

AN ANALYSIS ON DIFFICULTY LEVEL OF ENGLISH SUMMATIVE TEST FOR THE SECOND GRADE STUDENTS AT MTSN GOWA

Andi Khairul Fajri¹
Muhammad Nur Akbar Rasyid²
Nur Aliyah Nur³

^{1,2,3}Universitas Islam Negeri Alauddin Makassar
E-mail: fajriandikhairul@gmail.com

ABSTRACT

This research aimed to finding out the level of difficulty faced by students in English exams, especially in summative test and to find out how teachers prepare and evaluate in making summative tests. This research utilized quantitative method on quantifying variables and establishing relationships among them through statistical analysis. The method employed primarily involve the use of questionnaires to know how the teacher implement the difficulty level of test items. The questionnaire utilized as a primary instrument to gather information about teachers preparation and evaluation regarding the difficulty level of the English summative tests. Additionally, student test results collected to analyze the actual performance and difficulty levels of the test items. The difficulty level for Class VIII.5 revealed that all test items fall within the moderate category, with difficulty indices (P) ranging between 0.34 and 0.69. Meanwhile difficulty level for Class VIII.6 revealed that all test ranging between 0.44 and 0.66., this range was considered moderate too. Next, both English teachers demonstrated a high level of awareness and best practices in the implementation of summative tests. Their responses confirmed that they not only prepare assessments thoughtfully but also evaluate their effectiveness critically, aligning with core principles of assessment literacy. Based on the questionnaire results, there was a discrepancy between what teachers believed or perceived about their test making practices and the actual test results, which showed a lack of variation in test difficulty levels. Teachers felt they had designed the test items with consideration of difficulty and quality, and aligned them with assessment principles. The teacher must reevaluate the level of difficulty of the questions because the questions should vary in level of difficulty. However, the actual test results did not indicate any easy or difficult test items.

Keywords: Analysis, The difficulty level, English summative test.

1) INTRODUCTION

English is a widespread language due to the fact it is used by most nations within the world. According to Dyvik (2024), there were about 1.5 billion people around the world who spoke English either natively or second language. Apart from that, English is an international language that is important to learn and master. The development of education in Indonesia continues to progress rapidly, including in the field of English language education. One

Paper presented at The 5th ICONELE on April 16th-17th, 2026 English Education
Department, Faculty of Tarbiyah and Teacher Training
Universitas Islam Negeri Alauddin Makassar
South Sulawesi-Indonesia

of the important aspects of English material is the evaluation of students' abilities through summative test.

Summative test are one of the evaluation tools used to evaluate students' understanding the material that has been taught during a certain period. According to Karlina (2021), summative test is one of the types of the achievement test. Although summative tests have many advantages, there are also some criticisms of their use. Summative tests tend to focus on the result and often does not provide enough details about the student's learning process. According to Widyaningsih (2013), the creation of summative tests carried out by teachers has several obstacles, namely in each semester the material studied is too much so that it cannot be realized in the form of limited questions. This forces students to focus on the final grade rather than the actual learning process.

There are several previous research that have examined the summative test. The first researcher is Muspira (2016) with the title "Item Analysis of English Summative Test for Second Grade Student of MAN 1 Tanete Bulukumba", the results of the research showed that the summative test in MAN 1 Tanete is reliable and the level of difficulty shows that there were various items with different levels of difficulty. The next researcher is Fikran (2024) entitled "The Analysis of Index Facility, Item Discrimination, and Distractor Efficiency of English Summative Teacher-Mid Test (TMT) At MTsN Barru", the results of Fikran's research indicate the teacher to conduct, analyze and review their tests before giving to students. This can have an impact on the accuracy of the assessment, if the items are too easy, the test is not able to distinguish between students with different levels of understanding, while if it is too difficult, the results can be unfair and do not reflect the actual abilities of students (Inawati et al., 2006).

Therefore, an analysis of the difficulty level of the test is a must to ensure that the questions used are valid and dependable. By understanding the level of difficulty, teacher can assess the measure to which students are mastering the material being taught and make adjustments in the learning method or curriculum if necessary. Based on the previous background, some problems need to be answered from this research are what is the difficulty level of English summative test item for the second-grade students of MTsN Gowa? and How does the English teachers implement the summative test for the second-grade students at MTsN Gowa?

Language Testing

Language is divided into four macro skills, which is; listening, reading, writing and speaking. Reading is a key skill that needs to be developed and improved continuously. The government also pays special attention to this skill as evidenced by the National Literacy Movement since 2016. This government program is based on the Minister of Education and Culture Regulation No. 23/2015 which is intended to foster character (Sepniwati 2022).

According to Sepniwati (2022), English literacy skills can be measured and improved through reading comprehension competency tests. One of them is TOEFL (Test of English as a Foreign Language). TOEFL is a test used in many countries to measure English language skills, especially in the field of education. Related to reading comprehension, there are five types of questions that are tested, namely questions about main idea of the passage, directly answered questions, indirectly answered questions, vocabulary questions, and overall review questions. Each type of question has its own characteristics, types of language tests is divided by two, such as Formative test and Summative Test.

Item Analysis

Item analysis is a statistical technique used to evaluate the quality of individual items on an exam. Using valid data on a test can have a maximum influence on student performance in evaluation activities. The data helps to assess how good each item discriminates between high-scoring and low-scoring individuals. So, the teacher or test maker needs to identify the quality of the test by analyzing each item on the test, so that the test can be classified as a good test. Taylor & Wier (2008) stated that teachers have to ensure that a test results scores which are an accurate reflection of an examinee's ability in a specific material.

With the method of analyzing items in the test given by students to evaluate students' abilities, it can provide feedback to teachers in improving the quality of student knowledge and teachers can review points that need to be followed up. That way, teachers can obtain data that can be processed to become learning revision material that needs to be developed. So that the creation of the next question can be maximized properly.

Level of Difficulty on Summative Test

The level of difficulty in item analysis have many other titles, such as item easiness, item facility, item difficulty, facility value or p-value. According to Hartati & Yogi (2019), analysis difficulty level or Facility Value (FV) is the first stage in analyzing test items. Heaton (1998) stated that the FV of an item shows the item difficulty. That way, FV can show how easy or difficult the test is by seeing what the ratio of presentation of student answers is correct.

A good test is a test that has a level of difficulty, include easy, moderate, and difficult. Using FV in the test can help teachers in identifying items in the test which include easy to difficult. Sumarsono & Kadir (2020) assumed that a good test should have 1:2:1 for the easy, medium, and difficult levels of the test. So, a test should have 20% of the easy level, 50% of the moderate level, and 25% of the difficulty level. In fact, a good test question in a summative test must have a certain level of difficulty, not too difficult and not too easy. If the test questions are answered correctly by many students, it means that the test questions are easy and vice versa. Most importantly, by analyzing the test questions can be reviewed from the level of difficulty of the items, this way, the teacher can know the nature of the test whether the questions have an easy, medium and difficult level.

2) METHODS

This study utilized quantitative research which focused on quantifying variables and establishing relationships among them through statistical analysis. In this case, the research is designed to obtain relevant and precise information regarding the difficulty level of English summative tests at the eighth grade of MTsN Gowa. The method employed primarily involve the use of questionnaires to know how the teacher implement the difficulty level of test items, enabling the researcher to collect data that can be analyzed statistically to elicit conclusions regarding the test's difficulties. According to Gay (1991), descriptive research also used to compare how a group of men and women are about a certain issue or topic. The method used is quantitative to measure the difficulty level of the question by measuring it with several formulas. A general field consisting of subjects/objects with certain characteristics and qualities determined by the researcher to be studied and then drawn conclusions, said Sukmawati et al., (2023). The researcher has selected eighth grade students of MTsN Gowa. The number of population are 185 students and divided into 6 classes. Beside, there was 2 English teachers involved in this research. Subset of the

population that represent the characteristics of the population. In this study, the sample consist of two classes and two English teachers, which is from class VIII.5 with 32 students and class VIII.6 with 32 students. Cluster random sampling was used in this study, a technique used when the research object or data source is very broad, such as the population in an area or a large group. According to Sugiono (2013), this method involves selecting samples based on groups or clusters, not individuals directly. To collect data, the researcher utilized a questionnaire as the primary instrument. This questionnaire designed to gather information about the teachers preparation and evaluation regarding the difficulty level of the English summative tests. Additionally, documentations also collected in this research in the form of student test results to analyze the actual performance and difficulty levels of the test items.

The questionnaire included a structured questions that allow for quantitative analysis, providing valuable insights into how teachers prepare the material and perform on the evaluation. The questionnaire was based on Brown & Abeywickrama (2019), which emphasizes the importance of test construction, evaluation, and item analysis in language assessment. According to Brown & Abeywickrama (2019), an effective summative test should be Well-prepared – aligned with learning objectives, Evaluated for quality – analyzed in terms of difficulty level and Balanced in difficulty level – ensuring a mix of easy, medium, and difficult questions.

The researcher collects data using documentation and questionnaires as primary research instruments. The researcher gathers written documents related to the summative test, including Test Question Papers – containing English summative test items for the 8th grade, Answer Sheets – students' responses to the test questions and Answer Key – the correct answers used for scoring. These documents are used to analyze the difficulty level of test items using Brown & Abeywickrama's (2019) formula for item difficulty. The researcher administers a closed-ended questionnaire to: Two English teachers who prepared and evaluated the summative test, The questionnaire focuses on test preparation and evaluation based on principles of effective language assessment and Teachers respond using a Likert scale (1-5) to provide structured, measurable data.

To analyze the data, the researcher use quantitative technique to determine the difficulty level of English summative test, researcher applies the following formula from Brown & Abeywickrama (2019):

$$P = \frac{R}{T}$$

Where:

P = Difficulty index (proportion of students who answered correctly)

R = Number of students who answered correctly

T = Total number of students who answer the items

Table 3.1 The difficulty level classification

P	Classification
0.00 – 0.30	Difficult
0.31 - 0.70	Medium
0.71 - 1.00	Easy

Descriptive analysis is used to determine the trends of responses on each item. This helps identify the general perception of teachers regarding test preparation and evaluation.

Table 3.2 Likert Scale Categories

Scale	Score	Interpretation
Strongly Agree	5	Very Positive Perception
Agree	4	Positive Perception
Neutral	3	Neutral Perception
Disagree	2	Negative Perception
Strongly Disagree	1	Very Negative Perception

The questionnaire instrument was developed by referring to theoretical concepts of language assessment proposed by Brown & Abeywickrama (2019) and Nitko & Brookhart (2014). The instrument consists of two main sections representing test preparation and evaluation stages. Each item was designed to reflect specific indicators aligned with the research objectives. The instrument's accuracy and appropriateness were ensured through theoretical alignment and careful mapping of items to relevant constructs. This approach is acceptable for small-scale research where expert-based or theory-based instrument validation is considered appropriate.

3) RESULTS

The Difficulty Level of English Summative Test Item for the Second Grade Students of MTsN Gowa.

The researcher has collected the results of summative test from two classes, then analyzed the right and wrong answer from all items and categorized it based on the difficulty level formula. Results of the difficulty level showed below:

Table 4.1 Difficulty Level Analysis Results in Class VIII.5

Item Number	Correct Responses	Total Students	Difficulty Index (P)	Classification
1	16	32	0.5	Moderate
2	19	32	0.59	Moderate
3	15	32	0.47	Moderate
4	13	32	0.41	Moderate
5	12	32	0.38	Moderate
6	19	32	0.59	Moderate
7	16	32	0.5	Moderate
8	21	32	0.66	Moderate
9	18	32	0.56	Moderate
10	12	32	0.38	Moderate
11	16	32	0.5	Moderate
12	12	32	0.38	Moderate
13	13	32	0.41	Moderate
14	18	32	0.56	Moderate
15	13	32	0.41	Moderate
16	13	32	0.41	Moderate
17	21	32	0.66	Moderate
18	16	32	0.5	Moderate
19	21	32	0.66	Moderate
20	18	32	0.56	Moderate
21	15	32	0.47	Moderate
22	20	32	0.62	Moderate
23	14	32	0.44	Moderate
24	13	32	0.41	Moderate
25	17	32	0.53	Moderate
26	13	32	0.41	Moderate
27	14	32	0.44	Moderate
28	18	32	0.56	Moderate
29	16	32	0.5	Moderate
30	17	32	0.53	Moderate
31	22	32	0.69	Moderate
32	14	32	0.44	Moderate
33	20	32	0.62	Moderate
34	15	32	0.47	Moderate
35	17	32	0.53	Moderate
36	13	32	0.41	Moderate
37	16	32	0.5	Moderate
38	18	32	0.56	Moderate
39	18	32	0.56	Moderate
40	17	32	0.53	Moderate

The difficulty level analysis of the English summative test for Class VIII.5 reveals that all 40 test items fall within the moderate difficulty category, with difficulty indices (P) ranging between 0.34 and 0.69. According to Brown & Abeywickrama (2019), items within the 0.31–0.70 range are considered moderately difficult. This uniformity indicates that the test is generally well-calibrated and accessible to most students, yet it does not offer enough variation to effectively differentiate between higher and lower ability learners.

Table 4.2 Difficulty Level Analysis Results in Class VIII.6

Item Number	Correct Responses	Total Students	Difficulty Index (P)	Classification
1	14	32	0.44	Moderate
2	19	32	0.59	Moderate
3	18	32	0.56	Moderate
4	11	32	0.34	Moderate
5	15	32	0.47	Moderate
6	18	32	0.56	Moderate
7	19	32	0.59	Moderate
8	18	32	0.56	Moderate
9	16	32	0.5	Moderate
10	13	32	0.41	Moderate
11	17	32	0.53	Moderate
12	15	32	0.47	Moderate
13	14	32	0.44	Moderate
14	20	32	0.62	Moderate
15	15	32	0.47	Moderate
16	16	32	0.5	Moderate
17	21	32	0.66	Moderate
18	17	32	0.53	Moderate
19	21	32	0.66	Moderate
20	16	32	0.5	Moderate
21	16	32	0.5	Moderate
22	19	32	0.59	Moderate
23	15	32	0.47	Moderate
24	14	32	0.44	Moderate
25	20	32	0.62	Moderate
26	15	32	0.47	Moderate
27	18	32	0.56	Moderate
28	19	32	0.59	Moderate
29	16	32	0.5	Moderate
30	15	32	0.47	Moderate
31	21	32	0.66	Moderate
32	15	32	0.47	Moderate
33	19	32	0.59	Moderate
34	14	32	0.44	Moderate
35	19	32	0.59	Moderate
36	17	32	0.53	Moderate

37	16	32	0.5	Moderate
38	20	32	0.62	Moderate
39	18	32	0.56	Moderate
40	18	32	0.56	Moderate

The difficulty level analysis of the English summative test for Class VIII.6 reveals that all 40 test items fall within the moderate difficulty category, with difficulty indices (P) ranging between 0.44 and 0.66. According to Brown & Abeywickrama (2019), items within the 0.31–0.70 range are considered moderately difficult. This uniformity indicates that the test is generally well-calibrated and accessible to most students, yet it does not offer enough variation to effectively differentiate between higher and lower ability learners.

Implementation of Summative Tests by English Teachers at MTsN Gowa

The questionnaire was divided into two parts: Preparation of Summative Test and Evaluation of Summative Test.

a. Preparation of Summative Test (Teacher's Approach to Test Development)

The first section, both teachers responded very positively to most of the statements, predominantly choosing scale 5 (Strongly Agree) and scale 4 (Agree).

1) Statement 1: "I align summative test items with the learning objectives of the curriculum."

Both teachers selected scale 5.

This indicates that they consistently ensure that their test items were constructed in accordance with the national curriculum goals. This practice aligns with Brown & Abeywickrama (2019), who emphasize that alignment between test content and instructional goals is crucial for test validity.

2) Statement 2: "I consider students' proficiency levels when designing summative test questions."

Both teachers selected scale 5.

This suggests a strong awareness of learner diversity and the need to adapt the level of challenge in the test to the students' capabilities.

3) Statement 3: "I ensure that summative tests contain a balance of easy, moderate, and difficult questions."

Teacher 1 chose scale 5; Teacher 2 chose scale 4.

While both teachers agree with this practice, the slight difference indicates that one teacher might not always be consistent in balancing question difficulty. This could relate to the actual test results, which showed all items were of moderate difficulty.

- 4) Statement 4: "I review past test performances before designing new test questions."

Teacher 1 selected scale 5; Teacher 2 selected scale 4.

This implies that both teachers usually refer to student performance data before constructing new assessments, although one may do it more systematically.

- 5) Statement 5: "I consult other teachers or use references when preparing summative test items."

Both teachers selected scale 5.

This reflects a strong culture of collaboration and continual learning. Teachers do not work in isolation but seek input and support from peers or educational resources. Overall, the responses in this section show that the teachers engage in deliberate, curriculum-aligned, and student-centered test preparation. Their practices are consistent with the principles of sound assessment as suggested by Nitko & Brookhart (2014).

b. Evaluation of Summative Test (Teacher's Review and Analysis of Test Effectiveness)

In the second section, both teachers again showed highly positive responses to all five statements, which focused on how they evaluate and reflect upon their summative assessments.

- 1) Statement 1: "I analyze students' test scores to identify the difficulty level of each question."

Both teachers selected scale 5.

This indicates that they conduct item-level analysis post-assessment, which is a core part of effective test evaluation.

- 2) Statement 2: "I reflect on whether test items fairly assess all students' abilities."

Both teachers selected scale 5.

This shows their commitment to equity and fairness in assessment design, ensuring all students have a fair opportunity to demonstrate learning.

- 3) Statement 3: "I identify questions that are too difficult or too easy."

Both teachers selected scale 4.

This suggests that while they attempt to identify such questions, there may be limitations in practice, perhaps due to time, training, or familiarity with item analysis techniques.

- 4) Statement 4: "I ensure that test items measure distinct aspects of English skills (listening, speaking, reading and writing)."

Both teachers selected scale 5.

This indicates a strong awareness of skill integration, ensuring that the test reflects the full scope of language learning.

- 5) Statement 5: "I use students' feedback or performance data to improve future summative tests."

Both teachers selected scale 5.

This reflects a reflective teaching mindset where feedback and performance results are utilized for continuous improvement. In summary, both English teachers demonstrated a consistently high level of awareness and best practices in the implementation of summative tests. Their responses confirm that they not only prepare assessments thoughtfully but also evaluate their effectiveness critically, aligning with core principles of assessment literacy.

4) DISCUSSION

The analysis of the difficulty level of the English summative test items for class VIII.5 and VIII.6 at MTsN Gowa revealed that all test items in both classes were classified as moderately difficult, with difficulty index values ranging from 0.34 to 0.69 in class VIII.5 and from 0.44 to 0.66 in class VIII.6. According to Brown & Abeywickrama (2019), a difficulty index (P) between 0.31 and 0.70 is categorized as moderate, suggesting that these test items are appropriate in terms of challenge and are neither too difficult nor too easy for the students. This classification indicates that most students are able to enroll with the test items meaningfully.

The findings imply that the summative tests were designed in a way that aligns with the students' general comprehension level. From a pedagogical perspective, this reflects a well-balanced test structure. Heaton (1998) emphasized that item difficulty provides essential information on whether the test questions match the ability of the test takers. Since none of the

questions fell under the “easy” ($P \geq 0.71$) or “difficult” ($P \leq 0.30$) categories, the test can be considered to have consistent difficulty distribution across items but lacks variation.

However, while having moderate difficulty across all items may appear balanced, it also raises a concern. A good quality test should ideally have a variety in item difficulty to cater to students of differing proficiency levels. Hartati & Yogi (2019) and Sumarsono & Kadir (2020) suggest that an ideal test should include 20% easy, 60% moderate, and 20% difficult items. This allows the test to not only assess the general understanding of the whole class but also to differentiate between high-achieving and low-achieving students, a principle aligned with effective summative assessment design.

In both classes, the absence of easy and difficult level of items means the test may not effectively serve its diagnostic function, which is to identify students who excel and those who require further support. As Brown & Abeywickrama (2019) argue, an effective test must include items that vary in difficulty to ensure comprehensive assessment coverage. When all items fall in the same difficulty band, it limits the ability of the test to serve as a valid measurement across different learner levels.

This uniformity in difficulty might be due to the teacher’s emphasis on aligning the test content closely with what has been taught during the semester, resulting in students having a uniform level of readiness. While this approach aligns with the principle of test authenticity (Brown, 2004), it may unintentionally lead to a lack of item variety. Thus, even though the test is valid and reflects the taught material, it could benefit from increased complexity in item construction.

In summary, the test administered to both classes can be considered moderately well-constructed in terms of difficulty level. It is suitable for evaluating overall student learning outcomes but lacks differentiation power. In order to boost next summative tests, it is suggested that teachers adopt a planned distribution of item difficulties and integrate item analysis methods in the test development process. This ensure that tests are not only aligned with instructional goals but are also reliable, valid, and equitable for students of diverse proficiency levels.

The findings show that both English teachers at MTsN Gowa demonstrated strong awareness and application of good assessment practices in both the preparation and evaluation of summative tests. Most of their responses in the questionnaire reflected agreement or strong

agreement with all the statements, suggesting that they possess a sound understanding of assessment principles.

In the preparation stage, both teachers reported that they align their test items with curriculum objectives. This is strongly aligned with Brown & Abeywickrama (2019), who state that alignment between instructional goals and assessment tasks is essential to ensure content validity. By constructing test items based on what has actually been taught, teachers enhance the authenticity and instructional value of the test. This finding also reinforces Nitko & Brookhart's (2014) perspective, which emphasizes that assessment should be an extension of curriculum and instruction rather than a separate or isolated activity.

Furthermore, both teachers claimed to consider students' proficiency levels when designing summative test questions. This aligns with Heaton's (1998) assertion that good assessment practice must accommodate the heterogeneity of learners and avoid a one-size-fits-all model. Heterogeneity is a concept that describes the diversity, dissimilarity, or variability between elements in a group, data, or research sample. Recognizing learners' varying levels of language ability ensures that test items are neither too easy nor too difficult, which contributes to the test fairness. However, this contrasts slightly with the actual test result findings in the difficulty level analysis, where all items were classified as "moderate", suggesting that while the teachers believe they balance difficulty levels, in practice, the variation may be limited.

One particularly noteworthy point is the teachers' response to the statement regarding balancing easy, moderate, and difficult questions. While both teachers agreed with this principle, one responded with "agree" (scale 4) rather than "strongly agree." This may reflect a partial or inconsistent application of this principle in real classroom practice. According to Hartati & Yogi (2019), an ideal summative test should contain 20% easy, 60% moderate, and 20% difficult items to ensure it can effectively differentiate students' performance levels. The moderate-only distribution found in the actual test implies a potential gap between perceived practice and real implementation.

In terms of evaluation practices, both teachers stated that they conduct item-level analysis after administering the test. This reflects an understanding of post-assessment reflection, which Nitko & Brookhart (2014) identify as a key component of assessment for learning. Analyzing which items are difficult, too easy, or ineffective helps teachers refine their test instruments over

time. This reflective process is essential in moving from assessment merely as a measurement tool to a mechanism for instructional improvement.

The teachers also reported that they consider the fairness of the items and whether those items assess a set of English skills (listening, speaking, reading and writing). This was in line with Brown & Abeywickrama's (2019) view of integrated skills assessment, which advocates that English language testing should not isolate individual language components but rather reflect how they interact in authentic communication. Their practice demonstrates a holistic understanding of language proficiency and the importance of construct validity.

Moreover, both teachers stated that they use student feedback and performance data to improve future assessments. This response highlights their formative mindset, wherein the outcomes of a summative test are not just final judgments but inform next steps in teaching. This approach aligns with Nitko & Brookhart's (2014) assertion that even summative assessments can serve formative functions if followed by appropriate reflection and action.

In conclusion, the implementation practices of the two English teachers are largely consistent with the theoretical foundations discussed in Chapter II. While their responses indicate high levels of awareness and intent to apply good assessment principles, some discrepancies appear when these perceptions are compared to the actual test results—particularly in terms of item difficulty balance. This suggests a need for greater coherence between perceived practice and actual application, which may be addressed through further training in practical item analysis and test construction techniques.

5) CONCLUSION

Based on findings and discussions presented in above chapters, researcher draws the conclusions: The difficulty level of the English summative test items used for 2nd students at MTsN Gowa was uniformly classified as moderate. The analysis of 40 test items in both Class VIII.5 and VIII.6 showed that all items had a difficulty index (P) ranging from 0.34 to 0.69, which falls within the moderate classification. While this suggests that the tests were accessible and aligned with students' general comprehension, the lack of easy and difficult items indicates insufficient variation, which may reduce the test's ability to differentiate between students of varying proficiency levels. The implementation of summative tests by the English teachers at MTsN Gowa showed very positive practices. Based on the questionnaire results, the teachers

demonstrated strong awareness and effort in aligning the test items with learning objectives, evaluating test quality, and considering item difficulty during both the preparation and evaluation stages. However, despite their stated intention to balance item difficulty, the actual test results revealed no presence of easy or difficult items. This indicates a potential gap between teachers' perceptions and actual test construction practices, suggesting the need for further capacity building in item analysis and test development.

This study highlights the importance of strengthening the assessment literacy of pre-service teachers. The English Education Department should integrate deeper practical training on test development, item analysis, and summative assessment design into the curriculum. This ensure that future teachers not only understand theoretical concepts but are also equipped with the skills needed to construct balanced, valid, and reliable assessments aligned with pedagogical goals. This study highlights the importance of strengthening the assessment literacy of pre-service teachers. The English Education Department should integrate deeper practical training on test development, item analysis, and summative assessment design into the curriculum. This ensure that future teachers not only understand theoretical concepts but are also equipped with the skills needed to construct balanced, valid, and reliable assessments aligned with pedagogical goals. This research provides as a reference for researchers interested in exploring test quality and assessment practices in secondary education. Future studies can expand on this research by including additional dimensions of item analysis such as discriminating power and distractor effectiveness, or by examining the impact of teacher training and support systems on assessment outcomes. Researchers are also encouraged to involve a larger sample size and more diverse educational settings to increase generalizability.

REFERENCES

- Azizah, K. P., Sahuddin, & Zamzam, A. (2023). Item Analysis on English Summative Test For The 11th Grade Students in MAN 2 Mataram The Academic Year 2022/2023. *Jeff (Journal of English Education Forum)*
- Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford University Press
- Barokah, M. (2019). Manajemen Penilaian Sumatif Pada Ranah Kognitif Pembelajaran PAI Kelas X Semester Ganjir di SMA Negeri 2 Pontianak Tahun Pelajaran 2017/ 2018. *Al-Idarah: Jurnal Kependidikan Islam*, 9(2), 160–179.
- Brown, H. D., & Abeywickrama, P. (2019). *Language Assessment: Principles and Classroom*

- Practices (3rd ed.). New York: Pearson Education
- Brown, H. D. (2004). *Language assessment: principles and classroom practices*. Pearson/Longman.
- Dyvik Einar H. (2024, June 21). *The Most Spoken Languages Worldwide 2023*. Statista.
- Edmatiaara, M. (2019). *The Influence of Memory Matrix on Classroom Assessment Technique (CATs) to Improve Students English Vocabulary at The Junior High School in Academic Year 2018/2019*.
- Gay, L. R. (1991). *Educational Research*.
- Hartati, N., & Yogi, H. P. S. (2019). Item Analysis for a Better Quality Test. *English Language in Focus*. <https://jurnal.umj.ac.id/index.php/ELIF>
- Heaton, J. B. (1998). *Writing English Language Tests* (Longman Handbooks for Language Teachers).
- Inawati, Agustina, A., Awalludin, & Lasmiatun. (2006). *Analisis Tingkat Kesulitan Butir Soal Mata Pelajaran Bahasa Indonesia Buatan Guru*.
- Karlina, R. (2021). *The Items Analysis of Summative Test at The Eight Grade of SMPN 12 Lebong in The Academic Year of 2020/2021*. State Institute For Islamic Studies Bengkulu.
- Larianti, E. (2019). *The Analysis of Final School Exam Items on English Subject at SMPN 14 Seluma in Academic Year 2018/2019*.
- Li, R. (2020). A Study on the Application of Oral English Test in English Foreign Language Classroom. *Asia-Pacific Conference on Engineering Education, Advanced Education and Training (EEAET 2020)*, 72–75. <https://doi.org/10.25236/eeat.2020.015>
- Miles, B. M., Huberman, M. A., & Saldana, J. (n.d.). *Qualitative Data Analysis* (H. Salmon, Ed.; 3rd ed.). Sage.
- Moneterisqi, H., Rohmana, & Fatmawati, W. ode. (2021). An Analysis of English Summative Test Used at Grade Eleven of Madrasah Aliyah (MA) Bina Cendekia Tampo. *JTE Journal of Teaching of English*, 6(3), 2548–6810.
- Nitko, A. J., & Brookhart, S. M. (2014). *Educational assessment of Students*.
- Risnita, Zurnelli, N., Fitria, W., & Makrub. (2019). An Analysis of English Summative Test Items of Twelfth Grade Student of Man 2 Merangin Academic Year 2019/2020. *Journal of Research in English Language Teaching JR-ELT*, 4(2), 81–90. <http://e-journal.ftk.uinjambi.ac.id>
- Rizky, P. L. (2018). *Theachers Chalengges in Holding Group Works in EFL Classes*.
- Semiun, T. T., Wisrance, M. W., & Napitupulu, M. H. (2022). English Summative Tests: The Quality of Its Items. *Journal of English Teaching and Research*, 7(2), 119–127.
- Sepniwati, L. (2022). *Kemampuan Memahami Bacaan Bahasa Inggris dalam English Proficiency Test di IAKN Palangka Raya*. 4, 3297–3302.
- Suek, L. A. (2021). Item Analysis of An English Summative Test. *Pattimura Excellence Journal of Language and Culture*, 1(1), 9–18. <https://ojs3.unpatti.ac.id/index.php/pejlac>

- Sugiono. (2013). Metode Penelitian Kuantitatif Kualitatif dan R&D.
- Sukmawati, Salmia, & Sudarmin. (2023). Population, Sample (Quantitative) and Selection of Participants/Key Informants (Qualitative). 7(1).
- Sumarsono, D., & Kadir Bagis, A. (2020). The Forecasting Power of Task-Based Language Teaching and Self-Efficacy on Students' Speaking Performance. *JOLLT Journal of Languages and Language Teaching*, 8(4), 412.
<https://doi.org/10.33394/jollt.v8i4.2848>
- Sutikno, Y. (2019). Pentingnya Penilaian Formatif Dosen Terhadap Mahasiswa Pendidikan Keagamaan Budha. 12, 45–53.
- Taylor, lynda, & Wier, C. J. (2008). Multilingualism and assessment. Cambridge University Press.
- Widyaningsih Nalarastu Wayan Ni. (2013). The Analysis of Sumative Test Made by Indonesian Language Teacher of Eleventh Grade of Science Class of Sma Laboratorium Undiksha Singaraja in Academic. <https://doi.org/10.23887/jjpbs.v1i8.1218>