International Journal of English Language and Literature Studies
ISSN(e): 2306-0646
ISSN(p): 2306-9910
DOI: 10.55493/5019.v13i2.5007
Vol. 13, No. 2, 201-215

URL: www.aessweb.com

The use of statistical methods in the system of monitoring the qualimetry of level language tests

Saurbayev Rishat Zhurkenovich[1+]
Omarov Nurlan Ramazanovich[2]
Tekzhanov, Kairat Muhamedhafizovich[3]
Zhetpisbay Aliya Kozhamuratkyzy[4]

[1]*Department of Foreign Philology, Toraighyrov University, Pavlodar, Kazakhstan.*
[1]*Email: rishat_1062@mail.ru*
[2,3,4]*Higher School of Humanities Margulan Pedagogical University, Pavlodar, Kazakhstan.*
[2]*Email: omardos@mail.ru*
[3]*Email: tekzhanov.kairat@mail.ru*
[4]*Email: a_kanzhygaly@mail.ru*


(+ Corresponding author)

Article History
Received: 6 November 2023
Revised: 3 January 2024
Accepted: 5 March 2024
Published: 15 March 2024

Keywords
CAE
Cross-section tests
FCE
Qualimetry
Statistical method
Test
TOEFL.

ABSTRACT

This paper examines the problem of using statistical methods in the quality monitoring system of level language tests based on TOEFL, FCE and CAE materials. The majority of teachers and experts in qualimetry in general and educational qualimetry in particular are inexperienced which makes the use of test diagnostics at all stages of the learning process extremely challenging. The proposed approach involves the use of test diagnostics at every stage of the learning process to ensure accurate results. The quality of test tasks can be determined and the reliability coefficient of the tests can be calculated by employing experimental verification data and probabilistic factors. The statistical analysis of the results allows for the identification of qualitative characteristics of the test and provides insights into the heterogeneity of the subjects. This approach enables the grouping of subjects into homogeneous groups and ensures the adequacy of tests in a specific educational environment. This research employs the control cross-sectional analysis technique to identify areas of deficiency in understanding, ascertain the educational attainment of students and gauge their requirements. This method creates strategies to fill in any knowledge gaps and provides personalized support. Statistical methods are used in this research to collect, assess and analyse both quantitative and qualitative data.

Contribution/ Originality: The contribution of the study lies in the use of statistical methods to evaluate the precision of test scores and assess the validity and reliability of test questions. The findings of this research can serve as a foundation for further studies aimed at creating a comprehensive testing management system.

1. INTRODUCTION

It is necessary to emphasize that the in-test parameters of the language test given in the work below can and should only be fully implemented in standardized tests or experimental cross-section tests that summarize a lengthy period of study and claim to be more or less extreme measurement accuracy given the circumstances that the test methodology for obtaining the most objective data requires correct and accurate application. It is not possible to completely comply with the standards that require the continuous use of mathematical equipment in tests intended for general use and cover material of insignificant volume.

Using scientific experience from the past is crucial for the optimal development of language testing in modern times. Today, both testologists and practical teachers are unaware of the findings of several scientific studies due to a variety of factors, creating a knowledge gap in the field of linguodidactics. At the same time, there is no doubt that the ability to navigate the vast material of these theoretical and practical studies is a factor in the professional maturity, scientific competence and outlook of specialists in the field of test measurements at any level.

Test qualimetry is a crucial concept in the study of test quality. It enables us to develop a methodology for complex quantitative assessment of an object's quality particularly when it comes to language tests. We can ensure they are effective and accurate in measuring language proficiency by focusing on the quality and reliability of language tests. We can use test qualimetry to accurately assess language test quality and make well-informed decisions regarding their application.

During the 1970s, the field of language testing underwent significant evolution. Researchers, including Damico, Oller, and Tetnowski (1999) proposed a transition from the principles of discrete testing established by Lado (1986) and other testologists in the 1960s (Spolsky, 1968) to the principles of integrative testing. The rationale behind integrative testing was to integrate knowledge of relevant linguistic aspects such as pronunciation, grammar and vocabulary with an understanding of context. Additionally, Hymes (1972) introduced the theory of communicative competence leading to the development of testing methods that focused on communication skills. Testologists finally started doing studies in this area as well despite early setbacks.

The competition between two different approaches to language testing ended with the triumph of the communicative direction. Damico et al. (1999) and its proponents made several interesting works but they did not align with the generally accepted communicative approach to teaching foreign languages. As a result, they were either rejected or ignored but they still remain a subject of interest for many testologists (Johnsen & Sulak, 2021) who continue to explore various facets of pragmatic tests and actively introduce them into their work. Closed tests are one of the most popular objects of their research. Testologists' belief that traditional close exams can be used to evaluate both individual skills and general language proficiency is demonstrated by studies by Oller (1973), Darnell (1970), Conrad (1971), Stubbs and Tucker (1974) and Brown (1980) etc. We confirmed this hypothesis by showing a fairly high correlation with such well-known test batteries as TOEFL, UCLES (University of Cambridge Local Examinations Syndicate), the English Placement Examination, the General Examination in English as a Foreign Language and the MLA cooperative examination. .

2. LITERATURE REVIEW

2.1. Theoretical Framework

Testing allows for a relatively short period of time to assess the effectiveness of students' cognitive activity, i.e. the degree and quality of achieving learning goals. It is an effective way to check the level of knowledge on a particular topic or for a course of study. Tests are just one form of control in the educational process in order to save time and deepen the survey process objectively. They should be used to complement traditional forms of control while partially replacing them as needed. Inaccuracies in test readings can be quickly detected by the teacher while working with students on a daily basis.

All types of tests are characterized by certain intra-test parameters that are calculated statistically. The entire test is evaluated by two main parameters: validity (suitability) and reliability (consistency). Let us examine the many interpretations of "validity" that psychologists and methodologists have both domestically and abroad in order to prevent misunderstandings. The term has several definitions in the Russian methodological literature on tests. American psychologists and methodologists interpret it as the degree to which the test measures the quality or phenomenon that it is intended to measure (Mabel & McKeithan, 2022).

The most crucial aspect of the test that demonstrates what it measures and how effectively it does it is its validity (Buntins, Buntins, & Eggert, 2017). In other words, it shows the degree of validity of the test for its

intended purpose. The concept of validity is directly related to the problem of choosing the right material for testing. Validity is a specific characteristic of the test and not a general property. If a grammatical construction test measures only the knowledge or skills of using a grammatical construction, then it is suitable for measuring grammar but not for measuring word knowledge or reading comprehension since it does not test vocabulary or reading comprehension.

If the purpose of the test is to evaluate students' ability to correctly pronounce individual sounds, then the test should only include material related to this and exercises that consider the level of vocabulary proficiency.

Questions about the validity of a test are about the extent to which a sound conclusion can be drawn based on test scores. This parameter is related to the validity of conclusions drawn from test scores or other forms of evaluation. Many questions related to this parameter can be reduced to two:

1. What conclusion can be drawn about what was measured by the test?
2. What conclusion can be drawn about other behavior that differs from what is observed during testing?

Based on the above, it can be concluded that the validity of a language test depends on the linguistic content of the test, its purpose and the situation or method used to test the content. When measuring the validity parameter, it is possible to obtain numerical values that serve as quantitative characteristics of this parameter of the language test. There are four main methods for determining the validity parameter of the test:

1. A method that studies the content of the test.
2. A method based on correlation analysis where the calculated correlation coefficient serves as a numerical characteristic of fitness.
3. A method based on various experimental procedures or expert evaluation procedures where a numerical characteristic of validity is not obtained but a reliable conclusion can be made about the validity of the test based on the results of the procedure used.
4. A method based on a series of theoretical and experimental studies where hypotheses about the forms of manifestation of the structure are formulated and empirically tested. The purpose of the study is not to obtain a numerical indicator characterizing the validity of the methodology but to be able to draw an appropriate conclusion.

Reliability is the next important parameter for checking the effectiveness of a language test. In broad terms, reliability is understood as the degree of efficiency and stability of the formation, functioning and development of objects, phenomena, processes and systems.

Reliability in psychology and pedagogy refers to the consistency of evaluation findings for a skill or aspect that is periodically examined or the consistency of comparison results obtained by assessing the same competence across multiple tests. In our opinion, Merlo and Stevenson (2001) defined it most briefly and precisely". We call reliability the constancy or stability of estimates obtained from repeated observations" (2001).

Thus, reliability is determined by the constancy with which the test performs its function as a measuring instrument concerning tests. Reliability is calculated by comparing digital data reflecting the results of several tests performed by the same students. The test is unreliable if a group provides it twice in a brief period of time when language classes are not conducted. If this is the case, the results should change slightly from one another.

The study conducted by Arruarte, Larrañaga, Arruarte, and A. (2021) presented a visual learning analytics tool as an evaluation test to assess the quality and reliability of educational resources. Similarly, Hashemi and Daneshfar (2018) conducted a study that aimed to provide a descriptive review of the International English Language Testing System (IELTS) by focusing on various aspects such as reliability, validity and washback. It is noteworthy that a test can be reliable without being valid but it cannot be valid without being reliable at the same time. The reliability of a test is a necessary condition for its validity because a test that produces inconsistent results does not measure anything at all.

The reliability indicator is the value of the reliability coefficient which shows how stable the test indicators are when the same test or its parallel version is repeated. It should be noted that no test is a completely reliable tool.

Therefore, when they talk about the stability of test indicators, they mean relative stability. The stability of test indicators can be influenced by factors such as an insufficient number of test tasks, the limited time allotted for the test or misunderstanding of the test subjects' instructions or test questions. A test's reliability increases with its diversity and ability to differentiate between tasks as well as with how well its measured functions cover the test questions.

Determining the reliability of a test is crucial to ensure that its results are stable and accurate. In contrast to the validity coefficient, reliability is calculated without the help of an external criterion. The reliability coefficient measures how consistent the test results are when the same test or its parallel versions are repeated. There are different ways to determine reliability but the simplest method is to re-apply the same test. However, this approach is limited as results may be affected by familiarity, memory, exercise ability and student variability. The identical test is retested in parallel allowing for the tasks' equal difficulty and the correlation coefficients between each task and the test in order to remove these undesired variables. This approach is regarded as the best even though it is tedious since it reduces the influence of unfavourable elements that could alter the dependability coefficient's true value.

Another way to determine the reliability of a test is to use one form of the test and calculate the correlation coefficient between the two halves of the same test, namely the even and odd tasks. Different ways of alternating tasks in each half are used to ensure that the tasks follow increasing difficulty. The correlation coefficient is determined by specific mathematical formulas and indicates the reliability of half of the test.

Two more parameters are used to characterize each test question: facility value (FV) and discrimination index apart from the reliability coefficient. The former shows how simple or complex each test question is and allows us to judge how adequate the task of the test is for a given language audience. The ideal difficulty coefficient is 50% which is when half of the subjects in a given language group give the correct answer. On the other hand, the discrimination index indicates how the test task distinguishes a well-prepared subject from a poorly prepared one. These are the main internal test parameters that a language test must satisfy to serve as an effective and objective tool for testing the knowledge, skills and abilities of the subjects.

Modern testologists such as Stecker and Fuchs (2000), Cherepanov and Shikhov (2008), Espin et al. (2009), Cho, Capin, Roberts, and Vaughn (2019), Johnsen and Sulak (2021) and Mabel and McKeithan (2022) are developing new methods of statistical analysis of test results such as the rush method which is currently the most popular among testologists in the ALTE group. Although this technique is not yet widely used in the processing of linguodidactics test data, it shows great potential for improving the accuracy and reliability of test results.

3. METHODOLOGY

The following method is employed in our study: the control cross-sectional analysis that is carried out to identify gaps in knowledge, determine the level of educational achievement of students, study the needs of students to provide individual support and plan further work to fill the knowledge gaps. The application of the statistical method involves the presentation of general questions, the collection, measurement and analysis of statistical (quantitative or qualitative) data.

3.1. Statistical Analysis of Tests

In the classical method of post-test analysis, the four above-mentioned parameters are used which are calculated statistically.

The entire test as a whole is evaluated by two parameters: reliability and validity. The reliability parameter shows to what extent the scores shown by the subjects are constant. In other words, if a group passes the same test twice within a short period during which language classes are not conducted, the results should differ little from each other.

There are a large number of ways to determine the reliability of the test. Some are better, others are worse. Let us consider a method that does not take much time and has no major drawbacks for calculating the Half-Reliability Coefficient (HRC). It is necessary to determine the correlation coefficient between the two halves of the test even and odd to calculate the reliability of the test.

To do this, you should:

1) Divide the test in half odd (No. 1, 2, 3, 5) in one group and even (No. 2, 4, 6, 8i) in the other one.
2) Enter the results in tables.
3) Count the scores in both groups.
4) Assign two ranks to each examinee: one in an odd test, the other in an even one.
5) The correlation coefficient between the two halves of the test (even and odd) is calculated by the formula:

$$p = 1 - (\frac{6 \Sigma d^2}{N(N^2-1)})$$

Where

p is the correlation coefficient.

d is the difference between the ranks.

N is the number of examinees.

This coefficient indicates what the reliability of the test will be if it is twice as short as this one.

The Spearman-Brown correlation formula (Spearman, Brown) must be used to ascertain the test's overall reliability. $HRC = \frac{2r_{hh}}{1+r_{hh}}$

Where

HRC: Reliability.

rhh : Correlation between the two halves of the test.

The second parameter of the test is its validity. This parameter determines the degree of validity and objectivity of the test as mentioned above. Validity shows to what extent the material tests exactly the aspect that it is intended to test, for example, perception and understanding of speech by ear and not knowledge of grammatical rules. This parameter is important mainly for the test compiler as it gives a picture of the distribution of quantitative characteristics of test results and the ability to assess how difficult or conversely, easy the test is and what should be done to direct it. The validity coefficient is calculated as the correlation coefficient between two samples of text squared.

Other factors that distinguish each question independently are used to process the collected data i.e. the difficulty coefficient, the Facility Value (FV) and the discriminant coefficient (discrimination index). The difficulty coefficient shows how adequate the test task is for this audience. In other words, the FV of a particular test assignment is the percentage of examinees who answered this task correctly. If there are 100 examinees in the group and 50 of them answered the question correctly for this assignment, then the FV of the assignment will be:

$$50/100 \ x \ 100\% \ = 50\%.$$

This simple calculation allows test compilers to quickly determine how simple or complex a particular test item is.

The discriminant coefficient shows how the test task distinguishes a well-prepared subject from a poorly prepared one. It can be assumed that the assignment is correct if the examinees with the best knowledge of the language give more correct answers than the examinees with the worst knowledge.

If the test takers with the best knowledge incorrectly answer the questions of the task while the weak one does it correctly, the question arises about the correctness of the compilation of this task.

First, it is necessary to determine the ranks of the examinees to calculate the discriminant coefficient. If we arrange the points received by the examinees in descending order and assign the first examinee No. 1, the second

No. 2, etc., then these numbers will be considered ranks. If two or more examinees scored the same number of points, then the rank is calculated as the average between these two or more ranks. For example, if the scores were 8, 6 and 6.5, then the ranks would be 1.0, 2.5, 2.5 and 4.0 respectively.

There are many ways to calculate the discriminant coefficient. The simplest method is one that requires ranking the examinees according to the number of points they received as a result of the test followed by comparing the number of correct answers in the best third of the test takers with the number of correct answers in the worst third of the test takers.

For example, if the best group consists of ten examinees and seven of them answered the question correctly (0.7) while only two out of ten people in the weak group (0.2) answered correctly then D.I. =0.7 – 0.2=+0.5. It is considered that a task with a D.I. = +0.5 discriminates well against test takers.

A task whose discriminant coefficient is equal to or tends to zero does not allow us to identify differences in the level of strong and weak examinees. The following formula is used to get the discriminant coefficient:

$$(RT\text{-}RB)/N$$

Where

RT is the number of correct answers in the upper group.

RB is the number of correct answers in the lower group.

NT is the number of subjects in the upper group.

Such an analysis of tasks is not suitable for subjective tests such as writing a resume, essay or interview but preliminary testing of tasks is still necessary.

Sometimes it seems appropriate to rank the examinees not only by the results of the test for understanding foreign languages by ear but also by the results of the entire test (including speaking, reading and writing).

The above two parameters show how applicable the test is to this group of subjects. It is necessary to analyze the entire test after analyzing the individual tasks of the test. The range and standard deviation which indicate how widely the scores are dispersed as well as the arithmetic mean, mode and median which illustrate how the test results are related to one another should be found using statistics. The arithmetic mean is the sum of all the points received by the examinees divided by the number of examinees.

$$M = \frac{\sum x}{N}$$

Where

M is the average arithmetic mean.

x is the point.

N is the total number of examinees.

The mode is the most common number of points. It is useful to determine the mode in cases where the test is too simple or difficult or when people with different levels of knowledge take the test.

The median is the average term of an ordered series. It is necessary to arrange all the scores in descending order and then find the average member of the series to determine the median. The calculation of the median is necessary for the completeness and adequacy of the assessment of the test. If only one person from a group of 10 people who have earned 8-10 points gets 1 point, then the arithmetic means will drop sharply due to this random score. The median is calculated to prevent an incorrect reflection of the actual picture.

However, none of the above parameters shows the breadth of the spread of examinees' scores. The range is the difference between the largest and smallest values.

The range is fairly accurately estimated by variance but this value has one drawback. It does not reflect the presence of gaps in the distribution of points. The indicator that takes into account each individual score will be the *standard deviation* (SD). This is a very important statistical parameter. It shows the difference between the student's score and the arithmetic mean of all scores.

The mean square deviation is the square root of the sum of the squares of the differences between the scores and the arithmetic mean divided by the number of examinees minus one.

$$SD = \sqrt{\frac{\sum(x - M^2)}{(N - 1)}}$$

Where

SD is the mean square deviation.

x is the point.

N is the number of examinees.

M is the arithmetic mean.

Making a table is required in order to determine the mean square deviation. a) In the column under x, it is necessary to write out the points in descending order.

b) Under $(x\text{-}M)$, the differences between the number of points and the arithmetic mean are written out. M must be accurate not rounded. The sum of all values $(x\text{-}M)$ must be 0.

c) Then the values $(x\text{-}M)$ are squared and written in the third column under $(x\text{-}M)^2$

d) All the values $(x\text{-}M)^2$ are added and it turns out $(x\text{-}M)^2$

e) Then, the SD is calculated.

The listed parameters show how well the test meets the specified goals. For example, they allow you to determine and whether the test is suitable for the level of complexity whether it can identify differences between individual examinees and characterize the overall level of knowledge.

The aforementioned information leads to the conclusion that the testing process which has essentially not yet evolved into a system depends critically on the data processing (i.e., parameter calculation) of the test. This is understood by both methodologists and researchers of the testing problem as evidenced by a large number of scientific papers and reports devoted to improvements in the system for determining criteria for evaluating test results. The issue of prompt, correct and objective assessment of test results is a problem that continues to be not fully solved since the number of correctly (or incorrectly) completed test tasks even with strict mathematical processing of the results is only one side of the problem of measuring the level of knowledge, skills and abilities. The quality of the work remains outside the framework of the template (stencil) to check the results. Therefore, it is necessary to search for such quantitative indicators that would correspond to strictly defined qualitative indicators. An in-depth review of the participants' normal test-taking errors is required. This study might serve as the foundation for the search for these quantitative markers.

Therefore, it is vital to take into account two aspects of the problem of assessing knowledge, skills and talents when analyzing the objective evaluation of the subjects' results: 1) mathematical processing of the results and 2) linguistic analysis of typical errors of the subjects which together with each other would contribute to the creation of a basis for the emergence of an integrated system for checking and evaluating the knowledge, skills and abilities of the subjects. The following are examples of statistical analysis of some tests:

4. FINDINGS

4.1. Statistical Analysis of the Oxford Placement Test

The material for the analysis was Dave Allen's Oxford Placement Test 1992 (OPT) (Allen, 1992) consisting of two parts: a listening test and a grammar test. Each of which includes 100 questions. One point is awarded for each correct answer so the total maximum score is 200. According to the data provided by the Association of Language Testers of Europe (ALTE), the scores acquired are interpreted by level (from functionally bilingual to absolute beginner) and they also show a correlation with the outcomes of the Cambridge English for Speakers of Other

Languages (ESOL) scale. For example, a score of 130-140 points relates the result to the intermediate level, a lower score of 100-120 relates points to the waystage level, even lower than the threshold level.

The methodology of knowledge testing in OPT is a combination of multiple-choice and substitution methods. In the listening comprehension test, the subject must insert one of the two suggested words into the sentence following what he heard. Three options are available for the grammar test. The hearing comprehension test procedure including the instructions of the testing teacher and the preliminary reading of the text last 45 minutes including 35 minutes of audio recording without stopping the film. Direct questions with pauses for answers are 20 minutes of sound. The rate of pronouncing the text is quite high. One astronomical hour is allotted for the grammar test. The sentences of the grammar test unlike the listening comprehension test are related in meaning in the second part.

The subjects were students entering the special Phycology Department of the Faculty of the Humanities and Social Sciences who had the first higher non-philological education and training in English in secondary school or a non-linguistic university. The average age of the subjects is 30-32 years and all of them are native speakers of Kazakh or Russian. 110 tests were processed. The tests were conducted in a regular classroom without a special one in groups of 20 people. The test results have been collected for two years. The test results were subjected to statistical analysis. The listening comprehension test was considered. The reliability of the test was measured by calculating the correlation between the two halves of the test (even and odd questions). Each subject was awarded two ranks on an even or odd number of questions. The correlation between these two ranks was calculated using the formula given above.

The correlation between the two halves of the test was 69%. The overall reliability of the Spearman-Brown test is:

$$Rtt = 2Rhh\ (1+Rtt) = 0.82\ (82\%)$$

The reliability index of 100% is almost unattainable, 82% is a very high indicator. The first part of the OPT (listening test) can be used if necessary, independently of the second part (the grammar test). However, according to the results of the analysis, the grammar test cannot be used separately from the auditory test since its reliability coefficient is much lower. Furthermore, the test was subjected to statistical analysis in terms of its adequacy to the contingent of subjects.

The difficulty coefficient was used to identify test questions that were too difficult (FV<40%) or too simple (FV>90%) for the subjects. These thresholds are used when deriving final results in Cambridge University exams: above 90% (level A) and below 60% (level D) i.e. failure. There were 9% of difficult questions and 13% of simple ones. In question 74, the errors are caused by purely phonological reasons because native speakers of Russian accustomed to a noisier back-lingual {x} do not perceive it as a consonant. The same reason could be seen for question 81 but here the matter is complicated by ignorance of the realities and the initials of the author of the book about Watergate. The reality of rod users (anglers from fishing rod) in question 91 is also difficult to perceive since it is unlikely that errors are caused by not distinguishing the diphthong in the word rode and the monophthong in the word rod. The inability to read *the @ icon* explains a large number of errors in question 80. Question 95 causes difficulties in perception due to the identical sound of the phrases *free kick* and *freek kick*. In question 76, phonetic factors also act in the pronunciation of modern English, the final {I} is vocalized turning into {ɯ} which causes errors in perception. The error in question 63 may be caused by not distinguishing the degree of openness of the vowel although the choice of subjects may also be influenced by the low frequency of the phrase belly dancer. Not distinguishing shod\shot in question 44 can be caused by both the inability to recognize the positional length of a vowel and ignorance of the word shod a concept that is quite rare in modern languages.

Table 1. Difficulty in coefficients of extremely difficult and extremely easy questions of the auditory perception sub-test.

No	Questions	FV<40 (%)	FV>90 %
3	This beard of mine is awfully itchy. I'll be glad when it does and grows.		99
	Do you have any idea how long ago it was found?		95
21	Why and where are you going to live in London?		96
22	It is recommended that dyslexic students follow the remedial reading and writing option.		95
25	I can see and consent to it if it has to be done.		92
32	During his holidays, he spends most of his time at the Lotus test track watching and washing cars.		95
36	Do you think you could take us and talk us through the next bit of the film?		95
37	How many tests and texts are we going to need to get all the data we want?		92
39	Are you going to Penny's or Benny's tonight?		96
42	One of the lecturers in the sociology department is writing a book on the old board school or bosrtal system.		
44	This horse will have to be shod or shot immediately.	21	
61	I thought his behavior was unexceptional or unexceptionable.		95
66	Recent EC regulations have been disastrous for British fish stocks or docks.		98
63	Her ambition is to become a belly or ballet dancer.	36	
74	My brother-in-law left for Euston or Houston early this morning, so he should get here tonight.	15	
76	You can buy logs by the barrow or barrel load at the local timberworks.	37	
80	He works for a company called JMB or J@B JMB Realty a Chicago real estate company.	38	
81	Have you read the latest book on Watergate by H.A. or A.J.Haldman?	38	
91	In England, all rod or road users must have a license.	33	
95	England would never have scored if it had not been for that free or freak kick by Gascoigne.	17	

The analysis of errors reveals that other aspects of language proficiency influence the outcomes of the listening test apart from the challenge of recognizing certain phonological oppositions. This test essentially tests not only perception by ear but also the understanding of a written text. The results are also influenced by knowledge of reality and familiar vocabulary.

Consider the results of a grammar test. The difficulty coefficients of the test questions are presented in Table 1.

There were more difficult questions in the grammar test than in the auditory test. 19% and less simple questions were 5%. It is noteworthy that in both tests simple questions are located in the first part of the test and complex ones are closer to the end (fatigue factor effects).

Table 2. Intra-test parameters of American and British tests.

No.	Questions	FV<40 (%)	FV>90 %
6	Parts of Australia don't have some or any rain for long periods.		96
8	Climate is very important in most of, most or the most people's lives.	29	
11	Pele is still perhaps most, the most famous or the more famous footballer in the world.		92
12	He had been, in or was born in 1940.		98
20	The world cup finals were in 1958 and Pele was looking forward to play, to playing or to be playing.	32	

No.	Questions	FV<40 (%)	FV>90 %
21	But he hurt this, the or his knees in a game in Brazil.		97
31	Uruguay had won the Olympic football final in 1924 and 1928 and wanted be, being and to be world champions for the third time.		92
32	Four teams entered from Europe but with a little, few, little or success.	32	
33	It was the first time which, that, when professional teams had played for a world title.	20	
41	Italy, which, that or who won.	19	
42	Went on to win, winning or to have won the 1938 final.	30	
56	Children seem to find computers easy but many adults aren't used to work, the work or working with micro technology.	27	
58	The only way to become proficient is to practice a lot on your own, by your own or on yourself.	21	
59	You can pick up the basics quite quickly if you want to, would or are willing to make an effort.	20	
61	Some people would just rather, prefer or better not have anything to do with computers at all.	10	
62	A lot have resigned them to never even know, known or knowing how a computer works.	23	
74	Then, about eighteen months after she has arrived, to have arrived or arriving in Norwich.	19	
81	She said, told or explained to me that by the time.	28	
82	She would pay, would have paid or had paid the mortgage.	28	
86	It seemed like a good idea so after we'd agreed, we could agree or we agreed with all the details.	35	
88	At the end of this month, we have lived, we have been living or we'll have been living together for a year and a half.	29	
89	It's the first time I live, I'm living or I've lived with anybody before.	15	
90	But I should guess, I might have guessed or I'd have guessed what would happen.	32	
96	He's rarely been away for this long before is he, hasn't he or has he?	18	
99	We'd better not delay reading this any longer, should we, did we or would we?	23	
100	Now's hardly the time to tell me you didn't need a text at all, did you? Is it or wasn't it?	15	

Table 2 presents the analysis of grammatical errors that allows us to identify weaknesses in the language training of the subjects, namely:

1) The use of perfect tenses (questions 74,82,88,89 and 90).

2) The use of gerund forms (questions 20,31,42,56 and 62).

3) The use of which, that and who conjunctions (questions 33 and 41).

4) The use of a dividing question (questions 96 and 99).

The methodological complexity of performing a grammar test is that the test takers were faced with a triple choice they should not only know a certain grammatical rule but also be able to navigate lexical uses and word combinations (for example, in question 61b). In some cases, the set of words depends entirely on the context (question 81). The answers of the test takers are divided into three categories: 1) incorrect in terms of grammatical

basis. 2) Incorrect in terms of usage and consumption. 3) Meeting both criteria and context. Thus, the complexity of the test increases at least three times.

These data allow us to conclude that the grammar test of OPT when used in the audience of Russian or Kazakh-speaking applicants turns out to be objectively less reliable and more difficult than the listening comprehension test. The subjects need additional training during their passages. The auditory test is quite reliable and can correspond to the level of training of Kazakhstani applicants with minor revision and replacement of some questions despite the high speed of the speakers' speech and small pauses.

4.2. Statistical study of Cambridge Tests and TOEFL (Test of English as a Foreign Language)

An experimental study of the intra-test and cross-test characteristics was conducted on subtests for the receptive skills of TOEFL, FCE (First Certificate in English) and CAE (Certificate in Advanced English).

The object of the study was the results of testing students of the Faculty of Humanities and Social Sciences and students of the Law Department of Toraighyrov University on the above-mentioned linguistic and didactic tests (100 people were tested in total).

A rich corpus of data has been collected for each of the tests studied: 400 papers on TOEFL, 200 on FCE and 200 on CAE. A total of 26,400 responses were received.

The results obtained from measuring the in-test parameters are shown in Table 3. Since the purpose of the test compilers is to assess the formation of different language skills and abilities in the test taker and his weaknesses and strengths, the test questions cannot be the same in nature. Thus, we can talk about the stability and constancy of the corrected coefficients obtained and consequently about the reliability of the tests under study.

Calculations of the discriminant coefficient for testing receptive types of *RD* have shown that 16% of the tasks of the subtest on the perception and understanding of oral speech by ear (TOEFL), 10% of the tasks (FCE) and 20% (CAE) cannot adequately divide the control group of test takers into groups of strong and weak according to the degree of mastery of the skill of perception and understanding of oral speech by ear and 20% of tasks (TOEFL) and 31% of test tasks (FCE) (first cross-section), 18% of TOEFL tasks and 16% of test tasks (CAE) (second cross-section) according to the degree of proficiency in reading skills to distinguish a well-prepared subject from a poorly prepared one. It follows that these tasks do not meet the validity parameter and therefore, should be replaced by complex tasks. The questions of the sub-tests of the tests under study were identified which turned out to be too difficult (facility value: FV<40%) and too simple (FV>70%) for this control group by calculating the difficulty coefficient.

Table 3. Difficulty of coefficients of extremely difficult and extremely easy questions in the grammatical sub-test.

Test	Correlation coefficient		Listening		Reading		Discriminatory coefficient	
	Listening	Reading	Simple questions	Difficult questions	Simple questions	Difficult questions	Listening	Reading
TOEFL (1st cross-section)	0.92	0.94	44%	4%	25%	8%	16%	20%
TOEFL (2nd cross-section)	0.93	0.94	44%	2%	25%	3%	16%	18%
FCE (1st cross-section)	0.9	0.9	25%	2.5%	60%		10%	31%
CAE (2nd cross-section)	0.9	0.9	50%	4%	50%		20%	16%

Thus, the calculation of the level of complexity of the sub-test for testing reading skills showed that some tasks on the TOEFL, FCE and CAE tests turned out to be too simple for the subjects of this control group since the level of complexity of tasks exceeded 70% which indicates a high level of preparedness of students at Toraighyrov

University on this aspect. The presence of simple questions in the studied sub-tests makes the tests less objectively reliable. When testing non-English speaking students in non-linguistic faculties, it becomes necessary to replace test tasks (or complicate some tasks) that do not correspond to controlled objects and to remove tasks that are excessively easy or tasks that are completed by 70% of the subjects in order to obtain more objectively reliable indicators for all parts of the tests under study.

In our opinion, the introduction of such changes in the development of testing technology for non-English-speaking university students would improve the quality of tests as tools for measuring skills for the types of *RD* being tested.

The British tests that were studied turned out to be easy for a control group of subjects. These students had not taken any special courses to prepare for the CAE test. They studied according to the usual program and did not get acquainted with the formats of British tests. Based on these findings, it can be assumed that students should be offered the CAE test at the beginning stage of the transition from secondary education to higher education. After completing the English language program, they should take the certificate of proficiency which is the highest level test for general English proficiency.

The American TOEFL test has a universal (flexible) assessment scale that enables you to evaluate, as previously mentioned, both weak and strong subjects. Unlike the FCE and CAE, this test does not assume thresholds in the distribution of points, so we can discuss whether it is easy or difficult for the subjects. Statistical analysis of the results makes it possible to identify certain qualitative characteristics of the test identify trends in the heterogeneity of the contingent of subjects which allows them to be grouped into homogeneous groups and check the adequacy of tests in a specific educational environment.

These experimental tests, in particular the values of *FV* and *DI* show that sub-tests to test the skill of perception and understanding of oral speech on the tests under study especially the British FCE and CAE tests were inadequate for conducting in the conditions of this sample of the Russian-speaking audience since most of the subjects turned out to be sufficiently qualified. In other words, the tests were too easy for most of the representative sample of subjects. Nevertheless, the value of the reliability coefficient allows us to conclude that the tests are reliable since the *HRC* is within the norm of 0.5.

5. DISCUSSION

It can be stated that the tests are reliable but inadequate for this non-English-speaking audience. The above in no way detracts from the advantages of the tests used. There is neither an ideal group of subjects nor an ideal test. Correlators were calculated for subtests to test the skill of perception and understanding of oral speech and to test the skill of reading and comprehension of what was read according to the TOEFL/FCE and TOEFL/CAE tests based on the statistical processing of experimental verification data. Correlators meet the requirements identified by us in the work that is, K corresponds to the phenomenon of which it is a meter. It is expressed as a number which allows you to have a clear idea of the measured relative value and allows the simplest methods of measurement using counting equipment (micro calculators) or even dispenses with it (manually).

$$K = \frac{\sum(x_1 - M_r) \cdot (x_2 - M_r)}{n}$$

The values of the correlator for sub-tests for testing receptive types of RD according to the TOEFL/FCE and TOEFL/CAÉ tests were obtained.

According to the subtest to test the skill of perception and understanding of oral speech, the correlation coefficient (correlator) according to the TOEFL/FCE tests is 0.3.

TOEFL SAE is 0.13, according to the sub-test for testing reading skills and reading comprehension, the correlation coefficient (correlator) according to the TOEFL/FCE tests is 0.3 and according to the TOEFL/CAE test, it is 0.2.

The obtained values of *K* give only about 20-30% reliability (and according to TOEFL / CAE – 13%) that is, *K* will work only for 20-30% of this sample of subjects most likely on a group of sufficiently advanced subjects. Since the degree of reliability is less than 50%, it can be concluded that the tests under study are poorly correlated with each other as they test different abilities. According to statistical norms, *K* can be in the range between 0 and 1. The closer the value of *K is* to 1, the higher the correlation.

Nevertheless, the low cross-test correlation does not detract from the merits of each individual test but only shows that the tests test different skills since they belong to different testing systems and accordingly have many differences in formats, structure, specifics and evaluation methods which is especially noticeable in a small sample of subjects (100 people). (In a large sample of subjects, the differences between the tests under study would not be so pronounced).

5.1. The Main Recommendations for Compiling the Test

1. The implementation of testing methods in teaching is motivated by the need for objective indicators. Many higher education institutions in Kazakhstan have created various language tests and assignments but the test creators often lack sufficient knowledge of psychological and pedagogical measurement theory and methodology. As a result, the quality of the tests suffers which affects the accuracy and reliability of the measurements taken with these instruments. Ineffective testing can lead to poor academic performance. Teachers should take into account the following factors to make sure that test results are accurate and useful in representing students' progress: The test compiler must have a clear understanding of what material needs to be tested. The success of the test depends on the ability to determine what to test. The test must be relevant to the content taught during the learning process.
2. The theory of language testing posits that assessing proficiency in certain language aspects serves as an assessment of proficiency in skills. Therefore, it is necessary to test individual elements of the language separately depending on the goals of the testing conditions. For instance, if we need to conduct a diagnostic pronunciation test to highlight special difficulties that must be learned by the student, we should choose a test that tests the sound system of the language. If we want to determine the general level of language proficiency, we should offer a listening comprehension test and a writing test to assess the student's ability to take other subjects using a foreign language as a means of communication after studying them.
3. The test should be designed in such a way that it can be tested in a short time for a week, a month, a year or even several years. At the same time, a clear distinction should be made between language learning and the test. Therefore, the language is learned for a long time and the test requires a short time.
4. The test should mainly, if possible, contain materials that reflect the differences between the native language and the foreign language to be learned. It is necessary to test only the language difficulty, i.e. the knowledge of the difficulty is the knowledge of the language. It can be concluded that testing language proficiency consists of testing the mastery of language difficulties, i.e. units and models that are not present in the mother tongue or that are present but have a different structure or meaning.
5. When compiling the test, the most important statistical parameters must be taken into account which can be summarised as follows: the suitability coefficient (validity), the reliability coefficient (reliability), the complexity coefficient of the question and the discriminant coefficient.
6. The cost benefit ratio of the test. This is a requirement for the practicability of the test. If the test measures what we want to measure in the shortest possible time, then it is economical and practical.

7. Easy to calculate. Can the test results be easily calculated? If so, then the test meets this requirement. Subjective tests are not easy to calculate. The examiner finds it difficult, doubts which answer is better, more correct and more complete. There are no such difficulties with objective tests but even with objective tests the calculation varies with ease. For example, tests in which all answers are collected on a separate sheet are considered to be tests whose results are easier to calculate than tests whose results are scattered on all pages.

These are the requirements that the test must meet as one of the methods of objective control for student success.

6. CONCLUSION

The tests under study made it possible to measure and track the dynamics of the development of receptive skills by the end of the academic year, both for each individual subject and for the entire group as a whole. At the same time, the dynamics of success on British tests (especially on the CAE listening comprehension subtest) are much higher than American ones since students are taught in the British version of English throughout the year, teachers adhere to the British norm of spoken English and they use British textbooks, audio and video cassettes.

Probabilistic factors of the general distribution of the test subjects' scores reflect the heterogeneous composition of the sample of the contingent of subjects, (students' group of the Faculty of Humanities and Social Sciences), (students' group of the Faculty of Economics and Law), belonging to different genders, the different difficulty of tasks, the degree of heterogeneity of the initial level of training of the subjects and different levels of development of thinking processes and memory. Attention: different psychological preparedness (readiness) to pass tests, minimum time gap between passing comparable tests (emotional and physical condition) and so on. All these factors act in the same direction and group the subjects.

The reliability coefficient is equal to 0.86 and 0.87 according to the test (TOEFL) and 0.9 (according to the FCE/CAE tests) according to the listening sub-test for checking the perception and understanding of oral speech (listening) as well as 0.88 (first or second cross-sections) (according to the TOEFL test), 0.9 (first or cross-section) according to the FCE (CAE test) according to the reading and comprehension skill test (reading). The values of the standard deviation are 5.95 and 5.82 (TOEFL), 5.5 and 16 (FCE and SLE), 6/21 and 6.07 (TOEFL) and 7.0 (FCE and SAE) according to the reading that the tests are reliable.

The parameters of the difficulty coefficient as well as the discriminant coefficient indicate that 16% of the tasks (TOEFL), 10% of the tasks (FCE) and 20% (CAE) of the sub-test on the perception and understanding of oral speech (TOEFL), 20% of the tasks (TOEFL) and 31% of the tasks of the test (FCE) (the first section), 18% of TOEFL tasks and 16% of test tasks (SAE) (the second section) cannot adequately divide the control group of test takers into groups of strong and weak in terms of the degree of possession of receptive skills and need to be replaced due to their inadequacy to the level of the test takers in this sample of the Russian and Kazakh-speaking audience.

The calculated correlators of the TOEFL/FCE and TOEFL/CAE tests for subtests for testing receptive types of RD tests show that the British and American tests are aimed at measuring heterogeneous skills. The low cross-test correlation suggests that the tests are not quite of the same nature, they test different skills which is especially noticeable in a small sample of subjects.

Funding: This study received no specific financial support.
Institutional Review Board Statement: The Ethical Committee of the Toraighyrov University NJSC, Kazakhstan has granted approval for this study on 20 June 2023 (Ref. No. 1).
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

Allen, D. (1992). Oxford placement test 2, Grammar test, Part 1. In (Vol. 10). Oxford: Oxford University Press.

Arruarte, J., Larrañaga, M., Arruarte, A., & A., E. (2021). Measuring the quality of test-based exercises based on the performance of students. *International Journal of Artificial Intelligence in Education, 31*, 585–602. https://doi.org/10.1007/s40593-020-00208-0

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *The Modern Language Journal, 64*(3), 311-317. https://doi.org/10.1111/j.1540-4781.1980.tb05198.x

Buntins, M., Buntins, K., & Eggert, F. (2017). Clarifying the concept of validity: From measurement to everyday language. *Theory & Psychology, 27*(5), 703-710. https://doi.org/10.1177/0959354317702256

Cherepanov, V. S., & Shikhov, Y. (2008). Qualimetric monitoring of the quality of education: A conceptual and programmatic approach. *Education and Science, 2*(50), 64-72.

Cho, E., Capin, P., Roberts, G., Roberts, G. J., & Vaughn, S. (2019). Examining sources and mechanisms of reading comprehension difficulties: Comparing English learners and non-English learners within the simple view of reading. *Journal of Educational Psychology, 111*(6), 982. https://doi.org/10.1037/edu0000332

Conrad, C. (1971). The cloze technique and ESL proficiency. *Language Learning, 21*(2), 183-195.

Damico, J. S., Oller, J. W., & Tetnowski, J. A. (1999). Investigating the interobserver reliability of a direct observational language assessment technique. *Advances in Speech Language Pathology, 1*(2), 77-94. https://doi.org/10.3109/14417049909167162

Darnell, D. K. (1970). A procedure for testing English language proficiency of foreign students. *Speech Monographs, 37*(1), 36-46.

Espin, C., Deno, S., McMaster, K., Pierce, R., Yeo, S., & Mahlke, A. (2009). *Teacher use study: Progress monitoring with and without diagnostic feedback*. Technical Report No. 38. Research Institute on Progress Monitoring.

Hashemi, A., & Daneshfar, S. (2018). A review of the IELTS test: Focus on validity, reliability, and washback. *Indonesian Journal of English Language Teaching and Applied Linguistics, 3*(1), 39-52. https://doi.org/10.21093/ijeltal.v3i1.123

Hymes, D. H. (1972). On communicative competence in Sociolinguistics (Pride & Pholmes, Eds.). In (pp. 269-293). Harmoundsworth, UK: Penguin.

Johnsen, S., & Sulak, T. (2021). Screening, assessment, and progress monitoring. In M. R. Coleman & S. Johnson (Eds.), Implementing RtI with gifted students. In: Routledge. https://doi.org/10.4324/9781003235736-4.

Lado, R. (1986). Analysis of native speaker performance on a cloze test. *Language Testing, 3*(2), 130-146. https://doi.org/10.1177/026553228600300203

Mabel, R. O., & McKeithan, G. K. (2022). Progress monitoring of language acquisition and academic content for English learners. *Learning Disabilities Research & Practice, 37*(3), 216-225. https://doi.org/10.1111/ldrp.12290

Merlo, P., & Stevenson, S. (2001). Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics, 27*(3), 373–408. http://dx.doi.org/10.1162/089120101317066122

Oller, J. J. (1973). Cloze tests of second language proficiency and what they measure. *Language Learning, 23*(1), 105-118.

Spolsky, B. (1968). Language testing: The problem of validation. *Tesol Quarterly, 2*(2), 88-94.

Stecker, P. M., & Fuchs, L. S. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research & Practice, 15*(3), 128-134. https://doi.org/10.1207/SLDRP1503_2

Stubbs, J. B., & Tucker, G. R. (1974). The cloze test as a measure of English proficiency. *The Modern Language Journal, 58*(5/6), 239-241.