LedTure
Journal of Language, Education, Literature and Culture Vol.3, No.2: December 2025
https://bing.ukaw.ac.id/ejournal/index.php/ledture. e-ISSN Online: 3026-0302 DOI: 10.33323/l.v3i2.71

AN ANALYSIS OF ENGLISH FINAL SEMESTER TEST ITEMS AT SMK KENCANA SAKTI HAUMENI

Asrit Elusry Sanam[1*], Norci Beeh[2], and Alfred Snae[3]
[1,2,3]Universitas Kristen Artha Wacana
[*)]Corresponding author: Aveibrahim@gmail.com
Received date: 13/112025; Accepted date: 31/12/2026

Abstract This study analyzed the quality of English final semester test items developed by the English teacher of SMK Kencana Sakti Haumeni. Although many studies had examined teacher-made tests in urban contexts, few focused on rural vocational schools, where teachers often designed assessments without empirical validation. This study addressed that gap by evaluating the test's effectiveness through item difficulty and discrimination indices. Using a quantitative descriptive method, the researcher analyzed 50 multiple-choice items from the second-grade English final semester test. The analysis applied Heaton's (1988) formulas for facility value and discrimination power, with classifications from Sumarsono and Arikunto (Hartati & Yogi, 2019). The results showed that 18% of the items were difficult, 50% were moderate, and 32% were easy. Regarding discrimination power, 22% of the items were good, 30% satisfactory, 40% poor, and 8% bad or negative. These findings indicated that while most items met acceptable standards, nearly half failed to distinguish student ability effectively. The study contributed to assessment practice by emphasizing the importance of regular item analysis to help teachers design more valid, reliable, and diagnostic classroom tests aligned with learning objectives.

Keywords*: *item analysis, test quality, discrimination index, English assessment, vocational education.

INTRODUCTION

Assessment was one of the essential components of the teaching and learning process because it served as a tool to measure students' progress and achievement. Through assessment, teachers evaluated how well students understood learning materials and identified areas that required improvement. According to Kusumawati and Hadi (2018), educational evaluation enabled teachers to determine students' strengths and weaknesses, which supported the design of effective remedial or enrichment programs. As stated in the Indonesian Minister of Education and Culture Regulation on Educational Assessment Standards (2016), assessment referred to the process of collecting and analyzing data to determine the extent to which students achieved learning objectives. Therefore, a well- constructed assessment instrument was crucial to ensure valid and reliable learning outcomes.

In the context of language education, the quality of test items played a vital role in accurately reflecting students' abilities. Poorly constructed test items could lead to misleading interpretations of students' performance and hinder the effectiveness of the teaching process. According to Classical Test Theory (CTT), a student's observed score consisted of a true score and an error component, implying that any test must minimize measurement error to ensure accurate results (Heaton, 1988; Brown, 2004). Test validity and reliability were therefore essential indicators of assessment quality. Validity referred to the extent to which a test measured what it was intended to measure, while reliability referred to the consistency and stability of the results across time or equivalent forms (Maharani & Putro, 2020; Noviasmy &

Risma, 2024). One of the most effective ways to improve both validity and reliability was through *item analysis*, a process that evaluated the characteristics of test items—such as

difficulty level, discrimination power, and distractor efficiency—based on empirical evidence rather than subjective judgment (Hartati & Yogi, 2019; Nugraha & Suparman, 2022).

The final semester test was one of the most common forms of summative evaluation used to measure students' achievement at the end of an academic period. In English as a Foreign Language (EFL) classrooms, such tests often took the form of multiple-choice items because they allowed teachers to assess a wide range of materials efficiently and objectively (Suardipa & Primayana, 2018). However, despite their practicality, many teacher-made tests were developed intuitively, without empirical analysis to verify their quality. As a result, these tests might not accurately measure the intended language skills, which reduced their diagnostic and evaluative value (Putri & Rahmawati, 2021; Fitria, 2023).

In East Nusa Tenggara, particularly in rural areas such as South Central Timor Regency, studies on item analysis in English language testing remained limited. Teachers often designed test items based on intuition or adapted them from textbooks without empirical validation. Consequently, the quality of these tests varied, and some items failed to differentiate effectively between high- and low-performing students. This problem highlighted the importance of applying Classical Test Theory principles to evaluate item quality and strengthen the credibility of classroom-based assessments in rural contexts.

Although several national studies had investigated teacher-made tests (e.g., Pradanti, Martono, & Sarosa, 2018; Maharani & Putro, 2020; Noviasmy & Risma, 2024), their findings revealed several limitations that justified further research. First, most of these studies focused on urban or well-resourced schools and did not represent rural settings where teachers had limited access to professional training in test construction. Second, previous research often analyzed a large number of tests collectively, which limited in-depth understanding of individual test performance. Third, many of the earlier studies discussed item difficulty and discrimination indices but did not explicitly link their findings to the theoretical foundations of validity, reliability, and Classical Test Theory. Fourth, previous works rarely addressed the pedagogical implications of item analysis for improving teachers' assessment literacy and reflective practice. These limitations suggested the need for a more localized, theoretically grounded, and practice-oriented study.

Therefore, this study aimed to analyze the quality of English final semester test items developed by the English teacher of SMK Kencana Sakti Haumeni. It focused on two essential aspects of test quality—difficulty level and discrimination power—to evaluate the effectiveness of the test. The results of this study were expected to provide empirical evidence to guide teachers in designing valid and reliable English assessments and to encourage the regular use of item analysis as a reflective practice in classroom assessment, especially in rural and vocational school contexts. The study addressed the following research questions: (1) What were the levels of difficulty of the English final semester test items at SMK Kencana Sakti Haumeni? (2) What were the discrimination indices of the English final semester test items? and (3) What were the implications of the item analysis results for improving teacher-made English assessments in rural vocational schools?

METHOD

This study employed a quantitative descriptive research design, which focused on describing and analyzing numerical data to identify the characteristics of test items. According to Creswell (2009), quantitative research involved the systematic collection and analysis of numerical data to identify patterns, relationships, or trends. This approach was appropriate because the purpose of the study was to examine the psychometric quality of English test items developed by the English teacher of SMK Kencana Sakti Haumeni.

The data source consisted of the English final semester test and students' answer sheets from the second-grade class of SMK Kencana Sakti Haumeni during the 2024/2025 academic year. The test contained 50 multiple-choice items designed to measure students' comprehension of English materials taught throughout the semester. The items were developed based on the school's test blueprint, which followed the English syllabus of the

2013 Curriculum (*Kurikulum 2013*). The test blueprint covered three main competencies: (1) reading comprehension, which assessed students' understanding of short texts, dialogues, and notices; (2) vocabulary mastery, which measured word meaning and contextual usage; and (3) grammar understanding, which evaluated students' knowledge of sentence structure and function. The distribution of items was proportionally balanced across these competencies to represent the overall learning objectives of the English subject.

The data collection technique used in this study was documentation. As stated by Ary et al. (2009), document analysis is a research method used to study existing written or recorded materials to obtain factual information. Therefore, the researcher collected the original test papers and students' answer sheets to obtain authentic data for item analysis. Ethical approval for conducting this research was obtained from the English Education Study Program, Faculty of Teacher Training and Education, Artha Wacana Christian University. In addition, permission was formally granted by the principal of SMK Kencana Sakti Haumeni. All student data were treated confidentially and anonymized to ensure ethical compliance.

The research instrument consisted of the set of multiple-choice test items, which were analyzed to determine their difficulty level and discrimination power. The data analysis procedure followed several systematic steps. First, the total scores of all students were ranked from the highest to the lowest. Second, based on Heaton's (1988) recommendation, 27% of the total sample was selected as the upper group and 27% as the lower group. In this study, 12 students represented the upper group and 12 represented the lower group, which provided an optimal balance between group size and statistical stability for item analysis in small classroom samples. Third, each test item was analyzed using Heaton's (1988) formulas: the Facility Value (FV) to determine item difficulty and the Discrimination Power (DP) to evaluate how well each item distinguished between high- and low-performing students. Fourth, the results were categorized according to the classification proposed by Sumarsono and Arikunto (in Hartati & Yogi, 2019), which grouped items into categories such as easy, moderate, or difficult for FV, and excellent, good, satisfactory, poor, or bad for DP. Fifth, items were interpreted to determine whether they should be retained, revised, or discarded based on both indices.

This systematic procedure ensured that the analysis was objective, replicable, and empirically grounded. By combining Classical Test Theory principles with Heaton's analytical model, the method provided valid and reliable evidence of the

quality of the English final semester test items used at SMK Kencana Sakti Haumeni.

RESULTS

This section presents the quantitative findings from the analysis of the English final semester test items developed by the English teacher of SMK Kencana Sakti Haumeni. The analysis focused on two psychometric properties: (1) item difficulty and (2) discrimination power, using Heaton's (1988) formulas. The test consisted of 50 multiple-choice items administered to second-grade students during the 2024/2025 academic year. Students' total scores were ranked, and the top and bottom 27% (12 students each) were selected as the upper and lower groups for item analysis.

1.1 Difficulty Level of Test Items

The analysis revealed variation in item difficulty across the test. Facility Values (FV) were calculated for each item by dividing the total number of correct responses from both groups by the total number of students. Table 1 summarizes the classification based on Sumarsono's criteria (Hartati & Yogi, 2019).

Table 1. Classification of Item Difficulty

Category	Range	Number of Items	Percentage	Item Numbers
Difficult	0.00–0.30	9	18%	9, 13, 15, 31, 33, 40, 43, 45, 50
Moderate	0.31–0.70	25	50%	4, 5, 6, 7, 10, 11, 12, 14, 18, 19, 24, 25, 26, 28, 29, 32, 34, 36, 37, 38, 39, 42, 44, 46, 49
Easy	0.71–1.00	16	32%	1, 2, 3, 8, 16, 17, 20, 21, 22, 23, 27, 30, 35, 41, 47, 48

Half of the items (50%) were of moderate difficulty, which suggested a balanced level of challenge. However, the proportion of easy (32%) and difficult (18%) items slightly deviated from the ideal ratio of 25–50–25, indicating some inconsistency in the alignment between test items and students' proficiency levels

.

1.2 **Discrimination Power of Test Items**

Discrimination Power (DP) measured how effectively each item distinguished between high- and low-performing students. The values were calculated using Heaton's (1988) formula, and classifications followed the standards proposed by Sumarsono and Arikunto (Hartati & Yogi, 2019).

Table 2. Classification of Discrimination Power

Category	Range	Number of Items	Percentage	Item Numbers
Excellent	0.71–1.00	0	0%	—
Good	0.41–0.70	11	22%	6, 10, 11, 12, 14, 29, 32, 34, 36, 39, 42
Satisfactory	0.21–0.40	15	30%	4, 5, 8, 9, 13, 16, 20, 22, 25, 31, 35, 37, 44, 46, 50

Poor	0.00–0.20	20	40%	1, 2, 3, 15, 17, 18, 19, 21, 23, 24, 26, 27, 28, 33, 38, 40, 41, 43, 47, 48
Bad/Negative	<0	4	8%	7, 30, 45, 49

The findings indicated that 52% of the items (good and satisfactory) met acceptable standards, while 48% (poor and bad) did not effectively discriminate between high and low achievers. Four items had negative discrimination values, suggesting flaws in wording or content familiarity.

DISCUSSION

The results revealed both strengths and weaknesses in the quality of the English final semester test items. The dominance of moderately difficult items suggested that the teacher had designed questions suitable for most students' ability levels. Similar results were observed by Fitria (2023), who reported that teacher-made English tests often contained a majority of moderate items due to teachers' intuitive balancing of content difficulty. However, the presence of 18% difficult and 32% easy items indicated that the test lacked optimal proportionality, echoing findings from Nugraha and Suparman (2022), who emphasized that unbalanced difficulty distributions can distort students' true performance levels.

In terms of discrimination power, the finding that nearly half of the items performed poorly aligned with studies by Maharani and Putro (2020) and Noviasmy and Risma (2024), which found that teacher-made tests frequently fail to distinguish student ability effectively. Possible causes include unclear wording, ambiguous distractors, or a mismatch between item content and students' cognitive levels (Brown, 2004). The four negatively discriminating items suggested that some questions confused stronger students or favored weaker ones, consistent with McCowan and McCowan's (1999) view that weak distractors can significantly lower discrimination quality.

These findings contribute to *English Language Teaching (ELT) assessment practice* in several ways. First, they provide empirical evidence supporting the integration of Classical Test Theory (CTT) principles in classroom-based assessment, encouraging teachers to rely on item statistics rather than intuition. Second, they highlight the importance of ongoing item validation to ensure that each test fairly represents the intended competencies outlined in the curriculum. Third, they emphasize assessment literacy as a crucial skill for EFL teachers, particularly in rural contexts where professional support may be limited (Fitria, 2023; Putri & Rahmawati, 2021).

The results also reinforce that item analysis can serve as a professional reflection tool, helping teachers to revise, retain, or discard items based on objective evidence. This aligns with current trends in assessment reform in Indonesia, which stress formative reflection and data-driven decision-making (Kurniasih, 2022; Widodo, 2023). By applying regular item analysis, teachers in vocational schools can enhance the validity, reliability, and fairness of their English assessments, contributing to more accurate evaluation of students' learning outcomes.

CONCLUSION

This study aimed to analyze the quality of English final semester test items developed by the English teacher of SMK Kencana Sakti Haumeni, focusing on item difficulty and discrimination power. The findings showed that half of the items were

moderately difficult, while the remaining items were either too easy or too difficult. In terms of discrimination power, only 52% of the items met acceptable standards. These results suggest that the test was reasonably balanced in difficulty but limited in its ability to differentiate student performance effectively.

The study's conclusions are consistent with its objectives: to provide empirical insights into teacher-made test quality and promote assessment improvement in EFL contexts. The results underline the pedagogical importance of regular item analysis in ensuring valid, reliable, and fair classroom tests. Teachers should use statistical evidence to refine test items, rather than relying solely on intuition or textbook sources. Future research could expand the scope of item analysis by including distractor efficiency and reliability coefficients to provide a more comprehensive evaluation of test quality. Comparative studies across different schools or regions would also help identify contextual differences in assessment practices. Furthermore, experimental studies integrating teacher training on assessment literacy could examine whether professional development improves item-writing quality.

REFERENCES

Ary, D., Jacobs, L. C., & Sorensen, C. (2009). *Introduction to research in education* (8th ed.). Belmont, CA: Wadsworth Cengage Learning.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices.* White Plains, NY: Pearson Education.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Hartati, T., & Yogi, A. (2019). An analysis of teacher-made English test items based on item difficulty and discrimination index. *Journal of English Language Teaching and Linguistics, 4*(2), 89–101. https://doi.org/10.21462/jeltl.v4i2.234

Heaton, J. B. (1988). *Writing English language tests.* New York, NY: Longman.

Kusumawati, D., & Hadi, M. (2018). The role of assessment in improving English learning outcomes. *International Journal of English Language Education, 6*(1), 45–57. https://doi.org/10.5296/ijele.v6i1.12703

Maharani, I. A. P., & Putro, N. H. (2020). An analysis of multiple-choice test items for senior high school students. *Indonesian Journal of English Education, 7*(2), 145–158. https://doi.org/10.15408/ijee.v7i2.17612

McCowan, R. J., & McCowan, S. C. (1999). *Item analysis for criterion-referenced tests.* New York, NY: State University of New York Press.

Noviasmy, F., & Risma, H. (2024). Evaluating teacher-made tests: A case study of English summative assessment in Indonesia. *TEFLIN Journal, 35*(1), 77–95. https://doi.org/10.15639/teflin.v35i1.77-95

Pradanti, D. A., Martono, A., & Sarosa, T. (2018). An analysis of English test items based on validity, reliability, and discrimination index. *Indonesian Journal of Language Teaching, 3*(1), 12–20.

Semiun, E., & Luruk, S. (2020). Teachers' evaluation practices in EFL classrooms: A study in Kupang. *Journal of Educational Research and Practice, 10*(4), 223–233.

Suardipa, I. P., & Primayana, K. H. (2018). The implementation of summative assessment in English language teaching. *Jurnal Pendidikan Bahasa Inggris, 6*(2), 112–122.

Sumarsono, P., & Arikunto, S. (2015). *Dasar-dasar evaluasi pendidikan.* Jakarta: Bumi Aksara