**IDEAS**

**Journal of Language Teaching and Learning, Linguistics and Literature**

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Volume 14, Number 1, June 2026
pp. 1433 - 1438



Issued by English study program of UIN Palopo

# Item Analysis of an English Summative Test: A Classical Test Theory Approach in Indonesian Junior High School Context

**Rahmawaty Ahmad[1], Rasuna Rasid Talib[2], Magvirah El Walidayni Kau[3]**

[1,2,3]Universitas Negeri Gorontalo

**Article Info**

**Received:** 2026-04-21

**Revised:** 2026-05-05

**Accepted:** 2026-05-06

***Keywords:***
*classical test theory,*
*discriminating power,*
*item analysis,*
*summative test,*
*validity*

**DOI*:***
10.24256/ideas. v14i1. 10243

**Corresponding Author:**
Rahmawaty Ahmad
innahmad08@gmail.com
Universitas Negeri Gorontalo

**Abstract**

*Assessment plays a crucial role in evaluating students' learning outcomes; however, the quality of summative test instruments often remains questionable due to the limited use of systematic evaluation procedures. This study aims to examine the quality of English summative test items administered to eighth-grade students at SMP Negeri 1 Suwawa Timur using Classical Test Theory (CTT). A descriptive quantitative design was employed, analyzing 40 multiple-choice items completed by 56 students using Anates V4 software. The analysis focused on item validity, difficulty level, discriminating power, and distractor effectiveness. The results revealed that only 55% of the items were valid, while 45% were invalid. Most items (62.5%) were of moderate difficulty, yet no easy items were identified, indicating an imbalance. Furthermore, a substantial proportion of items demonstrated weak discriminating power, and several distractors were ineffective. These findings suggest that although some test items meet acceptable standards, a considerable number require revision. This study highlights the importance of systematic item analysis in improving the validity, reliability, and overall quality of assessment instruments.*

## 1. Introduction

Assessment is a fundamental component of the teaching and learning process, as it provides essential information about students' achievement and instructional effectiveness. Among various forms of assessment, summative tests are widely used to evaluate students' learning outcomes at the end of an instructional period (Arifin, 2019). However, the effectiveness of such assessments depends largely on the quality of the test items.

A high-quality test must be valid, reliable, and capable of accurately measuring students' competencies (Arikunto, 2018). Inadequate test construction may lead to inaccurate measurement, where test items fail to distinguish between students who have mastered the material and those who have not (Suek, 2021). Therefore, item analysis becomes a crucial procedure in ensuring the quality of assessment instruments.

Item analysis refers to a systematic process of examining test items based on students' responses to determine their effectiveness (Ratnawulan, 2014). It evaluates key indicators such as validity, difficulty level, discriminating power, and distractor effectiveness. These components function as an integrated system in determining whether test items should be retained, revised, or discarded.

Previous studies have emphasized the importance of item analysis in improving test quality (Hartati & Yogi, 2019; Fiska et al., 2021). However, empirical evidence suggests that many teachers do not routinely conduct such analysis due to limited knowledge and time constraints. This issue is also evident in SMP Negeri 1 Suwawa Timur, where summative tests are rarely evaluated systematically, potentially affecting the accuracy of student assessment outcomes.

Despite the recognized importance of item analysis, studies examining the quality of English summative tests in Indonesian junior high school contexts remain limited. Therefore, this study aims to analyze the quality of English summative test items using Classical Test Theory (CTT), focusing on validity, difficulty level, discriminating power, and distractor effectiveness. The findings are expected to contribute to improving assessment practices and enhancing test quality in EFL contexts.

## 2. Method

This study employed a descriptive quantitative research design aimed at systematically analyzing the quality of test items. This approach allows for objective evaluation of test characteristics using statistical indicators (Creswell & Creswell, 2019). The data consisted of: 40 multiple-choice test items. Responses from 56 eighth-grade students. English summative test documents the study was conducted at SMP Negeri 1 Suwawa Timur during the 2022/2023 academic year. The sample was selected purposively, as it represented a typical summative test used in classroom assessment.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Data were collected through documentation, including: test items, answer keys, students' answer sheets

The data were analyzed using Anates V4 software based on Classical Test Theory (CTT), focusing on four indicators:

1. **Item Validity**
   Measured using point-biserial correlation. Items were considered valid if $r > 0.304$.
2. **Difficulty Level**
   Calculated using the formula:
   $P = B / T$
   where B = number of correct responses, T = total number of students.
3. **Discriminating Power**
   Used to determine how well items differentiate between high- and low-performing students.
4. **Distractor Effectiveness**
   Distractors were considered effective if selected by at least 5% of students.

## 3. Result

The findings are presented based on the four indicators of item quality:

*Item Validity*

Out of 40 items, 22 items (55%) were classified as valid, while 18 items (45%) were invalid. Several items showed negative correlations, indicating that they failed to measure the intended constructs effectively. This suggests that nearly half of the test items did not meet basic validity requirements.

*Difficulty Level*

The distribution of difficulty levels showed that:

- 62.5% of items were moderate
- 32.5% were difficult
- 5% were too difficult
- 0% were easy

This imbalance indicates that the test lacked variation in difficulty, particularly the absence of easy items, which may disadvantage lower-performing students.

*Discriminating Power*

The results revealed that:

- 25% of items had excellent discrimination
- 30% were good
- 12.5% were satisfactory
- 32.5% were poor

This indicates that a considerable number of items were ineffective in distinguishing between high- and low-achieving students, reducing the overall reliability of the test.

*Distractor Effectiveness*

While some distractors functioned effectively, many were not selected by students, suggesting that they were implausible or poorly constructed. Ineffective distractors reduce the diagnostic value of multiple-choice items.

**4. Discussion**

The findings indicate that the overall quality of the test items is moderate but requires significant improvement.First, the high proportion of invalid items suggests problems in test construction, particularly in aligning items with learning objectives. According to Classical Test Theory, validity is a fundamental requirement for accurate measurement, and invalid items undermine the credibility of the assessment (Puspitaningsih et al., 2019).

Second, the absence of easy items reflects an imbalance in test difficulty. A well-constructed test should include a range of difficulty levels to accommodate students with varying abilities (Choirunisa, 2021). The lack of easy items may contribute to student frustration and lower overall performance.Third, the low discriminating power of several items indicates that these items fail to differentiate between students' ability levels. This suggests that some items may be either too ambiguous or too easy/difficult, limiting their effectiveness as assessment tools (Novriyanti & Arthur, 2024).

Fourth, ineffective distractors indicate weaknesses in item design. In multiple-choice tests, distractors should be plausible enough to attract students who lack understanding. Poor distractor construction reduces item quality and affects test validity (Muluki, 2020). From a theoretical perspective, these findings illustrate how the four components of Classical Test Theory—validity, difficulty, discrimination, and distractor effectiveness—function as an interconnected system. Weakness in one component can affect the overall quality of the test.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

This study contributes to assessment research by providing empirical evidence of test quality issues in a real classroom context. Practically, it highlights the importance of teacher competence in test construction and the need for regular item analysis.However, this study is limited to one test and one school context, which may limit generalizability. Future studies are recommended to include larger samples and incorporate reliability analysis for more comprehensive evaluation.

## 5. Conclusion

This study concludes that the English summative test used at SMP Negeri 1 Suwawa Timur has moderate quality but requires significant improvement. Only 55% of items were valid, and many items showed weaknesses in difficulty level, discriminating power, and distractor effectiveness.

To improve test quality, it is recommended that teachers:

1. Revise or eliminate invalid items
2. Ensure balanced difficulty levels
3. Improve distractor construction
4. Conduct regular item analysis

Implementing these recommendations will enhance the validity and reliability of assessment instruments and support better educational outcomes.

## 6. References

Arifin, Z. (2019). Evaluasi pembelajaran. Remaja Rosdakarya.

Arikunto, S. (2018). Dasar-dasar evaluasi pendidikan (3rd ed.). Bumi Aksara.

Atmowardoyo, H. (2018). Research methods in TEFL studies: Descriptive research, case study, error analysis, and R&D. Journal of Language Teaching and Research, 9(1), 197–204.

Bani, M., & Masruddin, M. (2021). Development of Android-based harmonic oscillation pocket book for senior high school students. JOTSE: Journal of Technology and Science Education, 11(1), 93-103.

Creswell, J. W., & Creswell, J. D. (2019). Research design: Qualitative, quantitative, and mixed methods approach (5th ed.). Sage Publications.

Fiska, N., et al. (2021). Analysis of item validity in educational assessment. Journal of Educational Measurement, 12(2), 45–56.

Hartati, S., & Yogi, A. (2019). Item analysis in educational evaluation. Indonesian Journal of Education, 8(1), 67–75.

Himawan, R., et al. (2024). Distractor effectiveness in multiple-choice tests. Journal of Educational Assessment, 15(1), 23–35.

Mistiani, R. (2020). Classical test theory in educational measurement. Educational Research Journal, 10(2), 89–98.

Muluki, A. (2020). Constructing effective distractors in multiple-choice tests.

Language Testing Journal, 7(1), 55–63.

Novriyanti, D., & Arthur, R. (2024). Discriminating power in educational testing. Assessment in Education Journal, 14(2), 101–115.

Puspitaningsih, R., et al. (2019). Validity in educational assessment. Journal of Educational Studies, 11(3), 145–156.

Ratnawulan, E. (2014). Evaluasi pembelajaran. Pustaka Setia.

Rustan, E. (2025). Developing a Web-Based Mobile Game to Enhance Students' Motivation in learning English Vocabulary. IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature, 13(1), 46-73.

Sugiyono. (2013). Metode penelitian pendidikan. Alfabeta.

Suek, A. (2021). The importance of item analysis in improving test quality. Journal of Educational Evaluation, 9(2), 88–96.