COVID-19 Post-Pandemic. *Journal of Languages and Language Teaching*, *11*(1), 130. https://doi.org/10.33394/jollt.v11i1.6635

Ningsih, N. A., & Widowati, W. (2021). Utilizing Test Item Analysis to Portray the Quality of English Final Test. *English Teaching Journal : A Journal of English Literature, Language and Education*, *9*(2), 143. https://doi.org/10.25273/etj.v9i2.10867

Noviasmy, Y., & Risma. (2024). An Analysis of The English Summative Test: EFL Teacher-Made Test. *JELITA*, *5*(1), 41–50. https://doi.org/10.56185/jelita.v5i1.618

Pratiwi, W. Y., & Inayati, N. L. (2023). Integration of the Merdeka Curriculum and the Cambridge Curriculum (Case study of an international class at SMA Batik 1 Surakarta). *Muaddib: Studi Kependidikan Dan Keislaman*, *13*(2), 115–125.

Rohmah, N. (2019, January 1). Validity and Reliability Study on Teacher-Made Assessment for English Mid-Term Examination. *Proceedings of the Eleventh Conference on Applied Linguistics (CONAPLIN 2018)*. https://doi.org/10.2991/conaplin-18.2019.236

Semiun, T. T., Wisrance, M. W., & Napitupulu, M. H. (2022). English Summative Test: The Quality of Its Items. *English Education:Journal of English Teaching and Research*, *7*(2), 119–127. https://doi.org/10.29407/jetar.v7i2.18347

Sengkaton, I., Hambali, M., & Mirizon, S. (2020). Teacher-made Summative Test: An Analysis of Test Format, Index of Difficulty, Discrimination, and Distractors. *Lingua, Jurnal Bahasa & Sastra*, *20*(2).

Sugianto, A. (2017). Validity and Reliability of English Summative Test for Senior High School. *Indonesian EFL Journal: Journal of ELT, Linguistics, and Literature*, *3*.

Tamsiyati, E., Agissa, A. W., Yuniar, Y., & Junaidah, J. (2025). The Impact of Eliminating National Examinations on Educa-tional Evaluation Quality in Indonesia: A Policy Analysis on Equity and Accountability. *International Journal of Education and Literature*, *4*(1), 293–304. https://doi.org/10.55606/ijel.v4i1.216