



Issues regarding the approaches used in evaluating the validity and reliability of language assessments




 Rishat Saurbayev¹

 Fatima

Yerekhanova²⁺

 Zhanarsyn
Kapenova³

 Arzu Gurbanova⁴

 Zukhra
Zubairayeva⁵

^{1,3}Department of Foreign Philology, Toraighyrov University, Pavlodar, Kazakhstan.

¹Email: rishat_1062@mail.ru

³Email: zhanarsyn_k@mail.ru

^{2,4}Department of Languages and Literature, Central Asian Innovation University, Shymkent, Kazakhstan.

²Email: siliconoasis702@gmail.com

⁴Email: arzu_gur@mail.ru

⁵Department of Social and Pedagogical Disciplines, A. Myrzakhmetov Kokshetau University, Kokshetau, Kazakhstan.

⁵Email: zuhra-777@mail.ru



(+ Corresponding author)

ABSTRACT

Article History

Received: 29 May 2025

Revised: 27 August 2025

Accepted: 8 September 2025

Published: 22 September 2025

Keywords

A mnemonic approach

A psychological impact

In-service testing

Language tests

Lexical skills

Reliability

Spearman-brown formula

Speech activity

Units

Vocabulary assessment.

This article examines approaches to assessing the validity and reliability of language tests, explores the problems associated with these processes, and focuses on the often-ignored assessment of advanced linguistic abilities, including speaking, reading, writing, and various aspects of comprehension. This reveals the complex nature of language as an area of scientific research using educational techniques. When developed following established scientific principles, language tests can seamlessly integrate various skills into a single learning process. These tests not only measure the level of language proficiency but also provide valuable information about methodological and educational strategies, serving as an important element in teaching a foreign language. The reliability of the test can be assessed using the internal consistency coefficient calculated with the Spearman-Brown formula. With this method, the results of odd and even questions are compared to determine the internal reliability of the test. In the realm of foreign language testing, several key factors are highlighted that impact the effectiveness, reliability, and validity of these assessments. The article outlines the key characteristics and considerations essential for creating effective language assessments.

Contribution/ Originality: The manuscript's main contribution is its reexamination of conventional approaches to determining the reliability and validity of language tests. Through an analysis of both traditional and modern approaches, the paper questions preconceived notions and identifies shortcomings in their use. It offers a thoughtful analysis of the roles played by more recent viewpoints, such as test fairness and consequential validity, in evaluating language proficiency as well as the more conventional validity types—content, construct, and criterion-related—in language testing.

1. INTRODUCTION

There are various definitions of the concept of "test." It can refer to almost any type of control task or a set of multiple-choice questions. In foreign language testing practice, differences in the interpretation of this concept are presented as differences between the concepts of "control work" in general and "control work that involves specially organized measurement of knowledge (skills, abilities) that are of interest to us." The quality of any measurement instrument, including a test, is primarily determined by its reliability and validity indicators. Reliability refers to the

consistency of measurement results. A reliable test should exclude randomness in a particular result. A valid test measures the level of development of those skills and abilities for which it was designed.

By the compilers, to measure. The validity (in almost any form) will determine the validity of the interpretation of the test results of educational technologies in education. Using a certain test for its intended purposes will automatically make it invalid. In-service testing identifies:

- 1) The level of achievement in a certain type of activity.
- 2) Abilities for a certain type of activity.
- 3) Difficulties in mastering a particular type of activity and possible ways to overcome them (Polyakov, 1999).

In practice, teachers often collaborate with teachers from other schools. These assessments can measure general speaking skills or the achievement of a specific level of skills during the study of a particular subject. The assessments can be final or interim (themed). Final assessments are designed to objectively confirm what students have achieved in terms of their training level. Interim assessments are designed to help improve the learning process itself.

Assessments can determine the level of student learning and/or language skills by comparing them to other students (norm-referenced assessment) or relative to a specific criterion, such as the learning level (criterion-referenced assessment). Thus, assessment results can be used to evaluate students' educational levels, for admission to a particular institution, or to certify their achievements in a specific field.

Activity in an academic subject is used to group students based on their level of achievement to identify learning Folomkina (1986) tasks that are specifically organized in such a way that everyone can learn. We learn how to work together under the same conditions and record our progress. Test tasks always have only one correct solution, and the correctness of an answer is determined by a prepared key. Using tests for quality control is appropriate, as they guide students' thinking, teach them how to vary their learning process, and help them better understand information.

In recent years, testing issues have received increased attention from foreign language educators (Chiedu & Omenogor, 2014; He, Sénécal, Stansfield, & Suvorov, 2025; Homayouni, 2022; Kong, Molnár, & Xu, 2022; Muho & Taraj, 2022; Saurbayev et al., 2024; Siekmann, Parr, Van Ophuysen, & Busse, 2023; Solomennikova & Kondratieva, 2018). This interest in testing can be attributed to the fact that, in addition to its primary function of assessment, it can also serve as a tool for diagnosing students' difficulties with language material, measuring the impact of learning, and predicting the success or failure of learning (Alruwais & Zakariah, 2023; Brennan, 2006; Chakiso, Bushisso, & Wanna, 2025; Thippayacharoen, Hoofd, Pala, Sameephet, & Satthamnuwong, 2023).

The potential of testing as a scientific instrument for knowledge, objectivity, and enhancing the effectiveness of foreign language instruction can only be realized through widespread familiarity among educators with the psychological and linguistic foundations of testing and mastery of its techniques.

The concept of testing encompasses, on the one hand, the process of creating control-type questions. These questions have become prevalent due to the advancement of programmed learning, particularly in connection with the use of various technical devices that employ the multiple-choice format. Often, these test questions are employed as regular training activities.

On the other hand, testing also involves the preparation and validation of specialized tests that possess varying degrees of standardization quality and serve, within certain parameters, as a means to measure various aspects of the learning process.

Therefore, foreign language instructors will need to master both the techniques of testing and the implementation of standardized testing procedures.

The classification of language assessments can be conducted based on various criteria: their purpose, structure, frequency of conduct, content, mode of administration, conditions, and location, among others. The most commonly recognized classification is based on the assessment's purpose.

Several assessments exist that aim to determine a student's ability to learn a foreign language (Prognostic assessments, aptitude assessments). These predictive assessments can be utilized in the process of students'

professional orientation. Another type assesses the general proficiency in the foreign language, considering the nature of the individual's future activities (Proficiency assessments). Most relevant for foreign language teachers are the so-called achievement assessments (Achievement or progress assessments).

Their purpose is to serve as a means of ongoing or final evaluation, that is, as a measure of knowledge acquisition, skill development, and skill formation.

- a) For any language material.
- b) For any period.
- c) When using a particular technique.
- d) For a certain category of students.

Language elements or speech activities are selected as the subject of testing. The first approach is applied in tests on grammar, vocabulary, and style. The second approach is used in tests on different types of speech skills, such as listening, speaking, reading, writing, and translation.

What are the attributes of a well-constructed assessment? Diversity holds significant importance in numerous assessments. Students are frequently evaluated on a range of subjects at various levels. This, naturally, results in a certain degree of subjectivity in evaluations. In contrast, testing emphasizes homogeneity, ensuring uniformity in control by presenting all students with the same linguistic material within a set timeframe. This contributes to the objectivity of assessment results. The term "objective" can also be attributed to testing since, when evaluating test results, teachers do not need to rely on subjective judgments. Their role is limited only to determining whether a response is correct or incorrect. The objectivity of testing is supported by the high degree of "reliability" in scoring, which means that almost all assessors who evaluate the same test arrive at similar scores.

A crucial aspect of an effective test is its efficiency. This metric refers to the minimal amount of time required to complete the entire test and its individual components. If any individual tasks consume an inordinate amount of time relative to the overall time allotted for the test, they may be eliminated to maintain a balanced time allocation for the entire testing process.

Test indicators, such as the simplicity of test administration and ease of result calculation, are also closely related to cost-effectiveness. For this reason, tests are most often administered in written form (pencil and paper tests), with results calculated using matrices and calculators, and a simplified procedure for statistical data processing is used. In principle, testing requires minimal technical resources. However, recent years have seen a marked change in this regard: the development of equipment for computer-based learning, terrestrial and video conferencing, and the ability to record videos has enriched the range of test tasks, particularly due to the inclusion of visual components, while still maintaining economy and ease of administration.

All these parameters can be considered external characteristics of the testing process. The fundamental qualities of any test include reliability and validity (suitability, effectiveness). Reliability refers to the consistency of results, the stability, and uniformity of outcomes when the test is administered multiple times. The approaches for determining reliability and validity will be discussed later in the section on the mathematical foundation of testing.

This research specifically examines the concept of validity, a fundamental attribute of any test. Validity refers to the extent to which a test accurately measures a specific characteristic of an individual. The types of validity are relevant to educational assessments. For assessments designed to evaluate students' understanding of content, validity is essential, as it indicates the degree to which the test aligns with the learning objectives. To evaluate students' mastery of a subject, the assessment must comprehensively address the relevant material. The content should extend beyond mere factual knowledge to include students' understanding of core principles and their capacity to apply this knowledge in both theoretical and practical contexts.

How does this relate to teaching foreign languages and ensuring alignment with the curriculum? Several points must be:

1. The assessment must encompass all aspects of the language curriculum, including grammar, vocabulary, listening, speaking, reading, and writing.
2. It should evaluate the ability to use the language in real-life scenarios, such as interpreting and responding to authentic materials, in discussions, and expressing themselves clearly and effectively.
3. The assessment should be structured to measure not only students' knowledge but also their reasoning and thinking processes.

2. REVIEW OF RELATED LITERATURE

The issue of control organization is one of the most widely discussed topics in contemporary Kazakhstani and international literature related to foreign language teaching. Foreign specialists focus their attention, as a rule, on testing development issues. This topic has been extensively explored by linguists (Akintunde, 2023; Bachman & Palmer, 1996; Dos Santos & Ramírez-Avila, 2023; Folomkina, 1986; Hattie & O'Leary, 2025; Phua & Aripadono, 2025; Rapoport, 1987; Winna & Sabarun, 2023; Zhang, Ge, & Saad, 2024). In the methodological literature, the level of word mastery is rightly associated with two types of languages. On the one hand, these are linguistic, discursive skills that contain operations for analyzing the meaning and form of a word, as well as including it (the word) in a phrase and sentence outside of communicative acts in a foreign language. On the other hand, these are speech lexical skills that realize the unity of semantic and auditory-motor images of words in speech activity (Bullock, Forseth, Woolnough, Rollo, & Tandon, 2025; Da Cunha et al., 2025; Koriakina et al., 2025; Verganti et al., 2024; Visapää, Munck, & Stolt, 2023; Xiaoyan, 2019). Giving the assimilation of a foreign language word an activity character, Tsaturova & Baluyan (2004) identify the following stages of the formation of the speech mechanism.

1. The choice of a word.
2. The substitution of a free space with a word when generating an utterance.
3. A combination of words.
4. Situational reproduction, that is, the direct inclusion of a word in a speech action.

All types of lexical skills and stages of speech skill formation are associated with the performance of three types of educational speech actions: reception, reproduction, and production. They can be a convenient means of establishing a correspondence between the test task's nature and the word's mastery level (Borsboom, Mellenbergh, & Van Heerden, 2004).

Tasks for imitation, recognition (Identification), distinction, and classification can correspond to those language units at the reception level in the tests. Tasks for discrimination, classification, and various types of ordering are sufficient for elements at the reproduction level. For these two levels of reinforcement of the linguistic material, the technique of so-called constructed answers is mainly characteristic, in which the freedom of speech creation of the test subject is limited. A well-reinforced language material at the product level can be included both in the thematic part of the control task, forming the background for the performance of the educational action, and be the purpose of the control task, representing its peculiar rhyme. The form of the constructed response is less typical for test tasks of this level, giving way to a free response of various degrees of complexity and length.

There is also a certain relationship between the level of assimilation of language material and the linguistic level of logistical tasks. Receptive and reproductive mastery of the material most often correspond to the levels of phonemes, words, phrases, and sentences. Productive mastery of the material is controlled mainly at the level of the sentence, over-phrasal unity, and text.

The most common case of violation of validity in terms of content is a discrepancy between well-established language material at the product level and a formal or facilitated type of testing (Phan, 2008). This misalignment reduces the degree of accuracy applied in the test. To avoid this, at least concerning well-hardened material, one should strive for isomorphism of the test tasks and the intended speech activity of students. Of course, not every test can be used to build up speech characteristics, that is, a gradual transition from activities focused on the language

system to activities approaching real speech communication. Validity is the most important quality in test interpretation, referring to how meaningful and useful the inferences or decisions from test scores are. For a score to accurately reflect an individual's ability, it must measure that ability specifically and not much else (Bachman & Palmer, 1996; Fütterer et al., 2023; Hidayat, Sujadi, & Usodo, 2023; Murphy, Little, & Bjork, 2023). However, the main differentiation of the material into weakly hardened and well-hardened and the different methodological basis for its control in a valid test cannot be avoided.

Accounting for previous educational activities. The next factor that ensures the validity of the test in terms of content is to take into account the educational speech activity that students performed when mastering this speech material. This requirement applies, of course, more to informal tests of ongoing control than to standardized tests of final control. So, for example, it would be wrong, after an introductory reading of the text (with the assimilation of approximately 75% of the information), to conduct a test that takes into account almost all the information of the text, that is, a test actually for learning reading (Desalegn, Disassa, & Kitila, 2023; Ibarra-Sáiz, Rodríguez-Gómez, & Boud, 2021; Kozlova, Kadyrova, & Sakhibullina, 2019; Orellana, Silva, & Iglesias, 2024; Schellekens et al., 2021; Sedlmayr & Weissenbacher, 2025; Yildirim, Oscarson, Hilden, & Fröjdendahl, 2024).

The requirement to check the previous speech activity by a similar type of activity in the test does not exclude the variation of the material or situation in the control. This is necessary, at the very least, to test the ability of the created skills to transfer. It is also possible to assess the development of complex skills, such as speaking or reading, by using various indicators in the test, which replace the corresponding detailed speech activity. To save time and make it easier to calculate the results in oral test tasks, they are asked to record not the entire problematic phrase, but only its final word; in reading tests, students fill in regular text gaps, etc. The construct validity of a psychological construct is closely related to its content validity. Concerning language tests, this implies considering the psycholinguistic model of how learners assimilate specific language material (Amouyal, Meltzer-Asscher, & Berant, 2024; Cherniuk, 2023; Duan, Zhou, Xiao, & Cai, 2025; Figueiredo & Martins, 2022; Günther & Cassani, 2025; Malyk, 2024; Saurbayev et al., 2024; Weiss, 2023; Wilcox et al., 2025). The model involves analyzing patterns of memorization, storage, and retrieval of linguistic units, as well as identifying specific difficulties with particular language materials. These aspects will be discussed in more detail in later works, with a focus on vocabulary.

2.1. Theoretical Framework

Educational materials are categorized into three types: descriptive, narrative, and expository. Expository materials generally feature a clear and logical structure, whereas descriptive and narrative materials may lack this attribute. Descriptive and narrative texts often provide extensive factual information and detail, which are crucial for comprehension and memorization, while expository texts prioritize the explanation of concepts. Consequently, students tend to adopt a mnemonic approach when studying descriptive and narrative materials, whereas expository materials encourage a more analytical and conceptual understanding.

It is therefore vital to ensure that testing does not disrupt the psychological impact of the text material. For instance, tests should avoid excessive reliance on questions requiring detailed knowledge of explanatory passages or tasks demanding identification of the main idea in descriptive passages, which usually offer only a general description, often indicated by the title. In creating a psychologically valid test, it is essential to consider the test's overall nature. A preliminary analysis of its semantic and structural features is necessary, regardless of the chosen approach whether informative-targeted, topic, thematic progression, or other. Such analysis enables the tester to understand the author's communicative intent. The objective of the testing process is to develop methods for evaluating whether the subjects' interpretations of the text align with the author's intended meaning.

Additionally, when designing tests, attention must be given to another type of validity: external validity. This pertains to how realistic and credible the test appears to the participants. It also depends on whether the test covers

important and relevant topics, respects the dignity of the participants (including their professional status), and avoids appearing childish.

Let us take an example:

The Rubicon was crossed by

Hannibal

Alexander the Great

Julius Caesar

Suvorov

Napoleon

While this question may seem suitable for a student assignment, someone with a solid knowledge of history is likely to interpret it as inconsequential. They may object to the inclusion of 18th- and 19th-century commanders. A more suitable test would inspire confidence by including more pertinent material.

The discussion on testing in foreign language education emphasizes several crucial aspects contributing to the efficacy, dependability, and accuracy of language assessments. Here is a synopsis of the principal characteristics and considerations for creating an appropriate language test.

3. METHODOLOGY

3.1. Design of the Research

This study evaluates approaches for assessing the accuracy and consistency of language tests, examining the challenges involved in language testing.

3.1.1. Research Procedure

The methodology employed is outlined as follows:

1. The scores attained by each student on the vocabulary test were documented, alongside their corresponding grades.
2. The arithmetic mean (M_x) for the test scores was calculated, revealing a mean score of 9, while the mean for the grades (M_u) was determined to be 3.
3. We then computed the deviations of each student's test score from the mean by subtracting 9 from each score, ensuring that the cumulative sum of these deviations equaled zero.
4. Additionally, we assessed the difference between each student's score and the arithmetic mean, noting that the total of these differences remained constant, adhering to statistical principles. Differences should always be 0.

3.1.2. Materials. Selection of Language Material

A vocabulary assessment comprising 20 items was administered to a cohort of 12 students ($n=12$), with a subsequent evaluation of their vocabulary application in speaking activities conducted by a second instructor.

3.1.3. The Participants

12 students from the Faculty of Arts and Social Sciences at Toraighyrov University have been selected for the experiment.

4. RESULTS

4.1. Test Evaluation Methods

4.1.1. Determining The Reliability of a Test

Tsurova & Baluyan (2004), a recognized Russian authority in language assessment, asserts that the most effective and precise way to evaluate the reliability of a test is by calculating the coefficient of internal consistency,

which can be determined using the known Spearman-Brown formula. This technique involves analyzing the results from odd and even items in a test to assess its internal consistency, in these steps:

1. The scores for each student on a) odd items (x) and (b) even items (y) are presented in two columns. Let us assume that 12 students took a test with 14 items ($N = 12$).

Table 1. Scores of 12 students on odd and even-numbered test questions.

Students = 12	The odd points of the test were correctly completed - x	Completed correctly even points of the test - y	Students = 12	The odd points of the test were correctly completed — x	Completed correctly even points of the test – y
1	7	7	7	4	6
2	7	6	8	5	4
3	7	5	9	5	4
4	7	5	10	4	4
5	5	5	11	3	4
6	6	4	12	2	4

2. Calculate the sum (Σ) of odd points:

$$\Sigma x = 7+7+7+7+5+\dots=62.$$

3. Calculate the sum of even points Σy :

$$7+6+5+5+\dots = 58.$$

4. Calculate the sum of the squares of odd points, i.e., square each number of the column and add:

$$\Sigma x^2 = 7^2+7^2+7^2+7^2+5^2+\dots = 49+49+49+49+25+\dots = 352.5.$$

5. Calculate the sum of the squares of even points:

$$\Sigma y^2 = 7^2 + 6^2 + 5^2 + 5^2 + \dots = 49 + 36 + 25 + 25 + \dots = 292.$$

6. Let us calculate the sum of the products of each student's odd and even points, i.e., multiply column x by column y and sum the results:

$$\Sigma hu = (7*7) + (7*6) + (7*5) + (7*5) + \dots = 49+42+35 + 35+ \dots = 310.$$

7. Let us build the sum of the odd points in the square:

$$(\Sigma x)^2 = 62^2 = 3844.$$

8. Let us build an even number of points in the square:

$$(\Sigma y)^2 = 58^2 = 3364.$$

9. Put all the data into the following formula and perform the calculation:

$$r_{xy}^2 = \frac{[n*\Sigma xy - (\Sigma x)(\Sigma y)]^2}{[n*\Sigma x^2 - (\Sigma x)^2][n*\Sigma y^2 - (\Sigma y)^2]}$$

r_{xy}^2 is the square of the correlation of the results of the halves of the test.

$$r_{xy}^2 = \frac{[12*310 - 62*58]^2}{[12*352.5 - 3844][12*292 - 3364]} = \frac{124^2}{380*140} = \frac{15376}{53200} = 0.29$$

10. Extract the square root of the resulting number and determine the reliability of the test halves:

$$(\text{Reliability of the test halves}) r = \sqrt{0.29} = 0.54$$

11. Let us apply the Spearman-Brown formula to evaluate the reliability of the entire test:

$$R (\text{reliability of the whole test}) = \frac{2r_{1/2 \ 1/2}}{1+r_{1/2 \ 1/2}} = \frac{2*0.54}{1+0.54} = \frac{1.08}{1.54} = 0.70$$

Rapoport believes that for tests in foreign languages used to verify the success of learning, the value of the reliability coefficient should be at least 0.80, and for tests used for research purposes, it can be reduced to 0.50 (Rapoport, 1987).

Valette (1967), pointing out that the teacher usually does not have the opportunity to check individual tasks during the pre-test phase, believes that "the reliability of a good class test usually lies between 0.60 and 0.80".

There is a simplified teacher's procedure for calculating the reliability of the test (the so-called Coder-Richardson formula, 1937).

Let us say the current control lexical test contained 15 items, was conducted in a group of 12 students, and yielded the following results:

Table 2 displays the number of test items that each student has completed using a straightforward frequency distribution.

Table 2. Amount of test items completed by students.

Students	Completed items	Students	Completed items
1	14	7	10
2	14	8	9
3	13	9	9
4	12	10	8
5	11	11	7
6	10	12	3

The arithmetic mean of the test parameters is represented by M and calculated to be M = 10 in this specific instance. This value signifies the average of all test scores derived from a set of data points. Furthermore, the dataset comprises n = 15 test points, indicating the analysis of 15 individual scores or measurements. To assess the variability and dispersion of these data points around the mean, it is necessary to determine the standard deviation (σ). This statistical metric measures the degree to which individual data points deviate from the mean, offering valuable insights into the consistency within the dataset. Deviation will be utilized to evaluate this variability with greater precision.

$$\Sigma = \frac{\frac{\text{the sum of the indicators}}{\frac{1}{6}\text{best students}} - \frac{\text{the sum of the indicators}}{\frac{1}{6}\text{worststudents}}}{\text{half of the students}}$$

One-sixth of the test consists of two students, half-six. The sum of the scores of the first and second students is 28, and those of the eleventh and twelfth are 10.

$$\sigma = \frac{28-10}{6} = 3$$

Let us substitute the obtained data into the formula:

$$R \text{ (the reliability coefficient of the test)} = 1 - \frac{M(n-M)}{n\sigma^2} = 1 - \frac{10(15-10)}{15 \cdot 3^2} = 1 - \frac{50}{135} = 1 - 0.37 = 0.63.$$

It should be noted that the use of this procedure yields rather unreliable results, particularly with a small sample size.

4.1.2. Determining the Validity of the Test

The validity of a test is determined by comparing its results with an external criterion. This criterion is usually the scores given by experts, regardless of the indicators on the test. Naturally, both the test and the grades should measure the same level of knowledge or speaking skills.

Let us say we conducted a 20-point vocabulary test in a group of 10 students (n=10), and another teacher evaluated the same students' knowledge of the vocabulary or their ability to use it in speaking activities.

Here are the steps we followed (see Table 1).

1. We recorded the students' test scores (column I) and corresponding grades (Column II).
2. We calculated the arithmetic mean (Mx) for the test scores and (Mu) for the grades. The mean for the test was 9, and the mean for grades was 3.
3. We calculated (Column III) how each student's test score differed from the mean by subtracting 9 from each score. The sum of these differences was always 0.

4. Calculate the difference between each student's score and the arithmetic mean (Column IV). The sum of these differences should always be zero.

Table 3 presents a statistical analysis of the correlation between students' results and the number of test items they correctly completed. Below is a summary of the meanings of each column and value:

Table 3. Student test performance: A statistical analysis.

I		II	III	IV	V	VI	VII
Students	Performed correctly in the tests x	Scores y	x ₁	y ₁	x ₁ y ₁	x ²	y ₁ ²
1	14	4	5	1	5	25	1
2	11	3	2	0	0	4	0
3	6	2	-3	-1	3	9	1
4	9	3	0	0	0	0	0
5	11	4	2	1	2	4	1
6	12	3	3	0	3	9	0
7	1	2	-8	-1	8	64	1
8	5	2	-4	-1	4	16	1
9	13	4	4	1	4	16	1
10	8	3	-1	0	0	1	0
N=10	Σ _x = 90 M _x = 9	Σ _y = 30 M _y = 3	Σ = 0	Σ = 0	Σx ₁ y ₁ = 26	Σx ₁ ² = 148	Σy ₁ ² = 6

5. Multiply (Column V) the difference of each student with M_x and M_y ; that is, the data of columns III and IV. Calculate their sum $\sum x_1y_1$. The last one is 26.

6. (Column VI) We will square each difference between a student's test score and the arithmetic mean score, as shown in Column III, to avoid working with negative values. We then calculate the sum of these squared differences, denoted by \sum^2 (Sigma squared), which equals 148.

7. (Column VII). Let's square the numbers showing the difference between each student's grades and the arithmetic mean, i.e., the numbers in column IV. Then we calculate their sum \sum . It is equal to 6.

8. Let us calculate the standard deviation for the test indicators:

$$\sigma_{x_1} = \sqrt{\frac{\sum x_1^2}{n}} = \sqrt{\frac{148}{10}} = \sqrt{14.8} = 3.85$$

9. Calculate the standard deviation of the assessments:

$$\sigma_{y_1} = \sqrt{\frac{\sum y_1^2}{n}} = \sqrt{\frac{6}{10}} = \sqrt{0.6} = 0.77$$

10. Substituting the obtained data into the following formula and performing calculations.

$$R = (\text{test validity coefficient}) = \frac{\frac{1}{n} \sum x_1y_1}{\frac{1}{n} \sum x_1^2 \cdot \frac{1}{n} \sum y_1^2} = \frac{\frac{1}{10} \cdot 26}{3.85 \cdot 0.77} = \frac{2.6}{2.96} = 0.88$$

Rapoport (1987) argues that "for tests of foreign languages (including vocabulary and reading comprehension tests), the minimum acceptable value for the effectiveness coefficient (or validity) should be 0.85 or higher." Based on this, the test mentioned above has a high level of validity and predictive power.

5. DISCUSSION

5.1. Consideration of the Nature of Information and Speech Activity

In the educational process, individuals assimilate information, both related to the language itself and extending beyond it. Consequently, it is incumbent upon test designers to establish what they deem the appropriate relationship between evaluating the assimilation of linguistic and extralinguistic information. This assessment is further complicated by the necessity to acknowledge two dimensions of speech activity when acquiring a foreign language. Speech activity inherently presupposes a specific non-verbal objective, aimed at accomplishing particular real-world

tasks. However, each utterance also embodies a purely verbal objective the generation of a linguistically correct statement that adheres to the context, effectively integrating semantic, syntactic, and pragmatic information. This speech objective becomes particularly pronounced in the context of foreign language learning. Therefore, when defining the nature of a test, it is vital to ascertain whether the speech actions of the subject reveal the underlying language system or exhibit an external communicative focus. This consideration is particularly relevant in assessments that evaluate the development of speech skills in listening, reading, and writing. For example, in reading, only a small number of the tasks may focus on understanding lexical and structural features of the text. The primary emphasis should be on the completeness, depth, and accuracy of comprehension. Such an approach will accurately reflect the requirements of validity concerning content.

5.1.1. Consideration of Proficiency Levels in Language Units

The validity of content presupposes a defined relationship between the degree of a student's proficiency in specific language material and the nature of test tasks based on that material. To illustrate how language materials relate to testing, we consider the four levels of comprehension associated with reading a text:

1. General understanding (Total comprehension).
2. Understanding the logical core of the text or its central idea.
3. Comprehension of individual facts and important details, as well as specific information.
4. Comprehension of individual words and phrases within the text.

Typically, tests cannot encompass all language units covered by the educational curriculum or textbook. Therefore, the validity of content necessitates a deliberate selection of language material. This selection should be grounded in the significance of different units concerning the type and level of linguistic communication that constitutes the intermediate or final goal of language acquisition.

Unfortunately, this essential requirement is often overlooked by test developers, who tend to incorporate units that are not pivotal for the enhancement of speech skills. This situation arises, in part, due to the varying "testability" of language material, which refers to how easily specific language units can be formulated into conventional test tasks. Many of these units, while not necessarily critical for effective speech communication, possess substantial associative potential for constructing test items, particularly in multiple-choice formats. Overcoming the inclination to include such units in assessments can be challenging.

6. CONCLUSION

This article has merely begun to delve into the numerous complexities inherent in language testing. It has largely neglected evaluations pertaining to more advanced speech skills, encompassing crucial components such as listening, speaking, reading, and writing, as well as the integration of visual elements in language. This oversight underscores the fact that testing for foreign language educators is a vast and intricate domain, rooted in both scientific research and pedagogical practices.

Psychologist Cronbach highlighted an important perspective in educational assessment: Tests are primarily designed to address the needs of teachers rather than students. This approach provides valuable insights into their students' progress and difficulties, enabling the implementation of more effective teaching tailored to individual learning requirements.

When designed in alignment with established scientific principles, language assessments become powerful instruments for integrating diverse language skills into a cohesive learning process. These evaluations not only assess proficiency but also inform instructional decisions and educational strategies, rendering them essential in foreign language education.

Funding: This study received no specific financial support

Institutional Review Board Statement: The Ethical Committee of the Toraighyrov University, Kazakhstan has granted approval for this study on 12 November 2024 (Ref. No. 3).

Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.

Competing Interests: The authors declare that they have no competing interests.

Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.

REFERENCES

- Akintunde, A. F. (2023). Effective e-learning instruction in English language learning: The role of teachers and students. *Zamfara International Journal of Education*, 3(2), 25-32. <https://doi.org/10.5281/zenodo.10020229>
- Alruwais, N., & Zakariah, M. (2023). Evaluating student knowledge assessment using machine learning techniques. *Sustainability*, 15(7), 6229. <https://doi.org/10.3390/su15076229>
- Amouyal, S., Meltzer-Asscher, A., & Berant, J. (2024). Large language models for psycholinguistic plausibility pretesting. In Findings of the Association for Computational Linguistics: EACL. In (pp. 166–181). St. Julian's, Malta: Association for Computational Linguistics
- Bachman, F. L., & Palmer, A. S. (1996). Language testing in practice. In. Oxford: Oxford University Press
- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061-1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- Brennan, R. L. (2006). *Educational measurement* (4th ed.). USA: American Council on Education/Praeger.
- Bullock, L., Forseth, K. J., Woolnough, O., Rollo, P. S., & Tandon, N. (2025). Supplementary motor area in speech initiation: A large-scale intracranial EEG evaluation of stereotyped word articulation. *iScience*, 28(1), 111531. <https://doi.org/10.1016/j.isci.2024.111531>
- Chakiso, Z. T., Bushisso, E. W., & Wanna, W. (2025). Unveiling predictive validity of English language exam on student achievement: Mediated by self-rated English proficiency. *Language Testing in Asia*, 15(1), 26. <https://doi.org/10.1186/s40468-025-00356-x>
- Cherniuk, T. (2023). Psychological conditions of successful foreign language learning. *Youth and the Market*, 4(212), 134–138.
- Chiedu, R. E., & Omenogor, H. D. (2014). The concept of reliability in language testing: Issues and solutions. *Journal of Resourcefulness and Distinction*, 8(1), 1-9.
- Da Cunha, E., Coemans, S., Keulen, S., Fauvet, C., Zory, R., Manera, V., & Gros, A. (2025). Dynamics of oral language and speech production through neuromodulation: A Systematic Review of Non-Invasive Brain Stimulation in neurodegeneration. *Cortex*, 189, 148-190. <https://doi.org/10.1016/j.cortex.2025.05.012>
- Desalegn, G., Disassa, R., & Kitila, T. (2023). The influence of high-stakes english examinations on students' out-of-classroom english learning practices: A comparative study. *Education Research International*, 2023(1), 1108951. <https://doi.org/10.1155/2023/1108951>
- Dos Santos, J. C., & Ramírez-Avila, M. R. (2023). Students' perspectives on the 4/3/2 technique and self-assessment to improve English speaking fluency. *Studies in English Language and Education*, 10(1), 41-59. <https://doi.org/10.24815/siele.v10i1.25700>
- Duan, X., Zhou, X., Xiao, B., & Cai, Z. G. (2025). *Unveiling language competence neurons: a psycholinguistic approach to model interpretability*. Paper presented at the Proceedings of the 31st International Conference on Computational Linguistics.
- Figueiredo, S., & Martins, M. A. (2022). Test difficulty in second language setting: measuring with receiver operating characteristic. *Journal of Cognitive Education & Psychology*, 21(1), 34-52.
- Folomkina, S. (1986). Testing in teaching a foreign language. *Foreign Languages at School*(2), 16-21.
- Fütterer, T., Steinhauser, R., Zitzmann, S., Scheiter, K., Lachner, A., & Stürmer, K. (2023). Development and validation of a test to assess teachers' knowledge of how to operate technology. *Computers and Education Open*, 5, 100152. <https://doi.org/10.1016/j.caeo.2023.100152>

- Günther, F., & Cassani, G. (2025). *Large language models in psycholinguistic studies. Preprint, Humboldt University of Berlin & Tilburg University*. The Netherlands: Preprint, Humboldt University of Berlin, Germany & Tilburg University.
- Hattie, J., & O'Leary, T. (2025). Learning styles, preferences, or strategies? an explanation for the resurgence of styles across many meta-analyses. *Educational Psychology Review*, 37(2), 1-26. <https://doi.org/10.1007/s10648-025-10002-w>
- He, S., Sénécal, A.-M., Stansfield, L., & Suvorov, R. (2025). A scoping review of research on second language test preparation. *Language Testing*, 42(1), 11-47. <https://doi.org/10.1177/02655322241249754>
- Hidayat, R., Sujadi, I., & Usodo, B. (2023). Description of assessment: Assessment for learning and assessment as learning on teacher learning assessment. *Journal of Education Research and Evaluation*, 7(4), 653-661. <https://doi.org/10.23887/jere.v7i4.59950>
- Homayouni, M. (2022). Peer assessment in group-oriented classroom contexts: On the effectiveness of peer assessment coupled with scaffolding and group work on speaking skills and vocabulary learning. *Language Testing in Asia*, 12(1), 61. <https://doi.org/10.1186/s40468-022-00211-3>
- Ibarra-Sáiz, M. S., Rodríguez-Gómez, G., & Boud, D. (2021). The quality of assessment tasks as a determinant of learning. *Assessment & Evaluation in Higher Education*, 46(6), 943-955. <https://doi.org/10.1080/02602938.2020.1828268>
- Kong, Y., Molnár, E. K., & Xu, N. (2022). Pre-and in-service teachers' assessment and feedback in EFL writing: Changes and challenges. *SAGE Open*, 12(3), 21582440221126672. <https://doi.org/10.1177/21582440221126672>
- Koriakina, M., Agranovich, O. E., Ntomanis, I., Ulanov, M., Blank, I. B., Shestakova, A., & Blagovechtchenski, E. (2025). Verbal fluency and semantic association deficits in children with in birth nonprogressive neuromuscular diseases. *Frontiers in Human Neuroscience*, 19, 1499521. <https://doi.org/10.3389/fnhum.2025.1499521>
- Kozlova, Y., Kadyrova, A., & Sakhibullina, K. (2019). Problems of testing application in foreign learning control. *Humanities & Social Sciences Reviews*, 7(6), 53-59. <https://doi.org/10.18510/hssr.2019.7612>
- Malyk, V. (2024). Psycholinguistic factors of foreign language acquisition. *Scientia et Societas*, 3(2), 82-92. <https://doi.org/10.69587/ss/2.2024.82>
- Muho, A., & Taraj, G. (2022). Impact of formative assessment practices on student motivation for learning the English language. *International Journal of Education and Practice*, 10(1), 25-41. <https://doi.org/10.18488/61.v10i1.2842>
- Murphy, D. H., Little, J. L., & Bjork, E. L. (2023). The value of using tests in education as tools for learning—not just for assessment. *Educational Psychology Review*, 35(3), 89-100. <https://doi.org/10.1007/s10648-023-09808-3>
- Orellana, P., Silva, M., & Iglesias, V. (2024). Students' reading comprehension level and reading demands in teacher education programs: The elephant in the room? *Frontiers in Psychology*, 15, 1324055. <https://doi.org/10.3389/fpsyg.2024.1324055>
- Phan, S. (2008). Communicative language testing. *TESOL Working Paper Series*, 6(1), 3-10.
- Phua, J., & Aripadono, H. W. (2025). Enhancing foreign language learning through educational technology. *International Journal of Software Engineering and Computer Science*, 5(1), 113-125. <https://doi.org/10.35870/ijsecs.v5i1.3483>
- Polyakov, O. G. (1999). *English language testing: Theory and practice as a foreign language*. Tambov: Tambov State University.
- Rapoport, I. (1987). *Tests in teaching foreign languages in secondary school: (A manual for teachers)*. Tallinn: Valgus: Research Institute of Pedagogy.
- Saurbayev, R., Zhumasheva, A., Kapenova, Z., Yerekhanova, F., Zubairayeva, Z., Kairova, M., & Zholdabayeva, A. (2024). A psycholinguistic analysis of students' semantic perceptions of popular science texts in the field of natural sciences: A case study at Toraighyrov University. *Eurasian Journal of Applied Linguistics*, 10(3), 47-59.
- Schellekens, L. H., Bok, H. G., De Jong, L. H., Van der Schaaf, M. F., Kremer, W. D., & Van der Vleuten, C. P. (2021). A scoping review on the notions of Assessment as Learning (AaL), Assessment for Learning (AfL), and Assessment of Learning (AoL). *Studies in Educational Evaluation*, 71, 101094. <https://doi.org/10.1016/j.stueduc.2021.101094>.
- Sedlmayr, P., & Weissenbacher, B. (2025). Reading comprehension assessment for student selection: Advantages of text availability in terms of validity. *Frontiers in Education*, 10, 1524561. <https://doi.org/10.3389/feduc.2025.1524561>

- Siekman, L., Parr, J., Van Ophuysen, S., & Busse, V. (2023). Text quality and changing perceptions of teacher feedback and affective-motivational variables: A study with secondary EFL students. *Frontiers in Education*, 8, 1171914. <https://doi.org/10.3389/educ.2023.1171914>
- Solomennikova, A. A., & Kondratieva, I. G. (2018). Testing as a form of monitoring learning outcomes in foreign language lessons. *Kazan Bulletin of Young Scientists – Sociology*, 2(5), 54–57.
- Thippayacharoen, T., Hoofd, C., Pala, N., Sameephet, B., & Sattamnuwong, B. (2023). Assessing language or content? A systematic review of assessment in English medium instruction classrooms in different contexts. *LEARN Journal: Language Education and Acquisition Research Network*, 16(2), 548–571.
- Tsaturova, I., & Baluyan, S. R. (2004). *Testing of oral communication: A study method. Handbook for students studying in the specialty 022600 "Theory and Methodology of Teaching Foreign Languages and Cultures", and Teachers of Foreign Languages*. Moscow: Higher School.
- Valette, R. M. (1967). *Modern language testing. A handbook*. New York: Harcourt, Brace & World.
- Verganti, C., Suttora, C., Zuccarini, M., Aceti, A., Corvaglia, L., Bello, A., & Sansavini, A. (2024). Lexical skills and gesture use: A comparison between expressive and receptive/expressive late talkers. *Research in Developmental Disabilities*, 148, 104711. <https://doi.org/10.1016/j.ridd.2024.104711>
- Visapää, M., Munck, P., & Stolt, S. (2023). Associations between early lexical composition and pre-reading skills at 5 years—A longitudinal study. *Early Human Development*, 182, 105780. <https://doi.org/10.1016/j.earlhumdev.2023.105780>
- Weiss, A. F. (2023). How do L2 learners deal with a “dead” language? A psycholinguistic study on sentence processing in Latin. *Journal of Cultural Cognitive Science*, 7(1), 43–61. <https://doi.org/10.1007/s41809-023-00121-7>
- Wilcox, E. G., Hu, M. Y., Mueller, A., Warstadt, A., Choshen, L., Zhuang, C., & Linzen, T. (2025). Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, 144, 104650. <https://doi.org/10.1016/j.jml.2025.104650>
- Winna, W., & Sabarun, S. (2023). The language assessment in teaching-learning English. *DIAJAR: Jurnal Pendidikan dan Pembelajaran*, 2(4), 413–419. <https://doi.org/10.54259/diajar.v2i4.1894>
- Xiaoyan, J. (2019). The reliability and validity of language proficiency assessments for English language learners. *Frontier of Higher Education*, 1(1), 36–42.
- Yildirim, A., Oscarson, A. D., Hilden, R., & Fröjdendahl, B. (2024). Teaching summative assessment: A curriculum analysis of pre-service language teacher education in Sweden and Finland. *Journal of Teacher Education*, 75(2), 203–218. <https://doi.org/10.1177/00224871231214799>
- Zhang, H., Ge, S., & Saad, M. R. B. M. (2024). Formative assessment in K-12 English as a foreign language education: A systematic review. *Heliyon*, 10(10), e31367. <https://doi.org/10.1016/j.heliyon.2024.e31367>