

Evaluating Listening Comprehension Tests Through Classical Test Theory: Analysing Item Difficulty and Discrimination

Abdullah Ahmed

Department of English Language Teaching, South Eastern University of Sri Lanka.

ABSTRACT: Listening comprehension plays a crucial role in second language (L2) acquisition, and reliable assessment is essential for measuring learner proficiency. This study examines the quality of a 25-item listening comprehension test taken by 34 first-year university students, using Classical Test Theory (CTT) for item analysis. Findings showed that 52% of items were moderately difficult, 28% easy, and 20% difficult, indicating overall balance but with some mismatches to students' proficiency levels. In terms of discrimination, 72% of the items effectively distinguished between high- and low-performing students, though some items showed weak discrimination, particularly at difficulty extremes. Distractor analysis further revealed issues with implausible options that affected test reliability. The study highlights the importance of moderately difficult items for stronger discrimination and recommends revising weak items and distractors. These findings offer practical insights for language educators and test developers in improving the validity and fairness of listening comprehension assessments.

KEYWORDS: Listening comprehension, Second language acquisition, Classical Test Theory (CTT), Item analysis, Test validity, Test reliability, Language assessment

I. INTRODUCTION

Listening comprehension is a critical component of language acquisition, particularly in second language (L2) learning contexts. It plays a foundational role in the development of other language skills, such as speaking, reading, and writing (Vandergrift & Goh, 2012). For students learning a new language, listening enables them to grasp vocabulary, syntax, and pronunciation patterns in real-time communication settings. Consequently, evaluating listening comprehension through standardized tests has become a significant tool in determining learners' proficiency and overall language competence (Buck, 2001).

Listening comprehension tests are widely used in educational settings to measure students' ability to understand spoken language. These tests usually present learners with audio materials, followed by a series of questions that assess their understanding of the content. Given its vital role in language education, listening assessment is an area that has drawn significant attention from researchers, educators, and language testing experts alike. Well-designed listening tests not only assess students' current abilities but also help teachers identify areas that need improvement, guiding future instruction (Field, 2008).

Classical Test Theory (CTT) has long been considered a valuable tool for test developers and educators in assessing the quality of language tests. One of the strengths of CTT is its simplicity and ability to provide clear, actionable results from test data (Lord & Novick, 1968). By calculating key statistics such as item difficulty and discrimination, CTT allows test developers to identify items that may not function as intended. For instance, if an item is too easy or too difficult, it may not contribute effectively to differentiating between test-takers of varying abilities (Gorsuch, 2013).

CTT is also useful in providing a straightforward framework for analysing test reliability. It allows for the identification of problematic items that could skew overall test results, helping educators and researchers refine their assessments (Hughes, 2003). Furthermore, CTT is relatively accessible to educators and researchers without advanced statistical knowledge, making it a practical approach for item analysis in listening comprehension tests.

The importance of this study is further underscored by the fact that listening comprehension is often regarded as one of the most difficult skills to assess (Field, 2008). The use of CTT in this context will help illuminate how listening test items can be optimized to enhance both their fairness and effectiveness, ultimately leading to improved assessments in language learning contexts.

A. Research Problem and Purpose

Despite their widespread use, developing reliable and valid listening comprehension tests remains a complex task. One of the most critical challenges lies in evaluating the quality of individual test items (Brown, 2012). Test developers must ensure that the items on a listening test are appropriate for the target population, discriminating well between high and low performers while also being

Evaluating Listening Comprehension Tests Through Classical Test Theory: Analysing Item Difficulty and Discrimination

neither too difficult nor too easy (Henning, 1987). If test items do not meet these criteria, the validity and reliability of the test can be compromised, potentially providing inaccurate reflections of students' listening abilities.

Moreover, test items should provide consistent measures of listening ability across different populations. This highlights the importance of analysing both item difficulty the proportion of test-takers answering a question correctly and item discrimination how well a particular item differentiates between higher- and lower-performing students (Hughes, 2003). These two factors are crucial for ensuring that a listening test accurately assesses students' abilities and provides meaningful results that can be used for instructional and evaluative purposes.

This study aims to evaluate the quality of a listening comprehension test by analysing the difficulty and discrimination of its items using Classical Test Theory. CTT is a widely used framework in test analysis that helps determine the reliability and validity of test items by focusing on observed scores, true scores, and error scores (Miller, Linn, & Gronlund, 2013). By employing this framework, we can examine how well the test items function in differentiating between students of varying proficiency levels and ensuring that the test is fair and effective.

The evaluation of item difficulty will provide insight into whether the test items are too easy, too difficult, or appropriately challenging for the intended group of test-takers. Simultaneously, item discrimination will assess the effectiveness of each test item in distinguishing between students who possess a strong understanding of the material and those who do not (Brown, 2012). Through this analysis, the study will contribute to a better understanding of how listening comprehension tests can be refined to improve their validity and fairness.

B. RESEARCH OBJECTIVES

The study was designed with the following objectives:

1. To evaluate the quality of listening comprehension test items by analysing their difficulty level, discrimination power, and distractor efficiency using Classical Test Theory (CTT).
2. To provide recommendations for improving listening comprehension test design in order to enhance validity, reliability, and fairness in L2 assessment.

These objectives emphasize both the analytical and practical dimensions of the study. The first objective focuses on examining test items in detail, ensuring that they function as reliable indicators of students' listening ability. The second objective moves beyond evaluation to highlight the importance of applying findings to strengthen future test design and improve assessment practices in second language classrooms.

C. RESEARCH QUESTIONS

In line with the objectives, the study seeks to answer the following questions:

- I. How effective are the listening comprehension test items in terms of difficulty, discrimination, and distractor quality as measured by Classical Test Theory?
- II. What insights can item analysis provide for improving the validity and reliability of listening comprehension assessments in second language (L2) contexts?

These questions guide the investigation by linking theory with practice. The first question directly addresses the technical quality of the test items, examining how well they challenge learners and differentiate between varying levels of ability. The second question extends the inquiry toward application, focusing on how item analysis findings can contribute to refining test construction, ensuring more valid, reliable, and fair listening comprehension assessments in L2 settings.

II. LITERATURE REVIEW

A. Listening Comprehension Testing

Listening comprehension is one of the core skills in language learning and is essential for effective communication in both first and second languages. Listening allows learners to understand spoken language in real-time and respond appropriately in conversations, lectures, and other auditory situations. As Field (2008) points out, listening is often seen as a passive skill, but in reality, it is an active and dynamic process that involves cognitive engagement, meaning construction, and information retention. In language learning, it plays a crucial role because it is usually the first language skill learners are exposed to, and it helps them build a foundation for other skills such as speaking, reading, and writing (Vandergrift & Goh, 2012).

Assessing listening comprehension has always been a challenging task. Listening comprehension tests usually involve audio materials, followed by questions that assess a learner's understanding. Traditional test formats include multiple-choice questions (MCQs), gap-filling tasks, true/false statements, and short answer questions. These tests aim to measure various aspects of listening, such as understanding the main idea, identifying specific details, making inferences, and recognizing the speaker's tone or intent (Buck, 2001).

The assessment of listening comprehension relies heavily on the test's ability to accurately measure a learner's listening skills, making the design and evaluation of test items crucial. According to Hughes (2003), the quality of a listening comprehension test

Evaluating Listening Comprehension Tests Through Classical Test Theory: Analysing Item Difficulty and Discrimination

depends on its validity whether it measures what it claims to measure and its reliability the consistency of the test results across different administrations. However, designing reliable and valid listening comprehension tests is often complex due to the various factors that affect listening, such as background noise, unfamiliar accents, speech rate, and content complexity (Field, 2008). Therefore, test developers need robust frameworks to ensure that test items accurately assess students' listening abilities.

B. Classical Test Theory (CTT)

Classical Test Theory (CTT) is one of the most widely used frameworks for evaluating the reliability and validity of test items. CTT focuses on analysing test scores and provides a simple yet effective way to assess the quality of test items (Lord & Novick, 1968). In CTT, a student's observed score is considered a combination of two components: the true score (which represents the student's actual ability) and the error score (which accounts for the variability in test scores due to factors like guessing, fatigue, or test conditions). This relationship is expressed in the formula:

$$\text{Observed Score} = \text{True Score} + \text{Error Score}$$

CTT emphasizes the importance of reducing error variance to improve test reliability. According to Miller, Linn, and Gronlund (2013), if the error variance is minimized, the observed scores will provide a more accurate reflection of students' true abilities.

One of the key strengths of CTT is its focus on item analysis, which examines individual test items to assess their difficulty, discrimination, and functionality. Through item analysis, test developers can identify poorly functioning items that either fail to measure what they are intended to or do not differentiate between students of varying abilities (Brown, 2012). Two critical components of item analysis in CTT are item difficulty and item discrimination.

C. Item Difficulty

Item difficulty is a crucial metric in test evaluation, referring to the proportion of test-takers who answer a particular item correctly (Henning, 1987). It is calculated using the following formula:

$$P = \frac{\text{Number of correct responses}}{\text{Total number of responses}}$$

Total number of responses

Where P represents the item difficulty index, which ranges from 0 to 1. A value closer to 1 indicates that the item is easy, while a value closer to 0 suggests that the item is difficult (Brown, 2012). Typically, an item is considered to have an appropriate difficulty level if its difficulty index falls between 0.30 and 0.70, meaning that around 30-70% of test-takers answered the item correctly (Hughes, 2003).

Previous research has shown that item difficulty plays a significant role in determining the overall quality of a test. Items that are too easy or too difficult do not contribute meaningfully to differentiating between students of varying abilities (Henning, 1987). For example, research by Alderson (2000) found that language tests with an optimal range of item difficulty were more effective in assessing students' actual proficiency levels. Moreover, listening comprehension tests often face challenges in balancing item difficulty due to the variability in factors such as audio quality, accents, and background noise (Buck, 2001). Studies emphasize the importance of continuously evaluating and adjusting item difficulty to ensure that tests provide a fair assessment of listening skills.

D. Item Discrimination

Item discrimination is another vital component of test evaluation, representing how well an item differentiates between high-performing and low-performing students (Brown, 2012). Items with high discrimination effectively distinguish between test-takers who have mastered the content and those who have not. The discrimination index is calculated using the following formula:

$$D = \frac{\text{High group performance} - \text{Low group performance}}{\text{Number of students}}$$

Where D is the discrimination index, and "high group performance" refers to the number of correct responses from the top-performing students, while "low group performance" refers to the number of correct responses from the bottom-performing students (Brown, 2012).

A well-discriminating item will have a positive discrimination index, meaning that students who perform well on the overall test are more likely to answer the item correctly. On the other hand, a negative or zero discrimination index indicates that the item does not effectively differentiate between high and low performers. Henning (1987) argues that items with low discrimination should be revised or removed, as they fail to provide meaningful insights into students' abilities.

Several studies have highlighted the role of item discrimination in improving test quality. For instance, Alderson (2000) demonstrated that test items with high discrimination indices contributed significantly to the overall reliability of language tests. Similarly, Buck (2001) emphasized the importance of item discrimination in listening comprehension tests, noting that well-discriminating items provide a clearer picture of a learner's listening skills. In language testing, the combination of item difficulty and discrimination helps ensure that tests are both challenging and fair, giving students of all ability levels an opportunity to perform well.

Evaluating Listening Comprehension Tests Through Classical Test Theory: Analysing Item Difficulty and Discrimination

D. Applications of CTT in Language Testing

Classical Test Theory has been widely applied in language testing to evaluate the reliability and validity of test items. One of the primary applications of CTT is in the development and refinement of listening comprehension tests. CTT allows test developers to assess how well individual items function, identify problematic items, and improve the overall quality of the test (Miller et al., 2013).

Numerous studies have used CTT to analyse language tests, particularly in listening comprehension. For example, Gorsuch (2013) conducted a study that applied CTT to evaluate the quality of a listening comprehension test for English language learners. The study found that using CTT to analyse item difficulty and discrimination allowed the researchers to identify and revise problematic test items, leading to a more reliable and valid assessment tool.

Similarly, Buck (2001) utilized CTT in his analysis of listening comprehension tests, emphasizing that CTT provides valuable insights into how test items perform and how they can be adjusted to better assess students' listening skills. In particular, CTT's focus on item analysis helps ensure that listening comprehension tests are appropriately challenging and effective in distinguishing between students of varying proficiency levels.

In addition to its application in listening tests, CTT has been used to evaluate other types of language assessments, including reading, writing, and speaking tests. Hughes (2003) notes that CTT's versatility and accessibility make it an ideal framework for educators and researchers seeking to improve language testing practices. By focusing on key metrics such as item difficulty and discrimination, CTT provides a clear path for refining test items and improving the overall quality of language assessments.

In conclusion, listening comprehension testing is a crucial element of language assessment, but developing reliable and valid tests can be challenging. Classical Test Theory provides a robust framework for evaluating the quality of test items, with a focus on item difficulty and discrimination. These two factors are essential for ensuring that listening comprehension tests accurately assess students' abilities and provide meaningful results. Through item analysis, CTT allows test developers to identify and revise problematic items, improving the reliability and validity of the test. As language testing continues to evolve, the application of CTT will remain a valuable tool for ensuring that listening comprehension tests are fair, reliable, and effective.

III. METHODOLOGY

A. Test Overview

This study is based on a listening comprehension test administered to 34 first-year university students. The test comprised 25 questions designed to evaluate various aspects of listening skills, including understanding the main idea, identifying details, and making inferences. The test questions were based on audio materials, which presented dialogues and monologues that simulated real-world conversations and lecture-style information. These materials were selected to ensure relevance to the students' academic and social contexts, thereby increasing the validity of the test in measuring their listening comprehension abilities.

The listening test covered different levels of comprehension, from basic recall of information to more complex tasks like inferencing and analysing the speaker's intent. The range of questions was intended to assess both lower-order thinking skills, such as remembering and understanding, and higher-order thinking skills, such as analysing and evaluating (Krathwohl, 2002).

B. Test Design

The design of the listening test was informed by Bloom's Taxonomy, a widely used educational framework that categorizes learning objectives into different cognitive levels (Anderson & Krathwohl, 2001). The questions were designed to align with various levels of Bloom's Taxonomy, ensuring a balanced assessment of both basic and advanced listening skills.

- **Remembering and Understanding:** These questions focused on recalling specific details from the audio, such as names, dates, and places. They tested the students' ability to comprehend the information presented in the recording.
- **Applying and Analysing:** Some questions required students to apply their understanding to new contexts or to break down information to determine relationships between ideas. For example, students had to infer the meaning of unfamiliar words based on context or analyse the speaker's tone.
- **Evaluating and Creating:** A few questions challenged students to evaluate the speaker's argument or form a conclusion based on the evidence provided. These higher-order questions assessed critical thinking and reflective listening skills.

By incorporating questions from various cognitive levels, the test aimed to provide a comprehensive assessment of the students' listening comprehension skills.

C. Data Collection

The listening test was administered during a regular class session, ensuring that the testing environment was familiar and free from unnecessary distractions. The students were instructed to listen to the audio materials, which were played twice, and then respond to the questions. They were given ample time to complete the test after the audio had been played.

Evaluating Listening Comprehension Tests Through Classical Test Theory: Analysing Item Difficulty and Discrimination

Once the students had completed the test, their responses were collected and manually scored. Each correct response received one point, and no partial credit was given. The maximum possible score was 25 points. The scores were recorded and prepared for further analysis using Classical Test Theory (CTT) to evaluate the quality of the test items.

D. Data Analysis

The primary framework used to analyse the test data was Classical Test Theory (CTT). CTT is a well-established model in educational measurement that helps assess the reliability and validity of test items (Lord & Novick, 1968). In CTT, two main statistics are used to evaluate test items: item difficulty and item discrimination.

- **Item Difficulty:** Item difficulty refers to the proportion of students who answered a question correctly. It is calculated using the following formula:

$$P = \frac{\text{Number of correct responses}}{\text{Total number of responses}}$$

Where P represents the item difficulty index. The values range from 0 to 1, where a value closer to 0 indicates that the item is difficult (few students answered it correctly), and a value closer to 1 indicates that the item is easy (most students answered it correctly) (Brown, 2012).

The item difficulty index was used to categorize the questions into three groups:

- **Easy:** Items with a difficulty index greater than 0.70 (answered correctly by more than 70% of students).
- **Moderate:** Items with a difficulty index between 0.30 and 0.70 (answered correctly by 30-70% of students).
- **Difficult:** Items with a difficulty index below 0.30 (answered correctly by fewer than 30% of students).

This categorization helped in determining whether the test items were appropriately challenging for the group of test-takers.

- **Item Discrimination:** Item discrimination measures how well a test item distinguishes between high-performing and low-performing students. A well-functioning item should be answered correctly by students who perform well on the overall test and answered incorrectly by students who perform poorly (Henning, 1987). The discrimination index is calculated as follows:

$$D = \frac{\text{High group performance} - \text{Low group performance}}{\text{Number of students}}$$

The "high group" consists of the top 27% of students based on their overall test scores, while the "low group" consists of the bottom 27%. The discrimination index ranges from -1 to 1:

- A value closer to 1 indicates that the item has high discrimination (strong students tend to answer it correctly, while weak students tend to answer it incorrectly).
- A value of 0 indicates that the item does not discriminate between strong and weak students.
- A negative value indicates poor discrimination, where weak students are more likely to answer the item correctly than strong students.

Items with a high discrimination index are valuable for distinguishing between students of varying abilities, while items with low or negative discrimination are generally considered problematic and may need to be revised or eliminated (Hughes, 2003).

E. Interpretation of Item Analysis

The results from the item analysis provided insight into the overall quality of the listening test. Items that were too easy or too difficult did not contribute effectively to differentiating between students' abilities, as they either provided no challenge or were beyond the capabilities of most students (Brown, 2012). Similarly, items with low discrimination indices were flagged for revision, as they did not effectively separate high-performing students from low-performing ones.

Based on the analysis, recommendations were made for improving the test. For example, items with very high or very low difficulty indices might need to be revised to ensure they are aligned with the target population's abilities. Additionally, items with poor discrimination indices could be reworded or replaced to better assess students' comprehension skills.

The methodology used in this study followed a systematic approach to evaluating the quality of the listening comprehension test. By using Classical Test Theory (CTT) to analyse item difficulty and discrimination, the study provided valuable insights into how well the test items functioned in assessing students' listening skills. The item analysis not only identified areas for improvement in the test design but also highlighted the importance of using a balanced range of item difficulties and well-discriminating questions to provide a fair and valid assessment of students' abilities.

IV. FINDINGS

A. Item Difficulty Analysis

The first step in evaluating the quality of the listening comprehension test involved analysing the difficulty levels of the 25 items. As described in the methodology, item difficulty is measured by the proportion of students who answered each item correctly,

Evaluating Listening Comprehension Tests Through Classical Test Theory: Analysing Item Difficulty and Discrimination

ranging from 0 (very difficult) to 1 (very easy). The results of the analysis revealed that the items fell into three distinct difficulty categories: easy, moderate, and difficult.

- **Easy items:** Out of the 25 items, 7 (28%) were classified as easy, with a difficulty index greater than 0.70. This indicates that more than 70% of students answered these items correctly. These questions generally focused on simple recall of facts or identifying specific details from the listening material.
- **Moderate items:** A total of 13 items (52%) fell within the moderate difficulty range, with a difficulty index between 0.30 and 0.70. These items required students to demonstrate a deeper understanding of the listening passage, such as making inferences or recognizing relationships between ideas. The balance of moderate items suggests that the test was neither too easy nor too difficult for the group of test-takers.
- **Difficult items:** There were 5 items (20%) that were classified as difficult, with a difficulty index below 0.30. These items were answered correctly by fewer than 30% of the students and typically involved higher-order thinking, such as evaluating the speaker's argument or interpreting complex information.

The distribution of item difficulty is illustrated in Figure 1 below. This balance of easy, moderate, and difficult items indicates that the test provided an appropriate range of questions, catering to different levels of student ability. However, the presence of a few difficult items suggests that some students may have struggled with certain aspects of the test, particularly higher-level comprehension tasks.

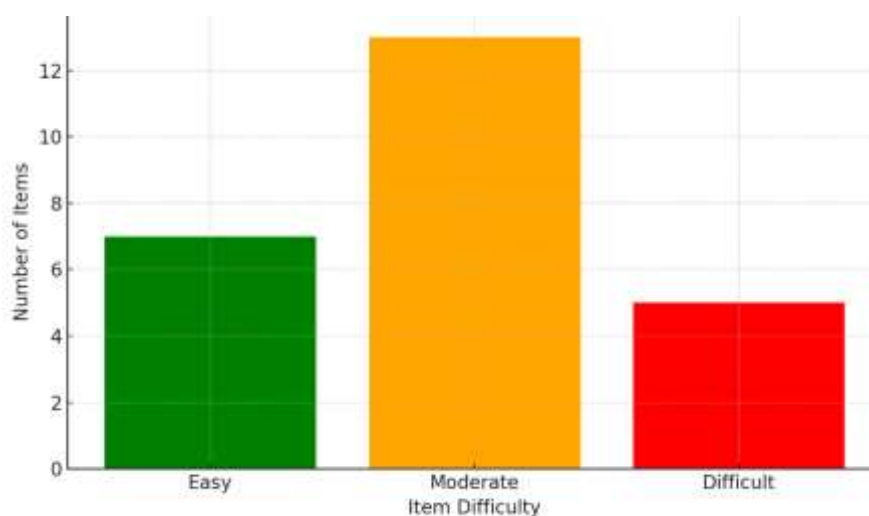


Figure 1: Distribution of Item Difficulty

Overall, the analysis of item difficulty showed that the majority of test items fell within the moderate range, providing a balanced assessment that was appropriately challenging for most students. However, further adjustments could be made to the more difficult items to ensure they are fair for all test-takers.

B. Item Discrimination Analysis

In addition to item difficulty, the analysis examined the discrimination indices for each of the 25 items. The discrimination index measures how well an item differentiates between high-performing and low-performing students. Items with a high discrimination index are those that are answered correctly by strong students and incorrectly by weaker students, making them valuable in assessing students' abilities.

The results of the discrimination analysis revealed the following classifications:

- **Very good discrimination:** 8 items (32%) had discrimination indices above 0.40, indicating that they effectively differentiated between high- and low-performing students. These items were mostly moderate in difficulty, requiring students to apply critical thinking and deeper comprehension skills.
- **Good discrimination:** 10 items (40%) had discrimination indices between 0.30 and 0.40, suggesting that they provided a reasonable level of distinction between students of varying abilities. These items included a mix of easier and moderately challenging questions.
- **Fair discrimination:** 5 items (20%) had discrimination indices between 0.20 and 0.29. While these items were somewhat effective in distinguishing between stronger and weaker students, they may need to be revised to improve their discriminatory power. These items included both easy and difficult questions, which may have led to inconsistent performance across different ability levels.

Evaluating Listening Comprehension Tests Through Classical Test Theory: Analysing Item Difficulty and Discrimination

- **Poor discrimination:** 2 items (8%) had discrimination indices below 0.20, indicating that they did not perform well in distinguishing between high- and low-performing students. Both of these items were classified as difficult and may have been too challenging for the majority of students, resulting in low discrimination scores.

The relationship between item difficulty and item discrimination is summarized in **Figure 2**. It can be seen that items with moderate difficulty tended to have higher discrimination indices, while items that were either too easy or too difficult generally exhibited lower discrimination values.

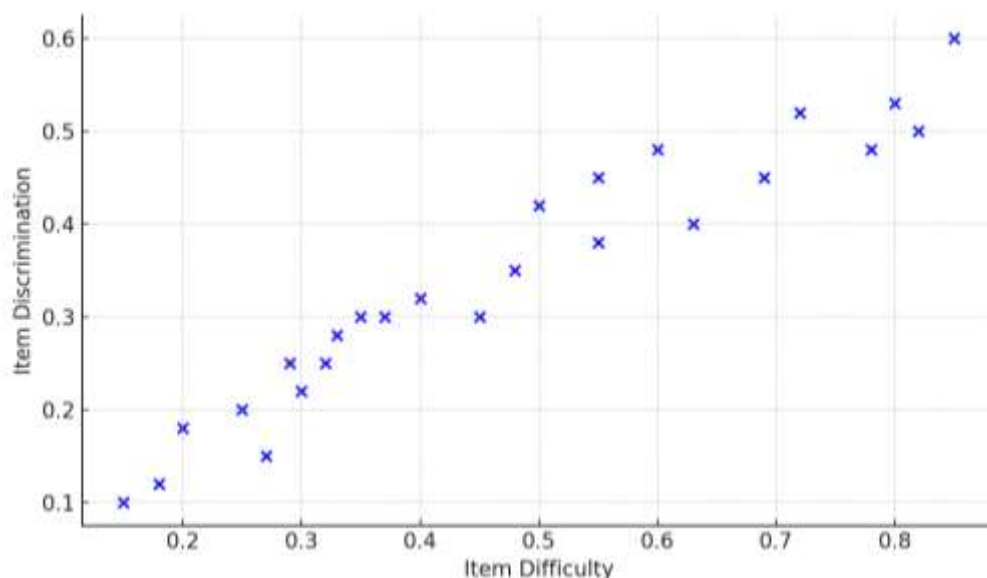


Figure 2: Relationship Between Item Difficulty and Discrimination

The analysis highlights that the most effective items in terms of discrimination were those of moderate difficulty, as they challenged students without being overly difficult. In contrast, items that were either too easy or too difficult struggled to differentiate between strong and weak performers, as nearly all students either answered them correctly or incorrectly.

C. Distractor Analysis

Finally, the distractor analysis was conducted to evaluate the efficiency of the multiple-choice options provided for each question. Distractors, or incorrect answer choices, play a crucial role in determining the effectiveness of a test item. Ideally, distractors should be plausible enough to attract students who do not fully understand the material, but not so confusing that they mislead students who do know the correct answer.

Table 1: Distractor analysis

Item no.	A	B	C
Q1	2	18	14
Q2	24	10	0
Q3	17	4	13
Q4	27	5	2

The distractor analysis identified several instances where the distractors did not perform as intended. Specifically:

- **Malfunctioning distractors:** In 4 items, one or more distractors were rarely chosen by students, indicating that these options were not sufficiently plausible. For example, in one item, over 90% of students chose the correct answer, while less than 10% selected any of the distractors. This suggests that the distractors were too obviously incorrect, reducing the overall difficulty and discrimination of the item.
- **Ineffective distractors:** In 3 items, the distractors were too similar to the correct answer, leading to confusion among students. In these cases, students often selected the distractors at nearly the same rate as the correct answer, resulting in low discrimination scores for these items.

Evaluating Listening Comprehension Tests Through Classical Test Theory: Analysing Item Difficulty and Discrimination

To address these issues, the distractors for the malfunctioning and ineffective items should be revised. For the malfunctioning items, more plausible distractors should be created to increase the difficulty of the item. For the ineffective distractors, adjustments should be made to ensure that they are distinguishable from the correct answer without being too misleading. These revisions will improve the overall effectiveness of the test by providing clearer distinctions between correct and incorrect responses.

In summary, the analysis of the listening comprehension test revealed a balanced distribution of item difficulty, with most items falling within the moderate range. The item discrimination analysis showed that many items effectively distinguished between high- and low-performing students, though a few items with poor discrimination require revision. Additionally, the distractor analysis identified several issues with malfunctioning or ineffective distractors, which should be addressed to improve the test's overall quality.

Through these findings, it is clear that while the test was generally effective in assessing students' listening comprehension skills, there is room for improvement. By refining the most difficult items, improving poorly discriminating questions, and revising ineffective distractors, future iterations of the test can provide a more reliable and valid measure of students' listening abilities.

V. DISCUSSION

A. Interpretation of Findings

The results of this study provide valuable insights into the quality of the listening comprehension test, particularly in terms of item difficulty and item discrimination. The analysis of item difficulty revealed that the majority of test items (52%) fell within the moderate range, while 28% of the items were classified as easy, and 20% were considered difficult. This distribution suggests that the test was relatively well-balanced, offering an appropriate mix of challenging and accessible items for the students. However, the presence of 20% difficult items indicates that some questions may have been too challenging for the target population, which consisted of first-year university students with varying levels of proficiency in listening comprehension.

From a pedagogical perspective, it is essential that test items are appropriately challenging, providing students with an opportunity to demonstrate their abilities without overwhelming them (Brown, 2012). The balance between easy, moderate, and difficult items in this test aligns with this principle, as it offers questions that assess both lower-order and higher-order listening skills. For example, the easy items allowed students to recall basic information, while the moderate and difficult items required more complex skills such as inferencing and analysis. This range of difficulty helps ensure that students across different ability levels are fairly assessed (Hughes, 2003).

However, the 20% of difficult items may indicate a potential misalignment between the content of those questions and the students' current level of language proficiency. In future tests, it may be necessary to review the content of these difficult items to ensure that they are not overly complex or ambiguous. As suggested by Henning (1987), test items that are too difficult for the majority of students may reduce the overall validity of the test, as they fail to accurately measure students' abilities.

In terms of item discrimination, the majority of the items (72%) showed good to very good discrimination, meaning that they were effective in distinguishing between high- and low-performing students. Items with high discrimination are essential for ensuring that the test accurately measures the differences in students' listening comprehension skills (Alderson, 2000). The items that demonstrated good discrimination were primarily those that fell within the moderate difficulty range, which aligns with previous research indicating that moderately difficult items are most effective at differentiating between students of varying abilities (Buck, 2001).

However, 28% of the items had fair or poor discrimination indices, suggesting that these items were less effective at distinguishing between strong and weak students. These items included both easy and difficult questions, indicating that extremes in item difficulty tend to reduce the discriminatory power of a test (Brown, 2012). This finding highlights the importance of designing items that are neither too easy nor too difficult, as this balance is crucial for ensuring that the test provides meaningful assessments of student performance.

B. Comparison with Previous Studies

The findings of this study are consistent with previous research on the role of item difficulty and discrimination in language testing. For instance, Alderson (2000) emphasizes the importance of using a range of item difficulties to ensure that a test is appropriately challenging for all students. The distribution of easy, moderate, and difficult items in this study aligns with his recommendation that around 60% of test items should fall within the moderate difficulty range, as this provides a fair assessment for students of varying proficiency levels. In this study, 52% of the items were classified as moderate, which is close to the optimal balance suggested by Alderson.

In terms of item discrimination, Buck (2001) highlights the importance of items that can effectively differentiate between high and low performers, particularly in listening comprehension tests. He argues that items with high discrimination indices are essential for ensuring that a test accurately reflects students' abilities. The fact that 72% of the items in this study demonstrated good to very

Evaluating Listening Comprehension Tests Through Classical Test Theory: Analysing Item Difficulty and Discrimination

good discrimination aligns with Buck's findings, suggesting that the test performed well in distinguishing between strong and weak students.

However, this study also found that a subset of items (28%) had fair or poor discrimination indices, which mirrors findings from previous research. According to Henning (1987), items with low discrimination indices may fail to provide useful information about students' abilities, as they do not differentiate between high and low performers effectively. This is particularly true for items that are either too easy or too difficult, as they tend to produce similar results across different ability levels (Brown, 2012). The presence of such items in this study underscores the need for careful item review and revision to ensure that all test items contribute meaningfully to the assessment of students' listening comprehension skills.

C. Limitations of the Study

While this study provides important insights into the quality of the listening comprehension test, it is important to acknowledge several limitations. First, the sample size was relatively small, consisting of only 34 first-year university students. This limits the generalizability of the findings, as the results may not be representative of a broader population of students with different backgrounds or proficiency levels. Future research should aim to replicate this study with a larger and more diverse sample to ensure that the findings are applicable to a wider range of learners (Field, 2008).

Second, the specific context of the test—focusing on first-year university students—may also limit the applicability of the results to other contexts. The students in this study were at the beginning of their academic journey, and their listening comprehension skills may differ from those of more advanced students. Additionally, the test content was designed to reflect the students' academic and social environments, which may not be relevant to other groups of learners. As Hughes (2003) points out, the content of a test plays a crucial role in determining its validity, and tests designed for specific populations may not be suitable for other contexts.

Finally, the test consisted of only 25 items, which may limit the depth and breadth of the assessment. While the test covered a range of listening skills, a larger number of items would allow for a more comprehensive evaluation of students' abilities. Future research should consider expanding the test to include additional items that assess a wider variety of listening skills, particularly those related to higher-order comprehension tasks such as evaluating and synthesizing information.

D. Recommendations

Based on the results of this study, several recommendations can be made for improving the design of the listening comprehension test. First, it is important to revise or eliminate the malfunctioning items, particularly those with low discrimination indices. Items that are too easy or too difficult should be reviewed to ensure that they are appropriately challenging for the target population. For example, the difficult items that were answered correctly by fewer than 30% of students may need to be simplified or clarified to ensure that they accurately assess students' abilities (Brown, 2012).

Second, the balance between item difficulty and discrimination should be carefully considered in future tests. As this study has shown, moderately difficult items tend to have the highest discrimination indices, making them the most effective at distinguishing between students of different ability levels. Test designers should aim to include a higher proportion of moderately difficult items, as these are likely to provide the most meaningful assessments of students' listening comprehension skills (Hughes, 2003).

Third, future tests should include a more diverse range of item types, particularly those that assess higher-order listening skills. While this test included questions that assessed both lower-order and higher-order skills, there may be opportunities to include additional items that challenge students to analyse, evaluate, and synthesize information from the listening passages. These higher-order tasks are particularly important in academic contexts, where students are often required to engage in critical thinking and problem-solving based on what they hear (Field, 2008).

Finally, future research should aim to replicate this study with a larger and more diverse sample of students. A larger sample size would allow for more robust statistical analyses and provide a clearer picture of the overall quality of the test items. Additionally, including students from different academic levels or language backgrounds would help ensure that the test is valid and reliable across a range of learning contexts.

The results of this study highlight the importance of balancing item difficulty and discrimination in listening comprehension tests. While the majority of the items in this test were appropriately challenging and effectively differentiated between high and low performers, there is still room for improvement. By revising malfunctioning items, focusing on moderate difficulty levels, and including a broader range of listening tasks, future tests can provide more accurate and meaningful assessments of students' listening comprehension skills.

VI. CONCLUSION

This study aimed to evaluate a listening comprehension test for first-year university students by analysing item difficulty and discrimination using Classical Test Theory (CTT). The findings indicate that the test was generally well-balanced in terms of difficulty, with the majority of the items (52%) falling within the moderate range, ensuring that the test provided an appropriate

Evaluating Listening Comprehension Tests Through Classical Test Theory: Analysing Item Difficulty and Discrimination

level of challenge for the students. However, the presence of some items that were classified as too easy or too difficult suggests that there is room for improvement in aligning all test items with the students' abilities.

In terms of item discrimination, most of the test items (72%) demonstrated good to very good discrimination, effectively distinguishing between high-performing and low-performing students. This is a critical aspect of test quality, as well-discriminating items ensure that the test accurately reflects students' listening comprehension skills. However, a subset of items with fair or poor discrimination suggests that those items may need revision to improve their effectiveness in differentiating students with varying levels of proficiency.

The use of CTT in this study proved to be a valuable tool for assessing the quality of the test items. By calculating item difficulty and discrimination, CTT allowed for a systematic evaluation of each item, highlighting areas where the test could be improved. The simplicity and accessibility of CTT make it an ideal framework for test developers, educators, and researchers who seek to ensure that their assessments are both reliable and valid (Lord & Novick, 1968). CTT also provides clear, actionable insights that can inform future test revisions, helping to refine the balance between item difficulty and discrimination (Brown, 2012).

Looking ahead, there are several directions for future research. First, expanding the sample size would allow for more robust statistical analyses and ensure that the findings are generalizable to a broader population of learners. This study was limited by its relatively small sample of 34 students, and future studies should aim to include a larger and more diverse group of participants to confirm the validity of the test across different contexts (Hughes, 2003).

Second, the application of CTT could be extended to other language skills, such as reading and writing. While this study focused on listening comprehension, the principles of item difficulty and discrimination are equally relevant to assessments of other language skills. By applying CTT to a broader range of language assessments, future research could help to refine test design across multiple domains of language learning, providing a more comprehensive understanding of students' abilities (Buck, 2001).

In conclusion, this study underscores the importance of using Classical Test Theory to evaluate the quality of language assessments. By focusing on item difficulty and discrimination, CTT provides valuable insights into the effectiveness of test items, helping to ensure that they offer a fair and meaningful assessment of students' skills. With further research and refinement, the listening comprehension test analysed in this study can be improved to better meet the needs of students and educators alike.

REFERENCES

- 1) Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press.
- 2) Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
- 3) Brown, H. D. (2012). *Language assessment: Principles and classroom practices* (2nd ed.). Pearson Education.
- 4) Buck, G. (2001). *Assessing listening*. Cambridge University Press.
- 5) Field, J. (2008). *Listening in the language classroom*. Cambridge University Press.
- 6) Gorsuch, R. L. (2013). *Factor analysis*. Psychology Press.
- 7) Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Heinle & Heinle Publishers.
- 8) Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge University Press.
- 9) Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, 41(4), 212-218.
- 10) Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- 11) Miller, M. D., Linn, R. L., & Gronlund, N. E. (2013). *Measurement and assessment in teaching* (11th ed.). Pearson Education.
- 12) Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. Routledge.



There is an Open Access article, distributed under the term of the Creative Commons Attribution – Non Commercial 4.0 International (CC BY-NC 4.0)

(<https://creativecommons.org/licenses/by-nc/4.0/>), which permits remixing, adapting and building upon the work for non-commercial use, provided the original work is properly cited.