International Journal of Multicultural and Multireligious Understanding

http://ijmmu.com
editor@ijmmu.com
ISSN 2364-5369
Volume 12, Issue 8
August, 2025
Pages: 297-303

Item Analysis of EFL Test Using AnBuSo: Difficulty, Discrimination Power, and Distractor Functionality

Vina Advianti; Yuyun Yulia

Universitas Negeri Yogyakarta, Indonesia

http://dx.doi.org/10.18415/ijmmu.v12i8.6963

Abstract

This study investigates the quality of a teacher-constructed English language multiple-choice test using item analysis through AnBuSo (Analisis Butir Soal). The test, which consisted of 30 items assessing vocabulary, grammar, and reading comprehension, was administered to 20 students from the WINA Choir Program at STKIP Widya Yuwana, Madiun. Utilizing a quantitative descriptive design, the research focused on evaluating item difficulty, discrimination power, and distractor functionality. Results showed that 83% of the items had moderate difficulty, while 73% demonstrated good discrimination indices. However, several items were found to have non-functional distractors or negative discrimination values, indicating the need for revision. The findings underscore the importance of item analysis in improving test quality and enhancing its diagnostic value. The study highlights how digital tools like AnBuSo can support effective assessment practices and guide teachers in revising items and designing targeted remedial instruction.

Keywords: *Item Analysis; Anbuso; Multiple-Choice Test*

Introduction

Assessment is a fundamental aspect of the educational process, serving not only as a tool to measure student achievement but also to guide instructional planning and curriculum development. In the context of English Language Teaching (ELT), assessment helps educators evaluate language components such as vocabulary, grammar, and reading comprehension. Hughes (2002) emphasizes that giving tests is one of the most commonly used classroom evaluation techniques, and Rame & Kesi (2023) further supports that assessment is an integral part of a teacher's responsibility.

Among the various types of test formats, Multiple Choice Questions (MCQs) are among the most widely used worldwide due to their efficiency, objectivity, and ease of scoring (Swanson et al., 2006). MCQs typically consist of a stem and several answer choices, one of which is correct. However, creating high-quality MCQs is not a simple task, particularly in designing plausible distractors that can effectively differentiate between students of varying proficiency levels (Maulina & Novirianthy, 2020). The use of well-constructed test items plays a significant role in accurately assessing students' competencies in specific domains. Karim et al., (2021) asserts that good test items allow teachers to measure students'

learning achievements effectively, while Darmawan et al., (2022) adds that the quality of a test significantly contributes to improving students' learning outcomes.

To ensure the effectiveness of MCQs, item analysis is often employed as a statistical method to evaluate individual item performance. Remmers et al., (1965) describe item analysis as a post-administration evaluation of test items, and this is supported Freeman (1955), who emphasize that the overall quality of a test relies heavily on the quality of its individual items. Through item analysis, teachers can identify items that are too easy, too difficult, or that fail to discriminate between high- and low-performing students.

Moreover, item analysis can enhance the efficiency of test development. According to Siri & Freddano (2011) improving test items through this process helps teachers and test designers save time and energy while maintaining accuracy. In line with this, Xu & Liu, (2009) argue that teachers' assessment knowledge is not fixed but is instead a complex and evolving process requiring continuous reflection and refinement. Thus, developing the skills to evaluate and improve test items becomes essential in fostering effective assessment practices.

In this study, AnBuSo (Analisis Butir Soal) was used as the primary tool to conduct item analysis. AnBuSo is a digital platform specifically designed to assist educators in evaluating test items based on established psychometric principles. By inputting student responses, the platform automatically calculates item difficulty, discrimination power, and distractor effectiveness providing valuable feedback for test improvement.

The test used in this research was developed as part of a final project in the *English Language Teaching and Learning Assessment and Evaluation* course at Universitas Negeri Yogyakarta. It comprised 30 multiple-choice items assessing vocabulary, grammar, and reading comprehension, administered to students in the WINA Choir Program at STKIP Widya Yuwana, Madiun. The student responses were analyzed using AnBuSo to identify the quality of each item and determine revisions where necessary.

By systematically analyzing each item, this study aims to identify strengths and weaknesses in the test design, offer recommendations for improvement, and contribute to better assessment practices in ELT. The research is guided by the following questions:

1. What is the distribution of difficulty and discrimination indices across the 30 multiple-choice items?
2. Which items require revision based on distractor inefficiency and item performance data?

Method

This study employed a quantitative descriptive design to investigate the quality of a teacher-developed multiple-choice English test. AnBuso (Analisis Butir Soal) is the primary tool for statistical evaluation. AnBuso is a platform designed to support educators in analyzing test quality through automatic calculations of key metrics such as item difficulty, discrimination power, and distractor functionality.

The participants involved in the study were 20 students from the WINA Choir Program at STKIP Widya Yuwana, Madiun, East Java. These students were selected through purposive sampling based on their participation in an extracurricular program that integrates musical experiences. Although they were not enrolled in a formal English course, their involvement in choir activities provided a relevant context for measuring English proficiency in vocabulary recognition, grammatical accuracy, and reading interpretation.

The instrument used in this research was a 30-item multiple-choice test developed by the researcher. The test focused on three major skills: vocabulary, grammar, and reading comprehension, and

was constructed using a Table of Specification (TOS) which provided on table 1. Each questions contained five alternatives, one of which was the correct answer, while the remaining four functioned as distractors. Prior to administration, the test was reviewed by a peer evaluator to assess the quality of content, construct, and language use. Based on the peer review, five items were revised to improve clarity and effectiveness.

The test was administered online using Google Forms. To ensure ethical administration, the researcher first obtained permission from the WINA Choir coach. The test link was then shared with the coach, who distributed it to choir members via the group's WhatsApp platform. Participants were instructed to complete the test independently. Their responses were collected automatically and compiled into a dataset for analysis.

After data collection, the responses were uploaded to the AnBuso platform for item analysis. The platform provided detailed item-level statistics, including the *difficulty index*, which indicates the proportion of students who answered correctly; the *discrimination index*, which reflects the item's ability to differentiate between high- and low-performing students; and an analysis of *distractor effectiveness*, identifying any options that were not chosen or were disproportionately ineffective. In addition to tabular data, AnBuso generated graphical representations that illustrated the overall distribution of item quality, difficulty levels, and discrimination power. These outputs were essential for determining which items were valid and reliable, and which required revision or elimination.

The results of the analysis were used not only to evaluate the technical quality of the test but also to design remedial instruction for students who did not meet the minimum mastery criteria. This methodological approach ensured that the assessment served both evaluative and formative purposes, ultimately supporting the enhancement of English language instruction and assessment practices.

Table 1. Table of Specification (ToS)

No	Material	Sub-material	Item Number
1.	Synonym and antonym	Contextual synonyms	1, 2, 3, 4, 5, 6, 7, 8
		Antonyms that are directly opposite	9, 20
2.	Tenses	Simple present tense	11, 13, 15, 18, 19
		Simple past tense	12, 14, 16, 17, 20
3.	Explanatory Text	Making a conclusion	21
		Gaining explicit and detailed information	22
		Gaining main idea of the Text	23
		Gaining Implicit and Detailed Information	24
		Guessing something through passage	25
4.	Recount Text	Making a Conclusion	26
		Gaining explicit and detailed information	27
		Gaining Main Idea of the Text	28
		Gaining Implicit and Detailed Information	29
		Guessing Something Through Passage	30

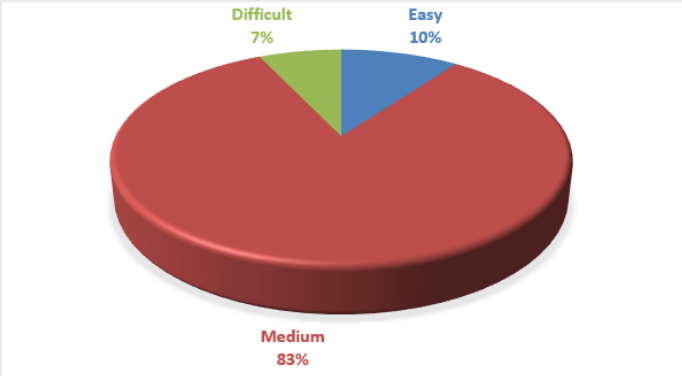
Findings and Discussion

The analysis of the 30 multiple-choice items revealed varying degrees of quality across key parameters: item difficulty level, discrimination power, and distractor effectiveness. Overall, the test demonstrated moderate effectiveness in evaluating students' English proficiency, but several items were identified for revision or improvement.

The overall results indicated moderate effectiveness of the test in assessing students' English proficiency. The average score obtained by the participants was 51.33, with scores ranging from 16.67 to 93.33. Only 35% of the students (7 out of 20) achieved the minimum mastery criterion of 60, while the remaining 65% (13 students) scored below this threshold. These findings suggest that the test posed a moderate level of challenge to the test-takers and necessitated a closer examination of item-level performance to ensure the assessment's reliability and diagnostic value. Here are the table of discrimination power and difficulty level of the questions.

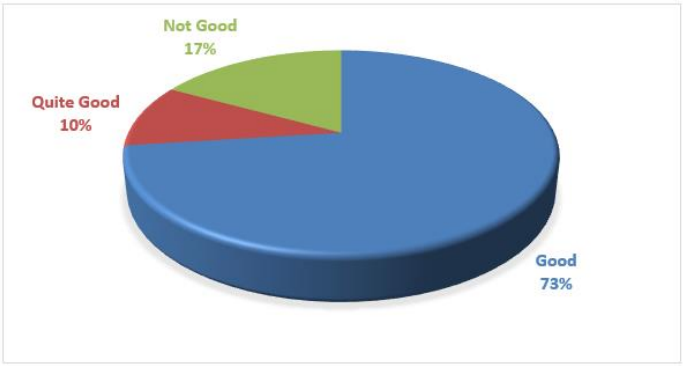
Table 2. Table of Discrimination Power and Difficulty Level

Item	Discrimination Power		Difficulty Level		Alternatives answers are ineffective	Information
	Coefficient	Information	Coefficient	Information		
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	0.260	Quite Good	0.900	Easy	CE	Distractor Revision
2	0.742	Good	0.800	Easy	C	Distractor Revision
3	0.382	Good	0.500	Medium	-	Good
4	0.114	Not Good	0.600	Medium	-	Not Good
5	0.499	Good	0.550	Medium	C	Distractor Revision
6	-0.173	Not Good	0.450	Medium	D	Not Good
7	0.118	Not Good	0.550	Medium	BE	Good
8	0.279	Quite Good	0.700	Medium	E	Distractor Revision
9	0.552	Good	0.350	Medium	-	Good
10	0.516	Good	0.550	Medium	D	Distractor Revision
11	0.433	Good	0.500	Medium	DE	Distractor Revision
12	0.309	Good	0.650	Medium	A	Distractor Revision
13	0.313	Good	0.500	Medium	DE	Distractor Revision
14	-0.287	Not Good	0.400	Medium	-	Not Good
15	0.298	Quite Good	0.350	Medium	-	Good
16	0.485	Good	0.500	Medium	D	Distractor Revision
17	0.607	Good	0.600	Medium	E	Distractor Revision
18	0.625	Good	0.350	Medium	-	Good
19	0.317	Good	0.400	Medium	-	Good
20	0.625	Good	0.550	Medium	-	Good
21	0.489	Good	0.750	Difficult	D	Distractor Revision
22	0.482	Good	0.600	Medium	CE	Distractor Revision
23	0.497	Good	0.350	Medium	-	Good
24	0.629	Good	0.250	Difficult	E	Distractor Revision
25	0.370	Good	0.350	Medium	-	Good
26	0.571	Good	0.600	Medium	D	Distractor Revision
27	0.054	Not Good	0.250	Difficult	E	Not Good
28	0.673	Good	0.650	Medium	E	Distractor Revision
29	0.546	Good	0.400	Medium	-	Good
30	0.583	Good	0.450	Medium	-	Good



Graph. 1 The Level of Difficulty Questions

In response to the first research question concerning the distribution of item difficulty and discrimination indices, the analysis revealed that 83% of the items were classified as having medium difficulty, with difficulty indices ranging from 0.30 to 0.70. Meanwhile, 10% of the items were categorized as easy (difficulty index > 0.70), and only 7% fell into the difficult category (difficulty index < 0.30). Although the distribution indicates a balanced test in terms of challenge, the low proportion of difficult items may reduce the test's capacity to accurately assess higher-level learners.

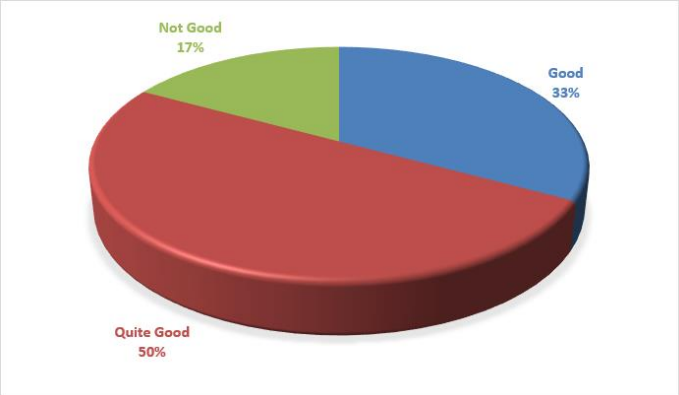


Graph. 2 The Power of Discrimination Questions

In terms of discrimination power, 73% of the items demonstrated good discrimination (index > 0.30), indicating that these items were effective in differentiating between high- and low-achieving students. Another 10% of items showed quite good discrimination (index between 0.20 and 0.29), while 17% were identified as having poor or negative discrimination. Among the problematic items were item 6 (discrimination index = -0.173), item 14 (index = -0.287), and item 27 (index = 0.054). These items failed to distinguish students' ability levels effectively and may indicate issues such as ambiguous phrasing, misaligned content, or incorrect answer keys. Interestingly, some items, such as item 2 were classified as easy (difficulty index = 0.800) yet showed a high discrimination index (0.742), indicating that even basic items can perform well in distinguishing learners when constructed appropriately.

Addressing the second research question regarding item revision based on distractor performance and item data, the findings showed that more than half of the test items contained one or more non-functional distractors options that were not selected by any student. Notable examples include item 1 (distractors C and E), item 2 (C), item 5 (C), item 6 (D), and item 11 (D and E). Such distractors are considered ineffective as they fail to attract students who have not mastered the material, thus weakening the item's ability to function diagnostically. Items that combined ineffective distractors with poor discrimination, such as item 6 and item 27, are particularly concerning and require revision. On the other hand, items such as item 3 (discrimination index = 0.382, difficulty index = 0.500) and item 30

(discrimination index = 0.583, difficulty index = 0.450) performed well across all parameters, suggesting that these items are both appropriately challenging and psychometrically valid.



Graph 3. The Quality of Questions

Visual representations generated by AnBuso further supported the findings. The difficulty level distribution graph confirmed the predominance of medium-difficulty items, while the discrimination index chart showed the concentration of items within the acceptable range, with a few outliers indicating low performance. The item quality classification graphic revealed that 33% of the items were categorized as "good," 50% as "sufficient," and the remaining 17% as "not good." These metrics provide clear insights into the overall quality of the test and highlight specific items for revision.

Pedagogically, these findings underscore the importance of conducting item analysis to enhance test quality. The presence of non-functional distractors and items with poor discrimination suggests that while the test design was generally sound, further refinement is necessary to ensure all items contribute meaningfully to student assessment. Poorly performing items should be revised to improve distractor plausibility, correct potential errors in answer keys, and adjust item wording for clarity. Additionally, increasing the proportion of higher-order or difficult items would allow for better differentiation among advanced learners.

Beyond summative evaluation, the results of this analysis also informed the development of individualized and classical remedial materials for students who did not meet the passing threshold. These materials focused on vocabulary development, grammatical accuracy (particularly verb tenses), and reading comprehension areas reflected in the content domains of the test. Thus, the item analysis not only supported the improvement of the test as a measurement tool but also served formative functions by guiding targeted instructional support.

Conclusion

This study investigated the quality of a teacher-developed multiple-choice English language test by conducting item analysis using the AnBuso platform. Through statistical evaluation of difficulty indices, discrimination indices, and distractor effectiveness, the study aimed to determine how well the test items functioned in assessing students' vocabulary, grammar, and reading comprehension skills. The analysis revealed that while the majority of test items demonstrated acceptable levels of difficulty and discrimination, a significant portion included non-functional distractors and a small number exhibited poor or negative discrimination, suggesting the need for targeted revision.

Specifically, 83% of the items fell within the medium difficulty range, indicating balanced overall item challenge. Additionally, 73% of the items showed good discrimination power, supporting the validity of the test in distinguishing between high- and low-performing students. However, 17% of the

items were categorized as poor, including items with ineffective distractors or negative discrimination values. These findings suggest that while the test was generally reliable, item-level revisions are essential to improve the quality and fairness of the assessment.

The use of AnBuso provided not only efficient and objective statistical analysis but also actionable insights that informed item revision and instructional planning. As a result of this analysis, specific items were flagged for reconstruction, and individualized and classical remedial materials were developed to address the observed learning gaps among students.

References

Darmawan, M., Sudarsono, Riyanti, D., Yuliana, Y. G. S., & Sumarni. (2022). A test-items analysis of English teacher-made test. *Journal of English Education and Teaching*, *6*, 498–512.

Freeman, F. S. (1955). *Theory and practice of psychological testing*.

Hughes, A. (2002). *Testing for Language Teachers*. https://doi.org/10.1017/CBO9780511732980.

Karim, S. A., Sudiro, S., & Sakinah, S. (2021). Utilizing test items analysis to examine the level of difficulty and discriminating power in a teacher-made test. *EduLite: Journal of English Education, Literature and Culture*, *6*(2), 256. https://doi.org/10.30659/e.6.2.256-269.

Maulina, N., & Novirianthy, R. (2020). Item Analysis and Peer-Review Evaluation of Specific Health Problems and Applied Research Block Examination. *Jurnal Pendidikan Kedokteran Indonesia: The Indonesian Journal of Medical Education*, *9*(2), 131. https://doi.org/10.22146/jpki.49006.

Rame, G., & Kesi, A. K. (2023). *An analysis of test items in English subject final test of ten grade students at SMA Negeri 2 Kupang*. *2*(2), 50–55.

Remmers, H. H., Gage, N. L., & Rummel, J. F. (1965). *A practical introduction to measurement and evaluation*. Harper & Row.

Siri, A., & Freddano, M. (2011). The use of item analysis for the improvement of objective examinations. *Procedia - Social and Behavioral Sciences*, *29*, 188–197. https://doi.org/10.1016/j.sbspro.2011.11.224.

Swanson, D. B., Holtzman, K. Z., Allbee, K., & Clauser, B. E. (2006). Psychometric characteristics and response times for content-parallel extended-matching and one-best-answer items in relation to number of options. *Academic Medicine*, *81*(10 SUPPL.), 93–96. https://doi.org/10.1097/01.acm.0000236518.87708.9d.

Xu, Y., & Liu, Y. (2009). Teacher assessment knowledge and practice: a narrative inquiry of a Chinese College EFL teacher's experience. *TESOL QUARTERLY*, *43*. http://www.ascd.org/publications/educational-leadership/mar06/vol63/num06/Needed@-A-Dose-of-Assessment-Literacy.aspx%5Cnhttp://www.tandfonline.com/doi/abs/10.1080/08878730.2011.605048%5Cnhttp://www.tandfonline.com/doi/abs/10.1080/00405840802577536