IMPROVING LISTENING TEST QUALITY THROUGH STATISTICAL ANALYSIS: A CASE STUDY USING SPSS

CẢI THIỆN CHẤT LƯỢNG BÀI KIỂM TRA NGHE THÔNG QUA PHÂN TÍCH THỐNG KÊ: MỘT NGHIÊN CỨU ĐIỂN HÌNH SỬ DỤNG SPSS

Tran Thi Tuyet Trinh[1,*], Pham Thi Hong[1], Nguyen Ngoc Quynh[2], Nguyen Phuong Thao[2]

DOI: http://doi.org/10.57001/huih5804.2025.035

ABSTRACT

This study investigates the quality of an English listening comprehension test administered to third-year students at a Vietnamese public university. The research uses SPSS statistical software to evaluate the test through descriptive statistics, reliability, and construct validity analyses. Results reveal that while the mean score indicates overall fair performance, the wide score distribution highlights inconsistencies in student achievement. The Cronbach's Alpha coefficient of 0.671 suggests moderate internal consistency, with several items showing low or negative item-total correlations, indicating potential flaws in the test design. Construct validity is partially supported through item correlations aligned with theoretical expectations. Based on these findings, the study proposes specific revisions, including rewording ambiguous items and removing poorly discriminating questions, to enhance the test's reliability and validity. By presenting a data-driven approach to test evaluation, this paper provides practical insights for educators aiming to improve language assessment practices.

Keywords: *Reliability, validity, SPSS, test assessment, English test, listening.*

TÓM TẮT

Bài báo này phân tích bài kiểm tra tiến độ của sinh viên năm ba tại một trường đại học công lập ở Việt Nam, tập trung vào kỹ năng nghe hiểu tiếng Anh. Nghiên cứu tiến hành phân tích toàn diện về chất lượng bài kiểm tra thông qua các thống kê mô tả, đo độ tin cậy và độ giá trị, sử dụng phần mềm thống kê SPSS. Cụ thể, nghiên cứu đánh giá chỉ số đồng nhất và độ chuẩn xác của thang đo nhằm xác định các vấn đề tiềm ẩn trong thiết kế và kết quả bài làm của người tham gia. Kết quả nghiên cứu chỉ ra điểm mạnh và điểm yếu của bài kiểm tra, đồng thời đưa ra các khuyến nghị dựa trên dữ liệu để cải thiện chất lượng đánh giá. Nghiên cứu này đóng góp vào lĩnh vực đánh giá ngôn ngữ bằng cách đề xuất một phương pháp hệ thống trong việc phân tích bài kiểm tra, có thể làm tài liệu tham khảo cho các nhà giáo dục mong muốn nâng cao chất lượng đánh giá của họ.

Từ khóa: *độ tin cậy bài kiểm tra, độ giá trị bài kiểm tra, SPSS, kiểm tra đánh giá, kiểm tra Tiếng Anh, k ỹ năng nghe.*

[1]School of Languages and Tourism, Hanoi University of Industry, Vietnam

[2]Faculty of Foreign Languages, Thang Long University, Vietnam

[*]Email: trinhttt@haui.edu.vn

Received: 08/01/2025

Revised: 18/02/2025

Accepted: 27/02/2025

1. INTRODUCTION

Language testing plays a crucial role in assessing and enhancing students' language proficiency, particularly in academic settings where structured evaluations inform both teaching and learning practices. A well-designed language test should not only measure students' abilities but also provide diagnostic insights that help educators identify areas requiring further development. As Brown emphasizes, an effective language assessment must be "fair, reliable, and valid," ensuring that test results accurately reflect learners' proficiency and offer meaningful feedback to support instructional decisions [1]. In the context of second language acquisition, listening comprehension is a particularly challenging skill to assess due to its cognitive complexity and the multiple

factors influencing comprehension, such as speech rate, accents, and background knowledge [2].

Despite the importance of listening assessments, research has highlighted common issues in test design, including problems with item difficulty, poor discrimination indices, and a lack of validity [3]. Many standardized and institutional tests fail to adequately differentiate between learners of varying proficiency levels, leading to inaccurate assessments of students' listening skills. Additionally, few studies have systematically analyzed the statistical properties of listening comprehension tests in Vietnamese university settings. Most studies focus on the factors affecting listening ability and listening comprehension. Therefore, there is a research gap in the field of language assessment regarding analyzing the statistical properties of listening tests in Vietnam.

To address this gap, this study investigates the quality of a listening comprehension progress test administered to third-year English major students at a Vietnamese public university. The study uses SPSS statistical software to evaluate the test's reliability and validity.

The study aims to provide empirical evidence for improving listening test design and contribute to best practices in language assessment. The findings will offer practical recommendations for educators and test developers seeking to enhance the quality of listening comprehension evaluations. Additionally, the study serves as a methodological reference for future research in language testing, particularly within the Vietnamese educational context.

2. LITERATURE REVIEW

2.1. Language testing

Listening plays a crucial role in the process of second language acquisition. Hence, the assessment of listening skills serves as an essential step to measure second language learners' communicative ability. However, Field claims that proper assessment of listening represents an extremely difficult task because existing theories and frameworks regarding listening are inadequate [4]. It is then emphasized by many researchers that there should be more studies to investigate in depth the learning and assessing listening skills [5, 6].

In the domains of language testing and language evaluation, reliability and validity issues matter significantly because they function as fundamental elements. The concept of reliability, according to Fulcher and Davidson, can be defined as "the degree to which a test consistently and precisely gauges the same underlying construct over time, across test forms, and/or within a single test, ensuring dependable and trustworthy results" (p. 30-32) [7]. The study by Shang, Aryadoust and Hou states that effective language tests must present consistent evaluation outcomes under different assessment conditions for proper test takers' proficiency evaluation. Shang et al. mention that unreliable tests produce random scoring results that might generate errors in determining the test-takers' performance evaluation [8]. Meanwhile, validity, according to the American Educational Research Association et al., refers to "the degree to which evidence and theory support the interpretations of test scores for the proposed use of tests" (p. 11) [9]. According to Chapelle, validity has traditionally been understood as the degree to which a test can measure accurately what it claims or purports to be measuring [10]. Validating a test means that language testers need to examine three types of evidence, including criterion-oriented validity, content validity, and construct validity [11]. When examining criterion-oriented validity, the tester is interested in computing the correlation between the results of a test and the results of other measures of the same criterion. Content validity can be identified by having experts judge the degree to which the test item is a representative sample from the domain that is to be tested. Construct validity is to examine the relationship between the performance in a test and the ability which is intended to be measured.

2.2. Previous studies

The review by Peng and Yuan found that prior studies mainly researched English listening proficiency evaluations among university students though researchers tend to focus more on national assessment than regional or institutional testing [12]. Zhao presented multiple perspectives on validity and reliability principles as they relate to language learning and education. He argues that present-day language assessment methods show a specific inclination for reliability yet advises they should instead focus on validity and work on maximizing it to the highest practical levels [13]. The song examines internal and

external construct validity and presents the conceptual meaning of different validity forms of evidence through theoretical investigation [14].

A thorough evaluation of listening examinations needs to be conducted at the school-based level. Modern academic studies about institutional assessments focus primarily on testing English major academic outcomes and evaluating listening tests through criterion-referenced language tests (CRTs). Jiang and Feng researched self-constructed English examinations designed by teachers while proposing nine essential questions related to proposition development, examination execution and management practices [15]. The research of Huang begins with an assessment diagnosis study for English proficiency tests at the college level where scores were evaluated between English diagnostic and final exams. Her research demonstrates college English standardized testing requires implementation because it can be successfully implemented [16].

Research initiatives explore the difficulties linked to the assessment of listening skills as the last component in this study. Scholars have identified several critical issues: a deficiency in authentic materials, wherein the authenticity of English language resources is insufficient, exemplified by a lack of titles and instructions that adversely impacts the validity of listening assessments [17]; a failure to incorporate diverse question formats [17, 18]; the predominance of multiple-choice questions in English listening assessments, which lacks adequate construct validity to effectively evaluate students' listening competencies; an insufficient focus on school-based assessments and classroom evaluations, raising concerns regarding the quality of the questions presented [19].

In this study, the authors analyzed data by using SPSS to investigate the reliability and validity of the listening progress test from a statistical perspective. SPSS (Statistical Package for the Social Sciences) provides descriptive statistics which present test-taker performance through mean and median values and mode calculations. Standard deviation acts as a statistic that determines score variability to show how test-takers distribute their performance results [20]. Liu et al. applied the assessment framework of Bachman and Palmer and analysed data by SPSS to identify problems of a listing final test based on the test results of 20 students and analysis of Cronbach's Alpha value, correlation coefficient, etc [21]. However, the number of participants in that study was still insufficient for frequency distribution. As a result, this current study, with more participants, is expected to provide a new insight into the literature gap.

3. METHODOLOGY

3.1. Participants

The study involved 105 third-year English major students from a Vietnamese public university. These students participated in a blended learning program that combined both online and in-class components. The online component focused on vocabulary acquisition and listening strategies, supplemented by various

Table 1. Demographic Breakdown of Participants

	Characteristics	Frequency	Percent	Valid Percent	Cumulative Percent
Class Code	20241FL6038002	25	23.8	23.8	23.8
	20231FL6038003	28	26.7	26.7	50.5
	20231FL6038004	27	25.7	25.7	76.2
	20241FL6038001	25	23.8	23.8	100
Gender	Female	84	80	80	80
	Male	21	20	20	100
Cohort	2023-2024	50	47.6	47.6	47.6
	2022-2023	55	52.4	52.4	100
Total		105	100	100	

Table 2. Types of items

	Section 1		Section 2		Section 3
Types	**MCQ**	**Short answer**	**Matching**	**Basic fill-in-the-blank**	**Advanced fill-in-the-blank**
Item	1,2,3,4,5,6,7	8,9,10	11,12,13,14,15,16	17,18,19,20	21,22,23,24,25,26,27,28,29,30
Numbers of items	7	3	6	4	10

exercises. Meanwhile, the in-class component emphasized practical listening skills, providing a balanced approach to language learning.

The participants were divided into four class codes: 20241FL6038002, 20231FL6038003, 20231FL6038004, and 20241FL6038001, with 25, 28, 27, and 25 students respectively. This distribution ensured a diverse representation of the student body. Gender distribution among the participants was 80% female (84 students) and 20% male (21 students), reflecting the typical gender ratio in language studies at the university.

Additionally, the participants were from two different cohorts: 2023-2024 and 2022-2023, with 50 students (47.6%) and 55 students (52.4%) respectively. This mix of cohorts provided a comprehensive dataset for analysing the effectiveness of the listening comprehension test, as it included students with varying levels of exposure to the blended learning program. This diverse group of participants offered valuable insights into the reliability and validity of the test, contributing to the overall goal of improving language assessment practices.

3.2. Research Design

The research employed a quantitative approach, utilizing SPSS statistical software to analyse the test data. The primary objective was to evaluate the reliability and validity of a listening comprehension test.

3.3. Data Collection

Data was collected through a progress test administered to the participants. The progress test is for the course "Listening Skills 5" at a public university in Vietnam. It is designed for 5th-semester English language students who have completed previous listening skills courses. The test aims to evaluate students' ability to understand main ideas and important details in relatively long and complex spoken texts on four topics (entertainment, technology, culture, and psychology).

The assessment is divided into three sections, each containing 10 questions, making a total of 30 questions. Each part of the test involves listening to a conversation, lecture, or discussion and answering questions in various formats, including fill-in-the-blank, short answer, multiple choice, and matching. Students listen to the audio twice and have a total of 45 minutes to complete the test. The types of items are shown in Table 2.

The questions are designed to assess students' listening comprehension at a B2 level, focusing on their ability to grasp key points and detailed information. The test is scored out of 30 points, with each correct answer worth one point. The final score is then converted to a 10-point scale for grading purposes.

The test was conducted under standardized conditions to ensure consistency and fairness. It took place in the classroom with minimal distractions, and high-quality audio equipment was used to ensure clarity. Clear instructions were provided, and the test was precisely timed. Standardized answer sheets were used. The test was reviewed by multiple instructors for clarity and accuracy.

After the test, responses were recorded and prepared for statistical analysis using SPSS version 26.0. Reliability and validity were assessed using established criteria, ensuring the test's consistency and accuracy. Ethical considerations included obtaining informed consent from all participants and ensuring the confidentiality of their data. The data analysis process followed a systematic approach: data entry, cleaning, descriptive statistics, reliability analysis, and inferential statistics, ensuring a thorough and accurate evaluation of the test results.

3.4. Data Analysis

The data analysis for this study was conducted using SPSS, focusing on several key areas to ensure a comprehensive evaluation of the listening comprehension test. The analysis included the following components:

3.4.1. Descriptive Statistics

Mean: Calculated to understand the central tendencies of the test scores, providing an average score for the test-takers.

Standard Deviation: Used to assess the variability of the scores, offering insights into the spread and dispersion of test-taker performance.

3.4.2. Reliability Analysis

Cronbach's Alpha: Employed to evaluate the internal consistency of the test items, ensuring that all items measure the same underlying construct.

Corrected Item-Total Correlation: Analysed to determine the correlation between each item and the total score, further validating the consistency of the test items.

3.4.3. Validity Analysis

Construct Validity: Statistical methods were applied to confirm that the test accurately measures the theoretical construct it was intended to assess. This included examining internal correlations and ensuring that the test components aligned with the overall construct.

By employing these statistical techniques, the study aimed to provide a thorough evaluation of the listening comprehension test, ensuring its reliability and validity. The findings from this analysis are intended to improve the quality of the test and offer valuable insights for language teaching and assessment practices.

4. RESULTS AND DISCUSSION

4.1. Descriptive Statistics Analysis

The descriptive statistics provide an overview of the test scores, including measures of central tendency and variability. The mean score of the listening comprehension test was calculated to be 7.2322, indicating the average performance of the students. The mean score suggests that, on average, students are performing fairly well. The standard deviation was 1.20970, reflecting the spread of the scores around the mean. A standard deviation of 1.20970 suggests that the scores are relatively close to the mean, but there is still some variability. The scores vary by 5.67 points, showing some diversity in performance. The range and standard deviation indicate that while most students' scores are close to the average, there is still a noticeable spread in the scores, meaning some students are performing significantly better or worse than others.

The skewness of the distribution was -0.148 with a standard error of 0.236, indicating a slight left skew (Table 3). This suggests that the distribution of scores is slightly skewed to the left, meaning there are a few lower scores pulling the mean down. The kurtosis was -0.562 with a standard error of 0.467, indicating a relatively flat distribution compared to a normal distribution. This suggests that the scores are more evenly spread out with fewer extreme values.

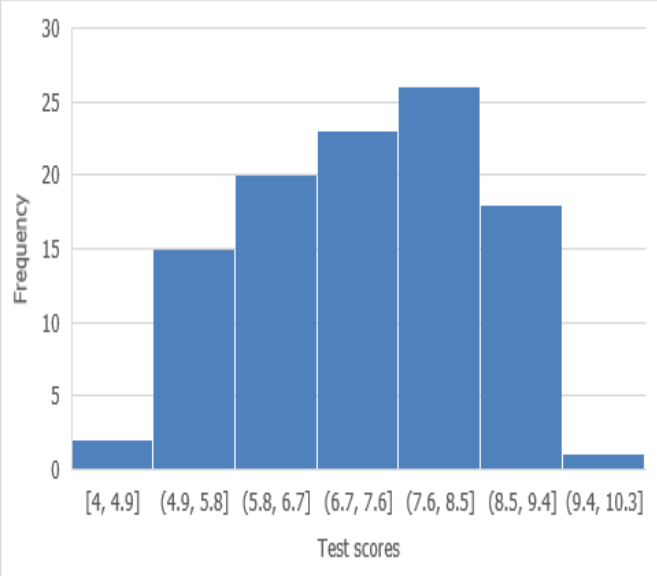

Figure 1. Histogram of Test Scores

To better understand the distribution of scores, a histogram (Figure 1) was created. The histogram shows that the majority of students scored between 6 and 8, with fewer students scoring at the extremes. Additionally, a box plot (Figure 2) was used to identify any outliers and to visualize the interquartile range.

Table 3. Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Skewness	Kurtosis
Points	105	5.67	4.00	9.67	7.2322	1.20970	-0.148	-0.562
Valid N (listwise)	105							

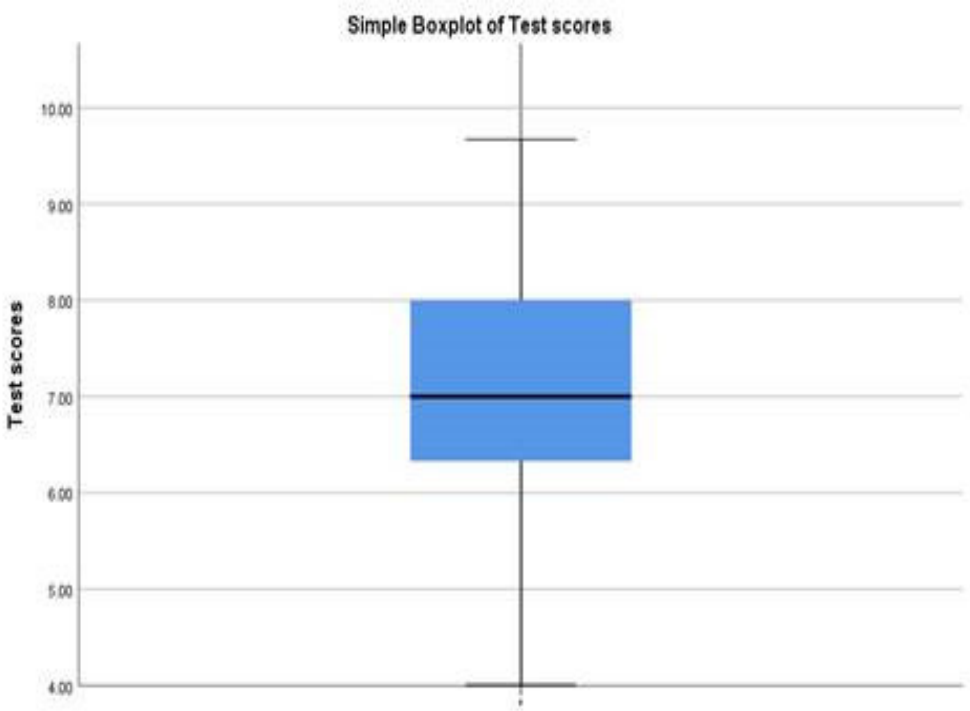

Figure 2. Box Plot of Test Scores

4.2. The analysis of reliability

The reliability of the listening comprehension test was assessed using Cronbach's Alpha. Table 4 shows that Cronbach's Alpha coefficient was 0.671, suggesting that the test items have a moderate level of internal consistency.

Table 4. Reliability Statistics

Cronbach's Alpha	N of Items
0.671	30

Table 5 provides detailed item-total statistics, including the scale mean if an item is deleted, scale variance if an item is deleted, corrected item-total correlation, and Cronbach's Alpha if an item is deleted. The Item-Total Statistics table provides crucial insights into the performance of individual test items and their contribution to the overall reliability of the test. Here's a detailed breakdown of the key components:

Scale Mean if Item Deleted: This column shows the mean score of the test if a particular item is removed. It helps in understanding how each item affects the overall test score. For example, if item 1 is deleted, the scale mean is 21.1048.

Scale Variance if Item Deleted: This column indicates the variance of the test scores if a specific item is removed. Variance measures the spread of scores. A lower variance suggests that the scores are more consistent. For instance, the variance if item 1 is deleted is 11.652.

Corrected Item-Total Correlation: This statistic measures the correlation between the score on an individual item and the total score on the test (excluding the item itself). Higher values indicate that the item is consistent with the overall test. For example, item 1 has a corrected item-total correlation of 0.379, suggesting a moderate positive relationship with the total score.

Cronbach's Alpha if Item Deleted: This column shows the Cronbach's Alpha coefficient if a particular item is removed from the test. It helps identify items that may be negatively impacting the test's reliability. For instance, if item 1 is deleted, Cronbach's Alpha is 0.646, slightly lower than the overall alpha of 0.671, indicating that item 1 contributes positively to the test's reliability.

Table 5. Item-Total Statistics

Items	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
1	21.1048	11.652	0.379	0.646
2	21.1429	12.143	0.225	0.663
3	20.8381	12.406	0.261	0.66
4	20.819	12.938	0.054	0.674
5	21.0381	12.191	0.227	0.662
6	20.8095	12.425	0.288	0.659
7	20.8095	13.406	-0.142	0.686
8	21.2571	11.501	0.422	0.642
9	21.0381	11.883	0.324	0.652

10	20.7333	13.024	0.082	0.671
11	20.7429	12.885	0.159	0.668
12	21	11.712	0.395	0.646
13	21.0095	11.721	0.387	0.646
14	20.7429	12.847	0.184	0.667
15	20.819	12.361	0.303	0.657
16	20.8667	12.097	0.355	0.652
17	21.2381	12.164	0.218	0.663
18	20.9333	12.428	0.187	0.666
19	21.0857	13.079	-0.041	0.689
20	21.581	12.65	0.186	0.665
21	20.8571	12.604	0.165	0.667
22	20.8857	12.218	0.291	0.657
23	21.3714	12.274	0.206	0.664
24	20.9714	12.663	0.097	0.674
25	20.7143	12.975	0.183	0.668
26	20.7143	13.071	0.086	0.671
27	20.7143	13.245	-0.088	0.676
28	21.0952	11.741	0.353	0.649
29	20.8857	12.679	0.121	0.671
30	21.3429	12.554	0.115	0.673

Evaluation of Weak Items

Item 2 (What does Joey say about roller skating? A. He mastered the moves relatively quickly. B. He learnt how to do it, especially for the movie. C. He couldn't get used to wearing old-fashioned skates.): The corrected item-total correlation is 0.225, which is below the acceptable threshold of 0.3. Additionally, removing this item increases the overall Cronbach's Alpha to 0.663, indicating it may be weakening the scale.

Item 4 (When asked about his co-star, Joey says that A. he appreciated the help she gave him. B. He disliked her telling him what to do. C. He found her rather unfriendly.): The corrected item-total correlation is 0.054, suggesting it does not align well with the other items. Removing this item increases the overall Cronbach's Alpha to 0.674.

Item 7 (How does Joey feel about the future? A. He'd like to concentrate on acting work. B. He's keen to go back to being a rap performer. C. He thinks he's too young to have definite plans.): This item has a negative corrected item-total correlation (-0.142), indicating it is not contributing positively to the overall reliability. Removing this item increases the overall Cronbach's Alpha to 0.686.

Item 19 (According to Jen, surfing is different from sports such as tennis and golf in which hardly any participants get a lot of ___ from the sport.): The corrected item-total correlation is -0.041, and removing this item increases the overall Cronbach's Alpha to 0.689, suggesting it is weakening the scale.

Recommendations for Improvement: The following revisions can be made to improve the clarity and alignment of the questions with the content of the audio.

Item 2

Original: What does Joey say about roller-skating?

A. He mastered the moves relatively quickly.

B. He learnt how to do it, especially for the movie.

C. He couldn't get used to wearing old-fashioned skates.

Revised: What does Joey say about roller-skating in the movie?

A. He mastered the moves relatively quickly.

B. He learnt how to do it, especially for the movie.

C. He had to adjust to using old-fashioned skates.

Reason: The revision clarifies that Joey had to adjust to using old-fashioned skates, which is more specific and directly related to the audio content where Joey mentions the difference between inline skates and old-fashioned skates.

Item 3

Original: Because the film was set in the past, Joey had to:

A. wear clothes that didn't suit him.

B. talk in a way that made him laugh.

C. follow the instructions of acting coaches.

Revised: Because the film was set in the 1970s, Joey had to:

A. wear clothes that didn't suit him.

B. talk in a way that made him laugh.

C. follow the instructions of acting coaches.

Reason: The revision specifies the time period (the 1970s) to provide more context and make the question clearer, as mentioned in the transcript.

Item 4

Original: When asked about his co-star, Joey says that:

A. He appreciated the help she gave him.

B. He disliked her telling him what to do.

C. He found her rather unfriendly.

Reason: The original question was already clear and aligned with the audio content, so no changes can be made at this point. Further investigation needs to be done.

Item 19

Original: According to Jen, surfing is different from sports such as tennis and golf in which hardly any participants get a lot of _____ from the sport.

Revised: According to Jen, surfing is different from sports such as tennis and golf because only a handful of participants _____ from the sport.

Reason: The revision makes the question more specific and directly related to the audio content, which mentions that only a handful of surfers ever "strike it rich" compared to sports like tennis and golf. This should help improve the item's alignment with the overall test and its reliability.

Implications for Test Improvement: Items with low or negative corrected item-total correlations (e.g., items 7 and 19) should be reviewed and potentially revised or removed to improve the overall reliability of the test. Items with high corrected item-total correlations (e.g., item 8) are performing well and contribute positively to the test's reliability. The overall Cronbach's Alpha of 0.671 suggests moderate internal consistency, indicating that while the test is generally reliable, there is room for improvement by addressing the problematic items. By carefully analysing these statistics, educators and test developers can make informed decisions to enhance the quality and reliability of the listening comprehension test.

4.3. The analysis of validity

Construct validity can be assessed through several methods, including internal correlation, factor analysis, and the multi-traits multi-methods (MTMM) approach. This paper primarily emphasizes internal correlation, which can be broken down into three specific types:

1. Correlation Between Different Components: The correlation between different components of the test should be relatively low, ideally ranging from 0.3 to 0.5. This ensures that each component measures a distinct aspect of the construct.

2. Correlation Between Tasks Within a Component: The correlation between two tasks within the same testing component should be high, typically at least 0.5 to 0.7. This indicates that the tasks are consistently measuring the same underlying construct.

3. Correlation with the Total Score: Each component should have a high correlation coefficient with the total score, generally above 0.7. This demonstrates that each component significantly contributes to the overall measurement of the construct.

Based on Table 6, there are specific areas where the correlation values should ideally be higher or lower to improve the test's validity and reliability. For the correlation between different components, such as Section 1, Section 2, and Section 3, the ideal range is 0.3 to 0.5. The current values, like 0.209 between Section 1 and Section 2 and 0.300 between Section 2 and Section 3, are within the acceptable range but on the lower side. Slight adjustments might be needed to ensure these correlations consistently fall within the ideal range, ensuring each section measures distinct aspects of listening comprehension.

For the correlation between tasks within a component, the ideal range is 0.5 to 0.7. However, current values such as 0.371 between MCQ and Short Answer, 0.008 between Matching and Basic Fill-in-the-Blank, and 0.047 between Short Answer and Matching are lower than ideal. This indicates that the tasks within these components are not consistently measuring the same underlying construct. To improve this, ensure that both types of questions are aligned in terms of content and difficulty level, and revise items to ensure they assess similar skills, such as detailed comprehension or inference skills.

Regarding the correlation with the total score, each component should ideally have a value above 0.7. While Section 1 (0.730) and Section 2 (0.719) have strong correlations with the total score, Section 3 (0.689) is slightly below the ideal threshold. To improve this, the

Table 6. Correlation

	Test scores	MCQ	Short answer	Matching	Basic Fill-in-the-blank	Advanced Fill-in-the-blank	Section 1	Section 2	Section 3
Test scores	1	0.602**	0.622**	0.571**	0.449**	0.689**	**0.730****	**0.719****	**0.689****
MCQ	0.602**	1	**0.371****	0.198*	0.070	0.094	0.915**	0.195*	0.094
Short answer	0.622**	**0.371****	1	0.047	0.160	0.496**	0.705**	0.134	0.496**
Matching	0.571**	0.198*	0.047	1	**0.008**	0.161	0.177	0.844**	0.161
Fill-in-the-blank 1	0.449**	0.070	0.160	**0.008**	1	0.309**	0.118	0.540**	0.309**
Fill-in-the-blank 2	0.689**	0.094	0.496**	0.161	0.309**	1	0.289**	0.300**	1.000**
Section 1	**0.730****	0.915**	0.705**	0.177	0.118	0.289**	1	**0.209***	**0.289****
Section 2	**0.719****	0.195*	0.134	0.844**	0.540**	0.300**	**0.209***	1	**0.300****
Section 3	**0.689****	0.094	0.496**	0.161	0.309**	1.000**	**0.289****	**0.300****	1

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

items in Section 3 should be reviewed and possibly revised to ensure they effectively contribute to the overall assessment. This could involve ensuring that the questions are clear, directly related to the audio content, and appropriately challenging.

4.4. Discussion

These findings suggest that while the overall performance is satisfactory, there is room for improvement, particularly for students at the lower end of the score distribution. Educators might consider providing additional support and targeted interventions for these students to help close the performance gap. This could include differentiated instruction, supplementary listening practice, and formative assessments to monitor progress.

To improve the reliability of the test, it is recommended to review and potentially revise or remove items with low or negative corrected item-total correlations. For example, item 7 could be rephrased to better align with the test construct, or it could be replaced with a new item that more accurately measures the intended skill. Additionally, pilot testing revised items with a small group of students can help ensure that the changes improve the test's reliability.

Improving the reliability of the test will lead to more consistent and accurate assessments of students' listening comprehension skills. This, in turn, will provide more reliable data for educators to make informed decisions about student performance and instructional strategies. Reliable assessments are crucial for identifying students' strengths and areas for improvement, thereby supporting effective teaching and learning.

To enhance the construct validity of the test, ensure that each section and item type accurately measures distinct aspects of listening comprehension. Adjust correlations between different components to fall within the ideal range of 0.3 to 0.5, confirming that each section assesses different skills. For tasks within a component, improve internal consistency by aligning question types in terms of content and difficulty, aiming for correlations between 0.5 and 0.7. This ensures that tasks within each section consistently measure the same underlying construct.

Lastly, ensure each component has a high correlation with the total score (above 0.7). Review and revise items in sections with lower correlations to ensure they contribute effectively to the overall assessment. This will enhance the test's validity, providing a more accurate measure of students' listening skills.

Improving the construct validity of the test will ensure that it accurately measures the intended listening comprehension skills. This will provide more meaningful and actio nable data for educators to support student learning and improve instructional

practices. Valid assessments are essential for providing accurate feedback to students and guiding instructional decisions.

The findings of this study align with previous research on language assessment. For instance, Liu and Diao highlighted the importance of well-designed test items for reliable assessments, which is consistent with the current study's emphasis on revising or removing unreliable items [22]. Similarly, Shang, Aryadoust, and Hou emphasized the need for consistent and accurate assessments to evaluate language proficiency effectively [8].

Furthermore, the study's focus on construct validity aligns with the work of Chapelle, who stressed the importance of validating language tests to ensure they accurately measure the intended constructs [10].

5. CONCLUSION

This study aimed to evaluate the reliability and validity of a newly developed listening comprehension test for assessing students' English proficiency. Through a rigorous analysis, the findings indicate that the test demonstrates quite strong psychometric properties. The moderate level of Cronbach's alpha coefficient suggests that the internal consistency in measuring students' listening comprehension skills can be improved by addressing the problematic items. The construct validity of the test was supported through correlation analyses, demonstrating alignment with established theoretical frameworks in language assessment. The findings suggest that with some modifications, the test can provide a more accurate measure of listening comprehension.

Despite these promising results, the study acknowledges certain limitations, such as the sample size and the specific linguistic background of participants, which may affect generalizability. Future research could expand the participant pool and employ longitudinal studies to further validate the test's effectiveness over time. Moreover, incorporating qualitative feedback from test-takers and educators could provide deeper insights into test usability and fairness.

In conclusion, this study contributes to the ongoing effort to develop reliable and valid assessment tools for English listening comprehension. By ensuring rigorous test design and validation, educators and policymakers can enhance language assessment practices; therefore, language learning outcomes can be improved.

REFERENCES

[1]. Brown H. D., *Language assessment: Principles and classroom practices*. Pearson Education, 2004.

[2]. Rost M., *Teaching and researching listening (2nd ed.)*. Pearson Education, 2011.

[3]. Bachman L. F., Palmer A. S., *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press, 2010.

[4]. Field, J., *Cognitive validity: The interface between cognitive psychology and language testing*. Cambridge University Press, 2013.

[5]. Aryadoust V., "A cognitive diagnostic assessment of listening comprehension: A model comparison study," *Language Testing*, 35(4), 501–527, 2018. https://doi.org/10.1177/0265532217716733

[6]. Buck G., *Assessing listening*. Cambridge University Press, 2001.

https://doi.org/10.1017/CBO9780511732959

[7]. Fulcher G., Davidson F., *Language testing and assessment: An advanced resource book*. Routledge, 2007.

[8]. Shang, Y., Aryadoust, V., & Hou, L., "Assessing the reliability of language proficiency tests: A review of current models", *Language Testing Asia*, 14(1), 22-34, 2024.

[9]. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

[10]. Chapelle C. A., *Validity argument in language testing*. Cambridge University Press, 2012.

[11]. Cronbach L. J., Meehl P. E., "Construct validity in psychological tests," *Psychological Bulletin*, 52(4), 281–302, 1995

[12]. Peng S., Yuan Q., "A review of research on college English listening assessment in China (2005–2015)," *Foreign Language World*, (6), 86-93, 2015.

[13]. Zhao Y., "A study on the balance between reliability and validity in language testing," *Foreign Language Testing and Teaching*, (1), 45-49, 2000.

[14]. Song M. "The concept and development of construct validity," *Journal of Language Teaching and Research*, 2(3), 558-562, 2011. https://doi.org/10.4304/jltr.2.3.558-562

[15]. Jiang L., Feng X. "An exploration into college English teacher-made testing," *Foreign Language World*, (3), 62–68, 2003.

[16]. Huang P., "An analysis of the diagnostic function of CET listening tests," *Foreign Language World*, (4), 54–57, 2001.

[17]. Niu H. "A study of the authenticity of English listening test materials," *Foreign Language Testing and Teaching*, (4), 15-18, 2001.

[18]. Wang Y., "A comparative study of question types in listening comprehension tests," *Foreign Language Testing and Teaching*, (2), 20-25, 2004.

[19]. Qian W., "Issues in listening assessment in college English testing," *Foreign Language Testing and Teaching*, (2), 12-15, 2004.

[20]. Pallant J., *SPSS survival manual: A step-by-step guide to data analysis using IBM SPSS (7th ed.)*. Open University Press, 2020.

[21]. Liu Y., Zhang H., Zhou J., "An empirical study on the validity and reliability of a university listening test," *Foreign Language Testing and Teaching*, 3, 42–50, 2020.

[22]. Liu X., Diao J., "The importance of well-designed test items in language assessment," *Journal of Language Testing*, 38(2), 123-145, 2020.

THÔNG TIN TÁC GIẢ

Trần Thị Tuyết Trinh[1], Phạm Thị Hồng[1], Nguyễn Ngọc Quỳnh[2], Nguyễn Phương Thảo[2]

[1]Trường Ngoại ngữ Du Lịch, Trường Đại học Công nghiệp Hà Nội

[2]Khoa Ngoại ngữ, Trường Đại học Thăng Long