Proceedings of The Third National Seminar on English Language Teaching (NELT)
Volume: 3, 2025
Publisher: English Education Department,
Faculty Tarbiyah and Teacher Training, UIN Alauddin Makassar

EVALUATING THE QUALITY OF PRE-SERVICE TEACHER'S TEST: AN ITEM ANALYSIS APPROACH

Nur Fadillah Nurchalis[1]*, Muhammad Aswad[2]
[1]*English Education Study Program, Graduate Program, Universitas Negeri Makassar*
[2]*English Education Study Program, Teacher Training Faculty, Universitas Sulawesi Barat*
*Email: nur.fadillah.nurchalis@student.unm.ac.id

Abstract

Assessment design is a critical competency for pre-service teachers, alongside pedagogical skills. This study investigates the quality of test items constructed by a pre-service teacher who has recently completed a Language Testing course in the sixth semester. There are 30 multiple-choice questions which consisted of 10 items each on reading comprehension, grammar, and vocabulary were analysed using descriptive statistical methods to determine item difficulty, discrimination, validity and reliability. The analysis revealed that while the prospective teachers demonstrated a foundational understanding of test construction principles, many items lacked the precision and balance necessary for high-quality assessment. These findings highlight the need for continued instructional support and practical training in test development to ensure future teachers are well-prepared to create valid and reliable assessment tools in the classroom.

Keywords: *English, Item Analysis, Pre-Service Teacher, Test*

INTRODUCTION

Teachers play a pivotal role in the success of education, as they are directly responsible for the quality of learning outcomes (Muhammadiah et al., 2022). According to Law No. 14 of 2005 on Teachers and Lecturers (Indonesia, 2005), teachers are professional educators whose primary duties include educating, teaching, guiding, directing, training, assessing, and evaluating students. In line with these responsibilities, pre-service teachers who are currently undergoing teacher training must also be adequately prepared to carry out these professional tasks.

To ensure comprehensive preparation, teacher training programs integrate not only teaching practicums but also courses in educational assessment. One of the most essential and complex aspects of teacher training for pre-service teachers is learning how to assess student learning effectively (Guevarra et al., 2024). Without proper assessment skills, pre-service teachers may be failed to accurately measure students' understanding, skills, and progress. This can lead to misleading judgments about student performance, resulting in either overestimating or underestimating their abilities.

Assessment is fundamental to determining whether the intended learning objectives have been achieved (Marsevani, 2022). Beyond measuring student performance, assessments provide teachers with valuable feedback on learners' progress and serve as a basis for planning subsequent instructional activities (Dejong et al., 2002). Moreover, assessments are not only beneficial to students, they also offer teachers the opportunity to reflect on their instructional effectiveness. By analyzing assessment results, teachers can identify strengths and weaknesses in their teaching methods and make informed adjustments to enhance their professional practice (Tosuncuoglu, 2018).

Among various assessment methods, testing remains a common and practical tool. Brown (2003) classifies tests into two categories: standardized tests and teacher-made tests. The latter are designed by teachers based on curriculum objectives and lesson plans, aiming to evaluate students' mastery of specific instructional content. Teacher-made tests are commonly used for daily assessments, formative evaluations, and summative examinations. Despite their less formal structure compared to standardized tests, teacher-made tests must still uphold high standards of quality to ensure valid and reliable measurements of student performance (Arikunto, 2010).

One of the most frequently used formats in teacher-made tests is the multiple-choice question (MCQ) (Marsevani, 2022). An MCQ typically consists of three components: the stem (question prompt), the key response (correct answer), and the distractors (incorrect alternatives) (Namdeo & Sahoo, 2016). To maintain the quality of these assessments, conducting item analysis is essential. Item analysis involves evaluating the quality of individual test items by examining their difficulty level, discriminatory power, and functionality. This process contributes to improving test validity and building a reliable question bank (Danuwijaya, 2018).

Although numerous studies have examined item analysis in teacher-made tests (e.g., Hakim & Irhamsyah, 2020; Aliah, 2020; Liando et al., 2021; Darmawan et al., 2022), there remains a limited focus on item analysis of tests created by pre-service teachers. As future educators, it is crucial that pre-service teachers develop the skills to construct quality assessment tools that accurately measure learning outcomes. Addressing this gap, the present study aims to conduct an item analysis of teacher-made multiple-choice tests developed by pre-service teachers.

METHOD

This research employed a descriptive quantitative approach to examine the quality of multiple-choice questions (MCQs) designed as a formative assessment tool. The MCQs were developed by a pre-service teacher who had completed the English Language Testing course in the sixth semester of the English Education Study Program at STAIN Majene. The purpose of the study was to evaluate how well these test items met accepted standards of quality in terms of their validity, reliability, difficulty level, and discriminatory power.

The test consisted of 30 multiple-choice items, which were evenly distributed across three language components: reading comprehension, vocabulary, and grammar. Each component included 10 items. The test was administered to 20 first-year students at SMP 1 Majene, and their responses served as the data for the item analysis.

To carry out the analysis, each correct answer was given a score of 1, while incorrect answers were scored 0. These scores were then compiled into a data table for statistical analysis. The item analysis was carried out to measure four aspects of test quality: item validity, reliability, item difficulty, and item discrimination.

Item validity was assessed using the product moment correlation technique. This statistical method measures the relationship between the item scores and the total test scores. An item was considered valid if the calculated correlation coefficient (r) was higher than the critical value in the r-table. If the value was lower, the item was classified as invalid (Miterianifa & Zein, 2016).

The reliability of the test was determined using the Cronbach's Alpha formula. A reliability coefficient of 0.70 or higher was considered acceptable, indicating that the items consistently measured the intended learning outcomes (Miterianifa & Zein, 2016).

The difficulty level of each item was analyzed to determine how easy or hard it was for students to answer. The analysis used a specific index scale as follow:

Table 1. Index scale of item difficulty

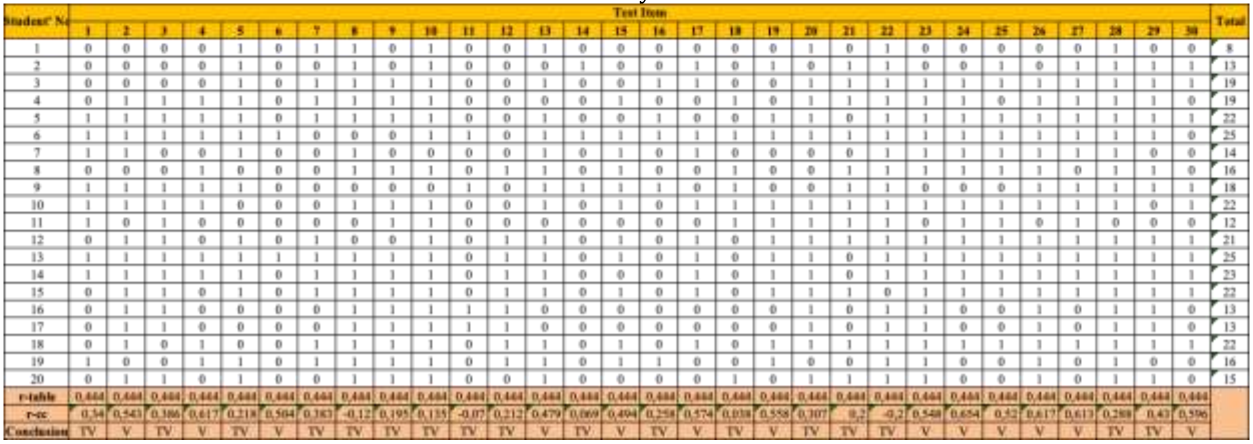
Index Value	Meaning	Code	Quality
0.81-1.00	Too Easy	TE	Ignored
0.61-0.80	Very Easy	VE	Fair
0.51-0.60	Easy	E	Good
0.5	Medium	M	Very good
0.40-0.49	Difficult	D	Fair
0.20-0.39	Very Difficult	VD	Good
0.00-0.19	Too Difficult	TD	Very good

The discrimination index was used to measure how well each item could distinguish between high-performing and low-performing students. An item was considered to have good discriminatory power if its index value was 0.30 or higher. Items with lower values were seen as less effective in distinguishing student ability.

RESULTS AND DISCUSSION

Validity

Table 2. Validity of the items



Student' No	Test Item																														Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
1	0	0	0	0	1	0	1	1	0	1	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	8
2	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	1	0	1	0	1	1	0	0	1	0	1	1	1	1	13
3	0	0	0	0	1	0	1	1	1	1	0	0	1	0	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	19
4	0	1	1	1	1	0	1	1	1	1	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	1	1	1	1	0	19
5	1	1	1	1	1	0	1	1	1	1	0	0	1	0	0	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	22
6	1	1	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	25
7	1	1	0	0	1	0	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0	1	1	1	1	1	1	1	0	0	14
8	0	0	0	1	0	0	0	1	1	1	0	1	1	0	1	0	0	1	0	0	1	1	1	1	1	1	0	1	1	0	16
9	1	1	1	1	1	0	0	0	0	0	1	0	1	1	1	1	0	1	0	0	1	1	0	0	0	1	1	1	1	1	18
10	1	1	1	1	0	0	0	1	1	1	0	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	22
11	1	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	1	1	0	1	0	0	0	12
12	0	1	1	0	1	0	1	0	0	1	0	1	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	21
13	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	0	1	0	1	1	0	1	1	1	1	1	1	1	1	1	25
14	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	0	1	0	1	1	0	1	1	1	1	1	1	1	1	1	23
15	0	1	1	0	1	0	1	1	1	1	0	1	1	0	1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1	22
16	0	1	1	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	1	0	13
17	0	1	1	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	1	0	13
18	0	1	0	1	0	0	1	1	1	1	0	1	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	22
19	1	0	0	1	1	0	1	1	1	1	0	1	1	0	1	1	0	0	1	0	0	1	1	0	0	1	0	1	0	0	16
20	0	1	1	0	1	0	0	1	1	1	0	0	1	0	0	0	0	1	0	1	1	1	1	0	0	1	0	1	1	0	15
r-table	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	
r-cc	0,34	0,543	0,386	0,617	0,218	0,504	0,383	-0,12	0,195	0,135	-0,07	0,212	0,479	0,069	0,494	0,258	0,574	0,038	0,558	0,307	0,2	-0,2	0,548	0,654	0,52	0,617	0,613	0,288	0,43	0,596	
Conclusion	TV	V	TV	V	TV	V	TV	TV	TV	TV	TV	TV	V	TV	V	TV	V	TV	V	TV	TV	TV	V	V	V	V	V	TV	TV	V	

The results presented in the table indicate that out of 30 multiple-choice test items, a total of 17 items were classified as valid, while the remaining 13 items were found to be invalid based on the item validity analysis. When examined across the three tested language components, reading, vocabulary, and grammar, distinct patterns emerged.

For the reading section (Items 1–10), only 3 items met the criteria for validity, indicating that a majority of the items in this section failed to effectively measure what they were intended to assess. This suggests that either the question construction or the alignment with learning objectives in this section may need significant revision.

In the vocabulary section (Items 11–20), 4 out of 10 items were valid, showing a slightly better quality than the reading section, but still falling short of the expected standard for a high-quality formative assessment. This result may reflect weaknesses in word choice, context clarity, or distractor effectiveness.

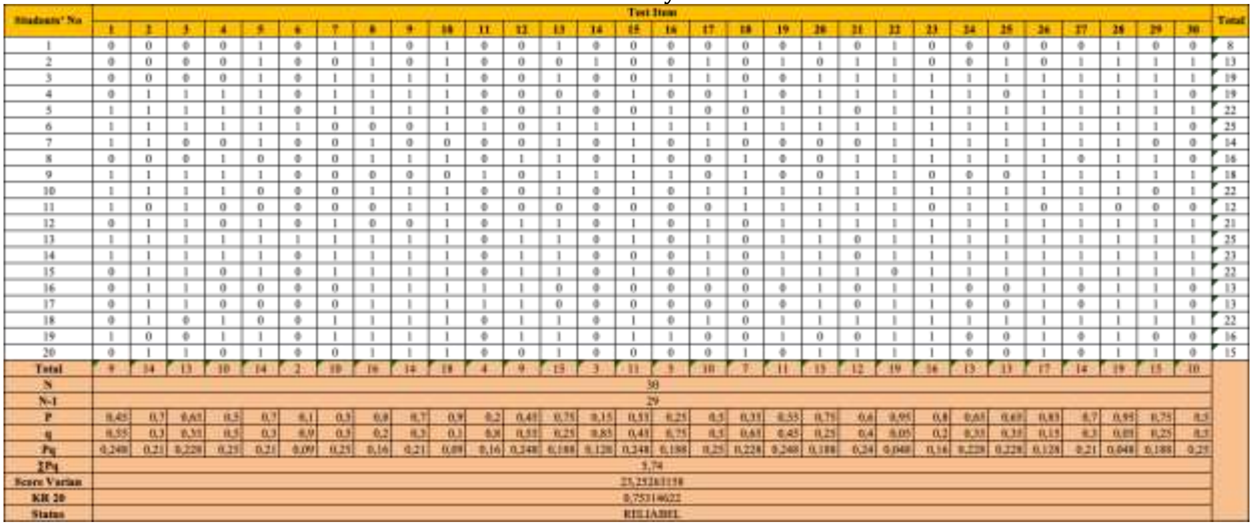
The grammar section (Items 21–30) showed the highest proportion of valid items, with 6 items meeting the validity criteria. This indicates that the grammar questions were relatively better constructed and more aligned with the testing objectives compared to the other two sections.

Overall, the analysis reveals that the grammar section exhibits stronger item quality in terms of validity than both the reading and vocabulary sections. This may suggest that the pre-service teacher had a better grasp of grammar item construction or that the grammar content was more concretely defined and easier to translate into well-structured multiple-choice questions. However, the relatively low number of valid items across all sections highlights the need for targeted training and feedback in test construction, particularly in crafting effective reading and vocabulary items.

According to Haladyna & Downing (2004), Poorly written or ambiguous items introduce construct-irrelevant variance, reducing item validity. Construct-irrelevant variance occurs when factors unrelated to the skill or knowledge being assessed influence test performance. In the context of this study, unclear wording, confusing distractors, or questions not aligned with the learning objectives may have caused students to answer incorrectly even if they had mastered the target material. For example, in the reading and vocabulary sections, where fewer valid items were found, the questions may have contained ambiguities in phrasing, misleading choices, or insufficient context, leading to students' misunderstanding of what was being asked.

Reliability

Table 3. Reliability of the test



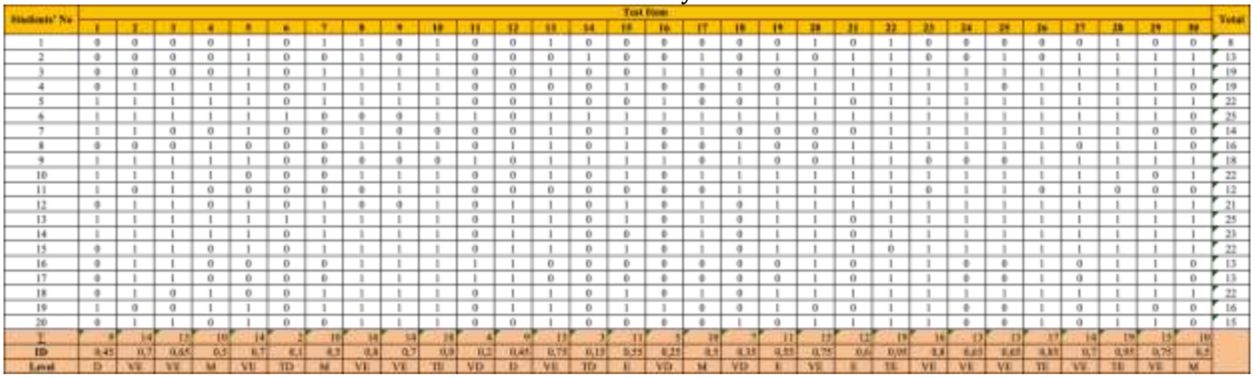
Students' No	Test Item																														Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
1	0	0	0	0	1	0	1	1	0	1	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	8
2	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	1	0	1	0	1	1	0	0	1	0	1	1	1	1	13
3	0	0	0	0	1	0	1	1	1	1	0	0	1	0	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	19
4	0	1	1	1	1	0	1	1	1	1	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	1	1	1	1	0	19
5	1	1	1	1	1	0	1	1	1	1	0	0	1	0	0	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	22
6	1	1	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	25
7	1	1	0	0	1	0	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0	1	1	1	1	1	1	1	0	0	14
8	0	0	0	1	0	0	0	1	1	1	0	1	1	0	1	0	0	1	0	0	1	1	1	1	1	1	0	1	1	0	16
9	1	1	1	1	1	0	0	0	0	0	1	0	1	1	1	1	0	1	0	0	1	1	0	0	0	1	1	1	1	1	18
10	1	1	1	1	0	0	0	1	1	1	0	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	22
11	1	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	1	1	0	1	0	0	0	12
12	0	1	1	0	1	0	1	0	0	1	0	1	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	21
13	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	0	1	0	1	1	0	1	1	1	1	1	1	1	1	1	25
14	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	0	1	0	1	1	0	1	1	1	1	1	1	1	1	1	23
15	0	1	1	0	1	0	1	1	1	1	0	1	1	0	1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1	22
16	0	1	1	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	1	0	13
17	0	1	1	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	1	0	13
18	0	1	0	1	0	0	1	1	1	1	0	1	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	22
19	1	0	0	1	1	0	1	1	1	1	0	1	1	0	1	1	0	0	1	0	0	1	1	0	0	1	0	1	0	0	16
20	0	1	1	0	1	0	0	1	1	1	0	0	1	0	0	0	0	1	0	1	1	1	1	0	0	1	0	1	1	0	15
Total	9	14	13	10	14	2	10	16	14	18	4	9	15	3	11	5	10	7	11	15	12	19	16	13	13	17	14	19	15	10	
N	30																														
N-1	29																														
P	0,45	0,7	0,65	0,5	0,7	0,1	0,5	0,8	0,7	0,9	0,2	0,45	0,75	0,15	0,55	0,25	0,5	0,35	0,55	0,75	0,6	0,95	0,8	0,65	0,65	0,85	0,7	0,95	0,75	0,5	
q	0,55	0,3	0,35	0,5	0,3	0,9	0,5	0,2	0,3	0,1	0,8	0,55	0,25	0,85	0,45	0,75	0,5	0,65	0,45	0,25	0,4	0,05	0,2	0,35	0,35	0,15	0,3	0,05	0,25	0,5	
Pq	0,248	0,21	0,228	0,25	0,21	0,09	0,25	0,16	0,21	0,09	0,16	0,248	0,188	0,128	0,248	0,188	0,25	0,228	0,248	0,188	0,24	0,048	0,16	0,228	0,228	0,128	0,21	0,048	0,188	0,25	
ΣPq	5,74																														
Score Varian	23,25263158																														
KR 20	0,75314622																														
Status	RELIABEL																														

The result of the reliability analysis shows that the test has a reliability coefficient of 0.83. This value indicates that the test is highly reliable. In other words, the items in the test consistently measure the intended skills across different students. A high reliability score means that if the same group of students were given the test again under similar conditions, their scores would likely be very similar (Arikunto, 2010). This consistency reflects the stability and dependability of the test as an instrument for assessing students' language abilities.

Although the test still contains some invalid items, the overall reliability remains strong. As noted by Anastasi and Urbina (1997), it is possible for a test to demonstrate high reliability even if some items do not accurately measure the intended construct. This indicates that, overall, the majority of the items function cohesively to assess students' reading, vocabulary, and grammar skills effectively. Furthermore, the high reliability coefficient suggests that the test can consistently provide dependable information about students' language performance, which is particularly useful for formative assessment purposes (Bachman, 1990). However, the existence of invalid items highlights the need for further refinement. While the scores remain stable, revising these problematic items would enhance the accuracy and validity of the test in measuring specific language components more precisely.

Item Difficulty

Table 4. Level difficulty of the items



Students' No	Test Item																														Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
1	0	0	0	0	1	0	1	1	0	1	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	8
2	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	1	0	1	0	1	1	0	0	1	0	1	1	1	1	13
3	0	0	0	0	1	0	1	1	1	1	0	0	1	0	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	19
4	0	1	1	1	1	0	1	1	1	1	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	1	1	1	1	0	19
5	1	1	1	1	1	0	1	1	1	1	0	0	1	0	0	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	22
6	1	1	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	25
7	1	1	0	0	1	0	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0	1	1	1	1	1	1	1	0	0	14
8	0	0	0	1	0	0	0	1	1	1	0	1	1	0	1	0	0	1	0	0	1	1	1	1	1	1	0	1	1	0	16
9	1	1	1	1	1	0	0	0	0	0	1	0	1	1	1	1	0	1	0	0	1	1	0	0	0	1	1	1	1	1	18
10	1	1	1	1	0	0	0	1	1	1	0	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	22
11	1	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	1	1	0	1	0	0	0	12
12	0	1	1	0	1	0	1	0	0	1	0	1	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	21
13	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	0	1	0	1	1	0	1	1	1	1	1	1	1	1	1	25
14	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	0	1	0	1	1	0	1	1	1	1	1	1	1	1	1	23
15	0	1	1	0	1	0	1	1	1	1	0	1	1	0	1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1	22
16	0	1	1	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	1	0	13
17	0	1	1	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	1	0	13
18	0	1	0	1	0	0	1	1	1	1	0	1	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	22
19	1	0	0	1	1	0	1	1	1	1	0	1	1	0	1	1	0	0	1	0	0	1	1	0	0	1	0	1	0	0	16
20	0	1	1	0	1	0	0	1	1	1	0	0	1	0	0	0	0	1	0	1	1	1	1	0	0	1	0	1	1	0	15
Σ	9	14	13	10	14	2	10	16	14	18	4	9	15	3	11	5	10	7	11	15	12	19	16	13	13	17	14	19	15	10	
ID	0,45	0,7	0,65	0,5	0,7	0,1	0,5	0,8	0,7	0,9	0,2	0,45	0,75	0,15	0,55	0,25	0,5	0,35	0,55	0,75	0,6	0,95	0,8	0,65	0,65	0,85	0,7	0,95	0,75	0,5	
Level	D	VE	VE	M	VE	TD	M	VE	VE	TE	VD	D	VE	TD	E	VD	M	VD	E	VE	E	TE	VE	VE	VE	TE	VE	TE	VE	M	

The analysis of item difficulty revealed that the test items varied in their levels of difficulty. Among the 30 test items, 4 items were classified as too easy, while 2 items were classified as too difficult. Items that fall into these extreme categories are generally considered less effective for accurately measuring student performance. The items that are too easy may not provide sufficient challenge for students and fail to distinguish between students with different levels of understanding. Conversely, the items that are too difficult may not reflect the students' level of mastery or may contain content that is beyond their current knowledge.

In the reading section, which includes items 1 to 10, there was one item categorized as too easy and one item categorized as too difficult. This suggests that while most reading items are appropriately challenging, a few items may need revision to better align with students' abilities.

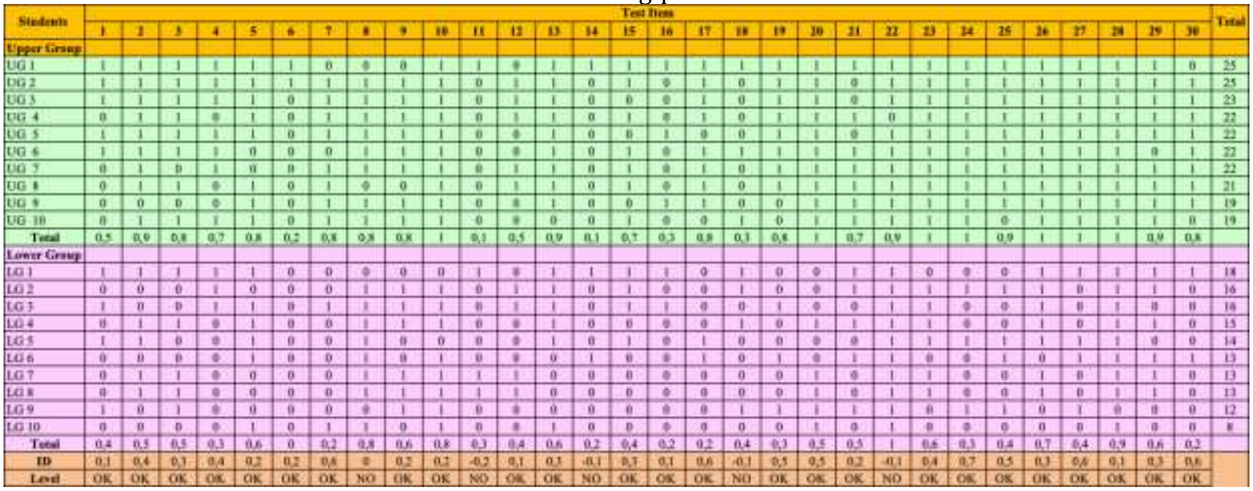
In the vocabulary section, covering items 11 to 20, one item was found to be too difficult. This indicates that some vocabulary used in the test may not be familiar to students or may require clearer context to be more accessible.

In the grammar section, consisting of items 21 to 30, three items were identified as too easy. This shows that certain grammar items may not adequately assess students' understanding and might need to be reconstructed to increase their level of challenge.

Overall, the majority of the test items fall within the acceptable range of difficulty, suggesting that the test is generally suitable for assessing students' language abilities. This balance in item difficulty allows for meaningful differentiation among students’ performance levels. However, the presence of several items in each section that are either too easy or too difficult indicates areas that require revision or replacement. As Ebel and Frisbie (1991) emphasize, items that are excessively easy or difficult provide little information about student differences, thereby diminishing the test’s overall validity. Refining these items will enhance the test’s capacity to more accurately measure the full spectrum of student proficiency.

Item Discrimination

Table 5. Discriminating power of the items



Students	Test Item																														Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
Upper Group																															
UG 1	1	1	1	1	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	25
UG 2	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	0	1	0	1	1	0	1	1	1	1	1	1	1	1	1	25
UG 3	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	0	1	0	1	1	0	1	1	1	1	1	1	1	1	1	23
UG 4	0	1	1	0	1	0	1	1	1	1	0	1	1	0	1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1	22
UG 5	1	1	1	1	1	0	1	1	1	1	0	0	1	0	0	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	22
UG 6	1	1	1	1	0	0	0	1	1	1	0	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	22
UG 7	0	1	0	1	0	0	1	1	1	1	0	1	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	22
UG 8	0	1	1	0	1	0	1	0	0	1	0	1	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	21
UG 9	0	0	0	0	1	0	1	1	1	1	0	0	1	0	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	19
UG 10	0	1	1	1	1	0	1	1	1	1	0	0	0	0	1	0	0	1	0	1	1	1	1	1	0	1	1	1	1	0	19
Total	0,5	0,9	0,8	0,7	0,8	0,2	0,8	0,8	0,8	1	0,1	0,5	0,9	0,1	0,7	0,3	0,8	0,3	0,8	1	0,7	0,9	1	1	0,9	1	1	1	0,9	0,8	
Lower Group																															
LG 1	1	1	1	1	1	0	0	0	0	0	1	0	1	1	1	1	0	1	0	0	1	1	0	0	0	1	1	1	1	1	18
LG 2	0	0	0	1	0	0	0	1	1	1	0	1	1	0	1	0	0	1	0	0	1	1	1	1	1	1	0	1	1	0	16
LG 3	1	0	0	1	1	0	1	1	1	1	0	1	1	0	1	1	0	0	1	0	0	1	1	0	0	1	0	1	0	0	16
LG 4	0	1	1	0	1	0	0	1	1	1	0	0	1	0	0	0	0	1	0	1	1	1	1	0	0	1	0	1	1	0	15
LG 5	1	1	0	0	1	0	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0	1	1	1	1	1	1	1	0	0	14
LG 6	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	1	0	1	0	1	1	0	0	1	0	1	1	1	1	13
LG 7	0	1	1	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	1	0	13
LG 8	0	1	1	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	1	1	0	13
LG 9	1	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	1	1	0	1	0	0	0	12
LG 10	0	0	0	0	1	0	1	1	0	1	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	8
Total	0,4	0,5	0,5	0,3	0,6	0	0,2	0,8	0,6	0,8	0,3	0,4	0,6	0,2	0,4	0,2	0,2	0,4	0,3	0,5	0,5	1	0,6	0,3	0,4	0,7	0,4	0,9	0,6	0,2	
ID	0,1	0,4	0,3	0,4	0,2	0,2	0,6	0	0,2	0,2	-0,2	0,1	0,3	-0,1	0,3	0,1	0,6	-0,1	0,5	0,5	0,2	-0,1	0,4	0,7	0,5	0,3	0,6	0,1	0,3	0,6	
Level	OK	OK	OK	OK	OK	OK	OK	NO	OK	OK	NO	OK	OK	NO	OK	OK	OK	NO	OK	OK	OK	NO	OK	OK	OK	OK	OK	OK	OK	OK	

The table above indicates that the majority of the test items exhibit good discriminating power, which reflects their effectiveness in differentiating between high-achieving and low-achieving students. This suggests that most items are functioning appropriately in assessing students with varying levels of proficiency. According to Gronlund and Waugh (2009), the discrimination index is a valuable indicator for evaluating how well a test item distinguishes between students of different ability levels. Therefore, the strong discrimination indices observed in this test contribute positively to its overall quality and ensure that the assessment provides meaningful information about student performance.

However, a small number of items were found to have poor discrimination. Specifically, in the reading test, 2 items did not meet the desired discrimination standard. Similarly, in the vocabulary test, there were also 2 items that failed to effectively differentiate student performance. In the grammar test, only 1 item was identified as having low discrimination power.

Overall, the findings suggest that while most of the items are valid for assessing student performance, a few items may require revision or replacement to enhance their ability to accurately discriminate among students with different proficiency levels.

Ensuring that all items have strong discrimination power is essential for improving the overall quality and fairness of the assessment.

CONCLUSION

This study aimed to evaluate the quality of multiple-choice test items developed by a pre-service teacher for assessing English language skills. The findings indicate that the test possesses moderate overall quality. Out of 30 items, 17 were valid, while 13 were invalid, suggesting that further refinement is necessary to improve the validity of several items. In terms of reliability, the test demonstrated satisfactory consistency, indicating its potential for dependable use. The analysis of item difficulty revealed a range of difficulty levels; most items fell into fair to very good categories, although some items were identified as too easy or too difficult, reducing the overall balance of the test. The item discrimination analysis showed that the majority of items effectively differentiated between high- and low-performing students, although a few items across reading, vocabulary, and grammar sections had weak discrimination power.

Despite these valuable insights, the study has certain limitations. The small sample size, limited to 20 students from a single school, may restrict the generalizability of the findings. Furthermore, since the test was developed by only one pre-service teacher, the results may not fully represent the broader population of pre-service teachers. Additionally, the scope of the test was limited to multiple-choice items assessing reading, vocabulary, and grammar, leaving out other important language skills such as listening, speaking, and writing.

Based on these findings, it is recommended that pre-service teachers receive more comprehensive training in test construction, with particular attention to developing valid, reliable, and discriminative items. Items identified as too easy, too difficult, or with weak discrimination should be reviewed and revised to enhance test quality. Future research should involve larger, more diverse samples and expand the scope of assessment to include various test formats and a broader range of language skills to provide a more comprehensive evaluation of assessment practices.

REFERENCES

Aliah, H. (2020). The Analysis of Junior High School Teacher-Made Tests for the Students in Enrekang. *FOSTER: Journal of English Language Teaching*, *1*(2), 122–138. https://doi.org/10.24256/foster-jelt.v1i2.14

Anastasi, A., & Urbina, S. (1997). Psychological testing, 7th ed. In *Psychological testing, 7th ed.* (p. xiii, 721-xiii, 721). Prentice Hall/Pearson Education.

Arikunto, S. (2010). *Dasar-dasar evaluasi pendidikan*. Bumi Aksara.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford University Press.

Brown, H. D. (2003). *Language assessment principle and classroom practice*. Pearson Education Inc.

Danuwijaya, A. A. (2018). Item analysis of reading comprehension test for post-graduate students. *English Review: Journal of English Education*, *7*(1 SE-Articles), 29–40.

https://doi.org/10.25134/erjee.v7i1.1493

Darmawan, M., -, S., Dwi Riyanti, Yohanes Gatot Sutapa Yuliana, & Sumarni. (2022). Test-Items Analysis of English Teacher-Made Test. *Journal of English Education and Teaching*, *6*(4), 498–513. https://doi.org/10.33369/jeet.6.4.498-513

Dejong, G., C. Lee, K., & and Kuntzleman, C. (2002). The Role of Assessment in Meeting the NASPE Physical Education Content Standards. *Journal of Physical Education, Recreation & Dance*, *73*(7), 22–25. https://doi.org/10.1080/07303084.2002.10607842

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational Measurement* (5th editio). Pearson College Div.

Gronlund, N. E., & Waugh, C. K. (2009). *Assessment of Student Achievement*. Pearson. https://books.google.co.id/books?id=g2x8QgAACAAJ

Guevarra, J. X. L., Bendicio, M. K. T., Aleroza, J. J., & Ilao, A. T. (2024). Assessment skills of pre-service teachers: Basis for the development of an assessment skills development guide. *International Research Journal of Education and Technology*, *06*(07), 21–31.

Hakim, L., & Irhamsyah, I. (2020). The analysis of the teacher-made test for senior high school at state senior high school 1 Kutacane, Aceh Tenggara. *JURNAL ILMIAH DIDAKTIKA: Media Ilmiah Pendidikan Dan Pengajaran*, *21*(1), 10. https://doi.org/10.22373/jid.v21i1.4120

Haladyna, T. M., & Downing, S. M. (2004). Construct-Irrelevant Variance in High-Stakes Testing. *Educational Measurement Issues and Practice*, *23*(1). https://doi.org/https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Indonesia, R. (2005). *UU RI No 14 Tahun 2005 tentang Guru dan Dosen*. DPR RI.

Liando, N., Serhalawan, E., & Wuntu, C. (2021). Analysis of Teacher-Made Tests Used in Summative Evaluation at SMP Negeri 1 Tompaso. *Jurnal Ilmiah Wahana Pendidikan*, *7*(8 SE-Full Articles). https://doi.org/10.5281/zenodo.5775342

Marsevani, M. (2022). Item Analysis of Multiple-Choice Questions: an Assessment of Young Learners. *English Review: Journal of English Education*, *10*(2), 401–408. https://doi.org/10.25134/erjee.v10i2.6241

Miterianifa, & Zein, M. (2016). *Evaluasi Pembelajaran Kimia*. Cahaya Firdaus.

Muhammadiah, M., Hamsiah, A., Muzakki, A., Nuramila, N., & Fauzi, Z. A. (2022). The Role of the Professional Teacher as the Agent of Change for Students. *AL-ISHLAH: Jurnal Pendidikan*, *14*(4), 6887–6896. https://doi.org/10.35445/alishlah.v14i4.1372

Namdeo, S., & Sahoo, B. (2016). Item analysis of multiple choice questions from an assessment of medical students in Bhubaneswar, India. *International Journal of Research in Medical Sciences*, *4*(5), 1716–1719. https://doi.org/10.18203/2320-6012.ijrms20161256

Tosuncuoglu, I. (2018). Importance of Assessment in ELT. *Journal of Education and Training Studies*, *6*(9), 163. https://doi.org/10.11114/jets.v6i9.3443