# CHAPTER II

# LITERATURE REVIEW

## A. Review of Literature

### 1. Multiple Choice Items Analysis

Item analysis is a crucial process for evaluating the quality and effectiveness of test items in measuring student learning. This process enables educators to examine the performance of multiple-choice items in terms of their difficulty level, discriminating index, and the effectiveness of distractors. According to Kumar et al. (2021), analyzing test items is essential for determining their effectiveness in assessing student abilities, while Muslim Darmawan et al., (2022) further noted that item analysis provides insight into how each question performs across these key aspects. Based on these perspectives, item analysis conducted through the examination of difficulty level, discriminant index, and distractor effectiveness serves as a critical tool for refining assessments and ensuring that test items accurately reflect students' level of understanding.

First, the difficulty aspects. The Level of difficulty aspect of multiple-choice questions refers to how challenging a question is for students. A well-constructed multiple choice item should be balanced, not too easy and not too difficult, but offer a reasonable challenge to assess students' understanding and cognitive abilities. According to Arikunto (2013), a question that is said to be feasible or good is a question with an average level of difficulty. Decent questions include easy, medium and difficult phases in the difficulty level.

Questions with moderate difficulty are included in the category of effective and good questions. Questions that are too easy or too difficult can weaken the quality of the problem and cannot obtain valid information about student learning achievement data. Then Arikunto (2013) also explained about the criteria for the level of difficulty that was declared feasible, namely the range from 0.00 to 1.00, the detailed table shows:

Table 2. 1 The Difficulty Level

| Index Level | Category |
|---|---|
| 0.00 – 0.30 | Difficulty |
| 0.31- 0. 70 | Moderate |
| 0.71- 1.00 | Esay |

In grouping in the index level, multiple choice items are determined by calculating the proportion of students who have answered multiple choice items with the correct answer. According to Brown, (2004) the formula for calculating the item difficulty index is as follows:

$$DL = \frac{\text{total number answering the item correctly}}{\text{number of student quetion sheets}}$$

Second, the discrimination aspects. The discrimination index aspect is a measure of how effectively a test item distinguishes between students with high and low performance. According to Brown (2004), the discrimination index is a statistical measure used to determine how well a test item distinguishes between high-achieving and low-achieving students, typically categorized into upper and lower groups based on their overall test

performance. In calculating the discriminate index, by dividing it into two groups based on predetermined categories, the following formula is used to calculate the discriminant index using the Brown formula (2004):

ID = DL high group –DL low group

This categorical discrimination index is also reinforced by Arikunto (2016) categorizing multiple choice items that are low is worth 0.00 - 0.20 then, items that discuss the sufficient category 0.21 - 0.41 then items that discuss the good category 0.42 - 0.70 and items that discuss the excellent category 0.71 - 1.00. the detailed table shows:

Table 2. 2 Discrimination level

| Index | Kategori |
|---|---|
| 0,00 to 0,20 | Poor |
| 0.21 – 0.41 | Satisfactory |
| 0.42- 0.70 | Good |
| 0,71 – 1.00 | Excellent |

In calculating the discriminate index, by dividing it into two groups based on predetermined categories, the following formula is used to calculate the discriminant index using the Brown formula (2004):

ID = DL high group –DL low group

Third, the distractors aspects. Distractors, or wrong answer choices are an important part of multiple choice questions. According to Arikunto (2019) in determining the distractor in multiple choice is 5% of the choice of all

students who fill in the question with the aim of being able to assess how feasible it is from a question that is used.

Table 2. 3 The Distractor Effectiveness

| Criteria | Category |
|---|---|
| DE ≥ 5% | Effective |
| DE < 5% | Ineffective |

To strengthen the foundation of this research, it is important to refer to previous studies with similar focus. The study conducted by Rahmaini & Taufiq, (2018) aimed to evaluate the quality of 30 multiple-choice items used in the final semester examination for the Islamic Education and Character Building subject at SMK Negeri 1 Sedayu during the 2017–2018 academic year. The analysis revealed that the overall quality of the test items was low. In terms of difficulty level, 70% of the items were categorized as easy, 23.3% as moderate, and only 6.7% as difficult, indicating an imbalance in item distribution. Regarding discriminating power, 57% of the items were classified as poor, 23% as fair, 20% as good, and none as very good, suggesting a limited ability of the items to differentiate between high- and low-performing students. Additionally, the effectiveness of distractors was found to be weak, with 50% of distractors not functioning, 20% functioning poorly, 30% functioning adequately, and none functioning effectively or very effectively. These findings highlight the importance of constructing test items based on essential quality criteria, including balanced difficulty, strong

discriminating power, and effective distractors, to ensure accurate and meaningful assessment outcomes.

Analysis related to level difficulty in multiple-choice questions in the textbook of study questions has also been carried out. The study conducted by Yusnida & Ayouni, (2023) aimed to examine the difficulty index of a set of prediction test items from the *Ekadanta* tutoring book. Using a descriptive qualitative approach, the study involved 20 test items and 11 students from the English Department at Universitas Iskandarmuda, Banda Aceh. The students were instructed to complete the test within 40 minutes, and the responses were analyzed using a standard difficulty index formula. The results indicated that 8 items were classified as difficult, 8 as moderate, and 4 as easy. Based on these findings, the test was considered acceptable in terms of item difficulty. The researchers recommended that future studies include a larger sample of test items and participants to strengthen the generalizability of the results.

Other research related to multiple-choice analysis was also carried out on university-level questions. The study by Mahjabeen et al., (2017) analyzed a midterm examination consisting of 65 multiple-choice questions (MCQs) to evaluate item quality based on difficulty index (DIF), discrimination index (DI), and distractor efficiency (DE). The findings showed that 81% of the items fell within the acceptable range of difficulty, while 2% were too difficult and 17% too easy. In terms of discrimination, 62% of the MCQs demonstrated excellent ability to differentiate between high- and low-performing students,

while the rest ranged from good to poor. Of the 260 distractors analyzed, 72% were functional, and 28% were non-functional. Furthermore, 71% of the MCQs had either zero or one non-functional distractor, indicating high distractor efficiency. Notably, the only difficult item achieved 100% DE, compared to a much lower average of 36.33% in the easy items. Overall, the exam showed a high proportion of items with acceptable DIF and DI values, and the study concluded that item analysis is a valuable tool for developing standardized MCQs with appropriate difficulty, strong discriminatory power, and efficient distractors, ultimately enhancing exam validity and student assessment.

Research related to the analysis of multiple-choice questions was also conducted for junior high school English try out questions in Indonesia. The study conducted by Jannah et al., (2021)analyzed multiple-choice items from a trial test in a state junior high school in Indonesia to evaluate item quality based on difficulty level, discriminating power, and distractor efficiency. The findings revealed that the difficulty levels varied across the test: 10 items were difficult, 33 were moderately difficult, and 7 were easy, suggesting an overall balanced level of difficulty. In terms of discriminating power, 25 items demonstrated good or satisfactory discrimination, effectively measuring students' abilities. However, the remaining 25 items were categorized as poor or rejected due to vague wording, with 16 requiring revision and 9 recommended for removal. Regarding distractor efficiency, the study found that many distractors were ineffective, as they tended to mislead high-

achieving students more than lower-performing ones. This indicates a flaw in distractor construction that could compromise the test's ability to accurately assess student performance. Overall, the study underscores the need for careful item revision, particularly in wording clarity and distractor design, to enhance test validity.

Analysis on multiple-choice questions for other subject subjects has also been often carried out. The study by Komariah & Rufi (2023) aimed to enhance the quality of learning and assessment in Grade VII at SMP Negeri 6 Bangkalan by evaluating the suitability of teacher-made multiple-choice questions in mathematics. Using a quantitative approach and SPSS 25.0 for analysis, the study assessed item validity, reliability, difficulty level, and discriminating power. The results showed that all 20 items (100%) were valid, indicating a high level of content validity. The test also demonstrated strong reliability, with a reliability coefficient of R1 = 0.937, suggesting consistent and dependable assessment outcomes. In terms of difficulty, 75% of items were of moderate difficulty, 15% easy, and none were classified as difficult. Regarding discriminating power, all items (100%) were found to have good ability to distinguish between high- and low-performing students. Based on these findings, the researchers emphasized the importance of conducting thorough item analysis prior to test administration, covering validity, reliability, difficulty, and discrimination, in order to ensure accurate and meaningful assessment of student competencies.

The collective findings from the reviewed studies emphasize the importance of comprehensive item analysis in ensuring the quality and effectiveness of multiple-choice assessments. Overall, the test items examined across various educational settings demonstrated a wide range of difficulty levels, with some tests showing an imbalance dominated by easy or moderately difficult items while others achieved more acceptable distributions. In terms of discriminating power, a consistent pattern emerged in which a portion of items functioned effectively to distinguish between high- and low-performing students, whereas others were found to be poorly constructed, often due to vague wording or insufficient clarity, necessitating revision or removal. Distractor efficiency also proved to be a critical issue in many cases, distractors failed to operate as intended, misleading even high-achieving students and thereby reducing test validity. Nevertheless, several assessments showed high levels of item validity and reliability, indicating that with rigorous testing and refinement, teacher-constructed items can meet psychometric standards. These insights collectively reinforce the need for educators to routinely conduct item validation processes including analyses of validity, reliability, difficulty index, discrimination index, and distractor functionality to develop high-quality assessments that yield reliable, meaningful measures of student learning outcomes.

## 2. Multiple Choice Items Made by Teacher

Multiple-choice questions created by teachers, grounded in previous lessons, are essential tools for evaluating student learning. There are several

studies emphasize the need for these questions to align with learning goals and to meet standards of validity and reliability. According Brown (2004) highlights that MCQs are efficient and objective methods for measuring student achievement. Supporting by Haladyna (2022) points out that well-designed multiple-choice items not only gauge students' performance but also aid them in better preparing for exams. Overall, creating effective multiple-choice assessments demands a combination of thorough content understanding and expertise in test design principles to ensure precise and meaningful evaluation results.

Research related to the analysis of this multiple choice test has been conducted several times. The research conducted by Sene et al., (2022) analyzed 40 multiple-choice items and 160 distractors from a junior high school semester exam to evaluate the quality of distractors. The results showed that 14.375% of distractors were categorized as very poor, 28.125% as poor, 18.125% as fair, 20% as good, and only 19.375% as very good. This means that more than half of the distractors failed to meet acceptable standards, indicating a need for item revision to improve the overall quality and effectiveness of the test.

Research on the analysis of multiple-choice questions can also be conducted based on the 18 norms of making a good multiple-choice tests. Research by Santy et al., (2020) analyzed 100 multiple-choice items from mid-semester tests created by English teachers at a school in Singaraja using content analysis and checklist comparison against standard norms for quality

item construction. The results showed that the majority of items demonstrated very good quality, with only 1% rated as sufficient and 8% as good. Five norms were fully met, reflecting basic competencies, avoiding opinion-based items, correct spelling, eliminating double negatives, and avoiding absolute options. However, recurring issues were identified in the use of punctuation and capitalization. Overall, the findings indicate that teachers generally adhere to test construction norms, though the quality of the test alone may not directly determine student achievement outcomes.

Research on the analysis of multiple choice tests for English has also been conducted. Research by Septi aimed to examined 125 items from four teacher-made multiple-choice tests used as summative assessments in English at SMP N 4 Singaraja. Using document analysis and interviews, each item was evaluated against established norms to determine its quality. The findings revealed that 124 items (99%) were classified as very good, while only one item (1%) was categorized as good. Despite this overall indication of excellent quality, the study emphasized the importance of implementing minor revisions by addressing specific criteria that had not been fully satisfied, in order to enhance the effectiveness and validity of the test construction.

In addition to English, research related to multiple choice test analysis was also carried out in other subjects. Research by Syahroi et al., (2023) aims to evaluated the quality of 30 multiple-choice items used in History assessments for Grade XI at SMAN 2 Pandeglang by examining their

validity, reliability, difficulty level, discriminating power, and distractor effectiveness. The findings revealed that only 43.3% of the items were valid, while the remaining 56.7% were invalid. The reliability coefficient was 0.49, indicating low reliability of both the items and test results. In terms of discriminating power, the majority of items (66.7%) fell into the weak category, with only one item categorized as good, and none rated as very good. The difficulty index showed that 86.7% of the items were difficult, with no items categorized as very difficult, suggesting an overall high level of difficulty. Despite these weaknesses, all items met the criteria for effective distractors. Additionally, expert validation confirmed that 27 items aligned with the indicators, 29 followed scientific principles, and all had a single correct answer. These findings suggest that although the content and distractor structure were acceptable, the test requires significant revision to improve its validity, reliability, and overall discriminatory capacity.

Analysis of multiple choice test questions can also be done for the level of elementary school education questions. Research by Marsevani, (2022) aims to asses the quality of 10 English multiple-choice questions used in a public elementary school by analyzing their difficulty level, discrimination power, and distractor efficiency. The results indicated that 80% of the items had an acceptable difficulty level, while 20% were considered fairly easy. The discrimination index was strong, with 80% of the items categorized as excellent and the remaining 20% as good. Out of 40 distractors, 32 (80%) were functional, whereas 8 were identified as non-

functional. Despite two items being completely inefficient in terms of distractor performance, the overall findings suggest that the test items demonstrated good quality, particularly in their ability to differentiate student performance and maintain appropriate difficulty levels.

The synthesis of the reviewed studies highlights a comprehensive understanding of multiple-choice test quality across different educational levels and contexts. The majority of the studies emphasize the importance of ensuring that test items meet essential psychometric criteria which are validity, reliability, difficulty balance, discriminating index, and distractor effective. Several investigations found that most items showed acceptable to high levels of difficulty and discrimination, with a significant proportion of functional distractors, indicating good test construction. In contrast, other studies revealed critical shortcomings, including imbalanced difficulty distribution (often skewed toward overly easy or difficult items), low reliability coefficients, weak discriminating indices, and high rates of non-functional distractors. These weaknesses frequently stemmed from poor item wording, vague stems, or distractors that confused even high-performing students. In contrast, a few studies demonstrated exceptional quality, with over 90% of items classified as very good based on formal norms, though even these acknowledged the need for revision in minor areas such as punctuation and capitalization. Collectively, these findings confirm that while some teacher-made assessments meet high standards, many still require improvement through rigorous item analysis procedures. As a result,

consistent application of item evaluation methods is essential to ensure test validity, fairness, and effectiveness in accurately measuring student learning outcomes.

**3. Multiple Choice Test of English Subject Made by Teacher**

Several previous studies have analyzed multiple-choice questions for English subjects made by English teachers. Research by Amalia & Nur, (2020) aim to The study examined the quality of an English final semester test from the 2018/2019 academic year of Junior highschool in Ponorogo by analyzing 151 student answer sheets using the Quest program, focusing on item difficulty, discrimination, and distractor effectiveness. The results indicated an imbalance in the distribution of item difficulty levels. However, the test showed strong discriminating power, with 39 items (97.5%) classified as excellent in distinguishing between high- and low-performing students. Additionally, 80% of the items featured effective distractors. These findings highlight the critical role of item analysis in ensuring the accuracy and quality of test instruments, which directly impacts the validity of students' scores.

The next study analyzed the level of difficulty and discriminating power of the teacher's English multiple-choice questions. Research by Karim et al., (2021) aim to assess 50 teacher-made multiple-choice test items and found that 16 items were rejected due to poor difficulty levels and low discrimination indices. Additionally, 12 items required revision for their moderate quality, while 11 were considered good. The remaining 11 items

demonstrated excellent quality, with discrimination power (DP) scores ranging from 0.44 to 0.78. These findings underscore the importance of evaluating teacher-constructed test items to ensure their effectiveness, particularly in the context of multiple-choice assessments.

English multiple choice questions for summative exams have also been analyzed before. Research by Semiun et al., (2022) aim to analyzed multiple-choice items of English summative tests constructed by junior high school EFL teachers in Kupang, NTT. The result conclude the English summative tests for Grades VIII and IX generally consist of well-constructed items that function effectively. Most items were categorized as easy, which may be attributed to students' familiarity with the content or prior exposure during instruction. In terms of discrimination index, the majority of items successfully differentiated between high- and low-performing students. Additionally, most distractors performed effectively, with only a few identified as poor. These findings suggest that the tests are of acceptable quality, though item difficulty may require better balance to enhance assessment rigor.

The combined results of the reviewed studies affirm the essential role of item analysis in evaluating the overall quality of teacher-made multiple-choice tests across educational settings. While a significant number of tests demonstrated high-quality items, particularly in terms of discriminating power and distractor effectiveness, several studies also reported imbalances in item difficulty, with many tests containing a disproportionate number of

easy items. This raises concerns about the tests' ability to adequately challenge students and measure a full range of abilities. Discrimination indices were generally strong in some cases, indicating that many items could distinguish between high- achieving and low-achieving students however, other findings revealed a substantial number of items with poor or mediocre discrimination, often requiring revision or rejection. In terms of distractor functionality, most studies noted a high proportion of effective distractors, though non-functional or misleading options were still present in some instruments, particularly those lacking clarity or alignment with the tested material. Collectively, these findings reinforce that consistent application of item analysis procedures is vital not only for enhancing the psychometric quality of assessments but also for supporting fair and accurate measurement of student learning outcomes.

## B. Theoretical Framework

In the context of educational measurement, item analysis has become an essential component in evaluating the quality of test items, particularly in formal assessments such as final-semester examinations. Item analysis refers to the systematic process of examining student responses to individual test items with the purpose of identifying the items' effectiveness, clarity, and ability to differentiate between different levels of student performance. By analyzing test items, educators can improve the quality of their assessments and ensure that they are aligned with learning objectives and standards.

Three major components of item analysis are commonly examined: difficulty level, discriminating power, and distractor effectiveness. The difficulty level of a test item refers to the proportion of students who answer the item correctly. It reflects how easy or difficult a question is and helps determine whether the item is suitable for the intended level of learners. An ideal test includes a mixture of easy, moderate, and difficult items to adequately measure the range of student abilities. Discriminating power, or the discrimination index, shows how well a test item can distinguish between high-achieving and low-achieving students. A good item will yield high scores for students who understand the material and low scores for those who do not, thus serving its diagnostic function. Distractor effectiveness is another critical aspect, particularly in multiple-choice tests. Distractors are the incorrect options offered in each item, and their function is to serve as plausible alternatives that can challenge students who have not mastered the content. Effective distractors attract the attention of less knowledgeable students and contribute to the discriminatory strength of the item. On the contrary, weak distractors that rarely chosen by students may indicate poor item construction and the need for revision.

The analysis of these three components is highly beneficial for multiple stakeholders. For teachers, item analysis provides actionable feedback on how well their tests reflect students' understanding and whether the questions they use are valid indicators of learning. For curriculum designers and academic administrators, it offers a basis for developing standardized,

fair, and effective assessment tools. Ultimately, it enhances the credibility and educational value of the assessment process.

This study is designed to conduct an item analysis on English mid-semester examination questions by evaluating their difficulty level, discriminating power, and the effectiveness of distractors. The results of this analysis are expected to provide insights that help improve future assessments and support educators in creating more valid, reliable, and student-centered test items.

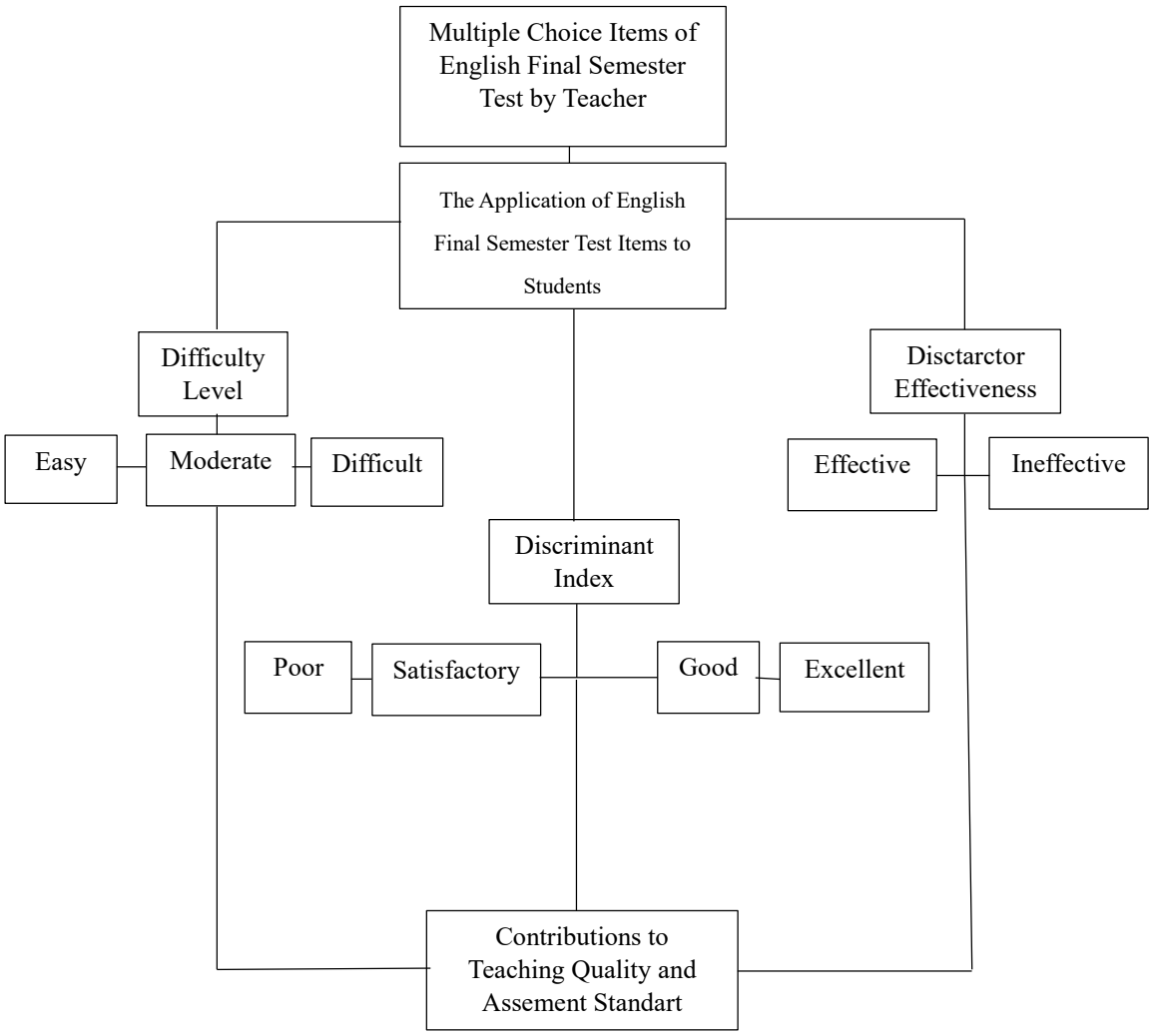


Figure 2. 1 Theoretical Framework