IRSTI 14.35.07 https://doi.org/10.63597/UTO3105-4161.2026.5.1.002

A. Serikbayeva[1*], D. Akizhanova[2], L. Shinetova[3], B. Abdrasilov[4]
[1]Al-Farabi Kazakh National University, Almaty, Republic of Kazakhstan
[2]Eurasian National University, Astana, Republic of Kazakhstan
[3,4]National Testing Center, Astana, Republic of Kazakhstan
*email: assiyaserikbayeva237@gmail.com

[1]ORCID 0009-0002-7516-482X, [2]ORCID 0000-0002-1300-0339
[3]ORCID 0000-0003-4280-7999, [4]ORCID 0009-0002-1371-6211

PSYCHOMETRIC VALIDATION OF A HIGH-STAKES ENGLISH LANGUAGE TEST: A RASCH-BASED ANALYSIS OF KAZAKHSTAN'S UNIFIED NATIONAL TESTING

The Unified National Testing is a high-stakes examination that serves as a critical gatekeeper for university admission. Psychometric evidence is necessary to justify score interpretations in high-stakes testing; however, its psychometric properties remain underexplored. The study aims to analyze the English section of UNT from a psychometric perspective, including the reliability, construct validity, item difficulty, discrimination indices, and how well the test corresponds to the curriculum and CEFR levels. This study examines whether the test adequately measures actual language proficiency. A pilot quantitative study was conducted with 212 secondary school students (grades 10-11) who selected English language subject. Within the framework of Classical test theory (CCT) discrimination indices were calculated while reliability, fit statistics and item targeting were examined using Rasch measurement analysis. Differential Item Functioning was used to compare results of 10^{th} and 11^{th} grade performances. Psychometric results reveal that UNT's English component indicated strong internal consistency (Cronbach's $\alpha = 0.879$; $\omega = 0.898$). The majority of items showed acceptable measurement as (person reliability = 0.836; item reliability = 0.972). However, several items were relatively easy in grammar and reading sections. The teacher survey revealed insufficient assessment of productive skills and a strong dominance of receptive skills (M = 4.50). The English section of the UNT is generally at an acceptable level in terms of reliability and structural stability. However, demonstrates limited representation of communicative competence and productive skills. Therefore, the findings emphasize the need for score alignment with standardized tests as IELTS with CEFR descriptions.

Keywords: Unified National Testing; English language assessment; construct validity; Rasch model.

Introduction

In today's contemporary educational contexts English language assessment plays an essential role in evaluating individuals' level of language proficiency using standardized exams and tests. According to Brown (2004) [1], assessment is an ongoing process that involves several different tasks.

In Kazakhstan, the Unified National Testing is a standardized test that examines the level of secondary school completion and the knowledge necessary for admission to higher institutions. The purpose of the Unified National Test's English section is to objectively assess students' lexical, grammatical, reading comprehension, and correct use of language structures through a standardized test format. The English language section of the UNT is selected by graduates intending to pursue careers such as English language teachers, future diplomats and translators who establish contacts in international relations. Recent developments in Kazakhstan's Unified National Testing reports (2025) [2] that over 210,000 participants attended UNT and approximately 76.35% of applicants met requirements for state-funded grants. From the standpoint of language evaluation, the quality of a test largely depends on its psychometric properties and the validity of score interpretation [3, 4, 5].

Messick (1995) [6] defines validity as the meaning of the test score and the validity of its use, not the test itself. Messick considers validity as a unified concept and considers it to consist of 6 aspects: content, theoretical/processual, structural, generalizability, external, consequential. Thus,

validity is not limited to psychometric indicators, it also includes the social implications of test results.

Kane (2013) [7] claims that validity is not the accumulation of evidence, but rather the systematic evaluation of the argument that explains a test score. According to Kane, validity assessment has two stages: the formative stage has clear interpretation and formulates predictions while summative stage empirically tests those predictions. Reliability and validity play a crucial role in evaluating language tests and ensuring meaningful interpretation of test scores. [8, 9].

The CEFR framework assesses receptive and productive skills such as listening, reading, writing and speaking [10]. However, UNT's English section does not directly measure productive skills whereas it measures grammatical knowledge and reading comprehension.

The significance of the study lies in an in-depth analysis of psychometric quality of English section UNT exam, high-stakes assessment that plays significant role for deciding university admission in Kazakhstan. Psychometric evaluation of large-scale tests as UNT is essential to ensure validity, reliability, fairness and transparent interpretation of test scores. This research contributes to the field of language assessment by offering concrete data on psychometric properties of UNT. Previous research works mostly concentrated on digitalization of UNT exam, governance reforms, societal impact. The aim of this study is to analyse the English section of UNT from a psychometric perspective, including the reliability, validity, item difficulty, discrimination indices, and how well the test corresponds to the curriculum and CEFR levels and the construct validity. This research aims to determine wheather the test measures real language skills.

In this study, the psychometric evaluation of the English section of the UNT is based on Messick's theoretical framework of construct validity and is also based on Kane's argument-based validation model at the interpretive level. Therefore, the factor structure and Rasch results are considered not only as statistical evidence, but also as empirical components of the argument supporting the interpretation of test scores.

Literature review

National examinations are crucial in shaping educational pathways, particularly when high-stakes tests determine access to university admissions. Previous studies highlighted drawbacks of UNT such as student stress, regional disparities and curriculum constriction [11]. However, recent research works explored students' perception on UNT, state scholarships and discrepancy between test results and university level competencies [12].

Furthermore, Mingisheva (2023) [13] analyzed the main challenges in UNT and examines transition of UNT format to international standards (PISA, SAT). The study highlights educational inequality, regional differences and data transparency however, the study has not thoroughly examined UNT's psychometric properties (reliability, validity, discrimination index) especially for the English language section. Didarbekova et al (2025) [14] claim that UNT tasks are still created manually using Microsoft Word which leads to errors and repetition, psychometric indicators are calculated after the test not when the task is created. The results of study proved that introduction digital platforms and automated item generation also real-time psychometric monitoring are the main mechanisms for improving the quality of the UNT. Additionally, Karelkhan & Alchenova, (2025) highlights the importance of subject preparation (computer science) in the UNT system. The results of study indicates the need to update the content of the test. Jumabayeva (2016) [15] examines the primary motivations for the UNT's implementation (time saving, obtaining school certificate, combating corruption as well as indicates some drawbacks psychological pressure for teenagers, the growth of private tutoring. Turlanbekova (2015) [16] claims that the UNT does not fully measure knowledge: majority of teenagers memorize answers to multiple-choice questions which lacks critical thinking, writing, communication skills. The result of study reveals that test causes high levels of stress, perceived less fair than international tests (IELTS, TOEFL,SAT). Abdrasilov et al. (2026) [17] analyzed validity of UNT results in Kazakhstan. Author used mathematic component of UNT system analyzed tests psychometric quality, reliability, task

difficulty and predictive validity. The study based on S. Messick's validity theory, classical test theory CTT and modern test theory IRT, Rasch models.

A review of international research indicates that construct underrepresentation occurs if a test does not fully capture the ability it is supposed to measure, this poses a threat to validity [6]. Bachman (1990) [18] developed concept of communicative competence and claimed that except grammar various components of language should be assessed together. According to Kane (2013, 2022) [7, 8], validity is not just a collection of statistical indicators, but a process of justifying the interpretation based on the test results. Especially in high-stakes exams like the UNT, where the test results affect the future opportunities of the candidates, the interpretation of the score should be based on scientific argument. In this sense, reliability coefficients, Rasch model results, factor structure, and CEFR correspondence are not independent evidence of validity, but rather are considered components of the interpretative argument.

Kunnan (2004) [19] closely links test fairness to validity, noting that validity and reliability are not sufficient in high-stakes testing. However, McNamara and Roever (2006) [20] describe the test as a social mechanism. Chapelle (1999) [21] argues that construct-task congruence is a key condition for validity in language assessment. These theoretical frameworks require an analysis not only of the psychometric quality of the English section of the UNT, but also of its social and interpretive implications.

The CEFR framework proposed by the Council of Europe (2018) [10] provides a comprehensive assessment of receptive and productive language skills, and any national test should, if it is to be internationally aligned, be consistent with the skill model outlined in the CEFR. North (2000) [22] argues that not only content similarity but also dimensional similarity are important when considering CEFR compliance. Alderson (2007) [23] criticizes the lack of sufficient empirical evidence to support the CEFR compliance of many tests. Moreover, expert judgment and thorough empirical evidence are needed for formal procedures that link exams to the CEFR [24]. Task authenticity and cognitive processing demands should be taken into account when aligning reading tasks in particular [25].

However, there is a research gap regarding the psychometric quality of English tasks in the UNT. In this regard, there is limited research of psychometric studies in the Kazakhstani context. Most studies provide only content descriptions. The exact level of correspondence of UNT English section tasks to the CEFR scale (B1–B2) is not fully determined, which makes comparison with international standards difficult. Psychometric analyses of English section UNT exam remains underexplored. To bridge this gap, this article presents psychometric evaluation of UNT's English language section, offering actual data on test quality, measurement reliability and validity within the Kazakhstani assessment context.

Accordingly, the following research objectives are formulated:

RO1: To investigate the psychometric properties of UNT's English section, with an emphasis on item-level performance and reliability.

RO1.1: To explore evidence-based construct validity in relation to CEFR proficiency levels.

RO2: To provide research-based suggestions for enhancing the quality of UNT's English section based on psychometric analysis.

Most studies on the UNT system examine its social impact, the process of digitization, and its impact on education policy. However, studies that specifically analyse the English language section are very limited. In Kazakhstan, the UNT is a high-stakes test that determines the chances of applicants entering higher education. However, the reliability, discriminant validity, construct validity, and correspondence to the CEFR levels of the English language section have not been systematically studied empirically. Therefore, this research work provides an opportunity to conduct a comprehensive psychometric analysis of the English language section, scientifically assess the quality of the test, and determine its compliance with international standards.

Materials and Methods

Participants. The quantitative part of study involved 212 students from the 10th and 11th grade who chose English language for the UNT. Participants are students from comprehensive schools. During pilot study data were collected on questions taken from the Unified National Testing website. All participants were school students preparing for the UNT within the same curriculum. The research data was processed anonymously, and the students' personal data was not used, which fully complies with the requirements of academic ethics.

The teacher survey involved 22 English language teachers. The majority of participants (n = 16) were aged 18–35 years, whereas 6 teachers were aged between 35 and 45 years. The survey was conducted among teachers working in comprehensive schools who prepare 11th grade students for the UNT. The teachers had at least 5 years of teaching experience and were familiar with the format and content of the UNT.

Instruments. The first tool was a pilot study of English language test items published on the official website of the Unified National Testing (UNT), each item's difficulty index, discriminant index, and internal reliability of the test (Cronbach's alpha) were calculated. UNT is a complex assessment system, within the framework of which candidates pass exams in three compulsory subjects (reading literacy, mathematical literacy, history of Kazakhstan) and two optional subjects, depending on the specialty.

The second tool was a questionnaire for teachers, which included the test format, listening and speaking skills, measurement validity and fairness, and suggestions for improving the UNT.

Data Collection. The research data was collected in two stages. First, the UNT English language test items were used in a pilot format with 10th and 11th grade students. Then, an online survey was conducted among English teachers. All data was collected anonymously, and the requirements of academic ethics were observed.

Ethical Considerations. Informed consent was obtained from all participants before the study was conducted. Since the participants were school students, permission was also obtained from the school administration. Participants were informed about the purpose of the study and were given full information about their right to withdraw from participation at any time. Participants were informed that their personal information would be kept anonymous and confidential throughout the study.

Data Analysis

All statistical analyses were conducted using software program Jamovi (Version 2.6.45.0). The quantitative data were processed using psychometric analysis methods. The difficulty index, discriminant index, and internal reliability of the test (Cronbach's alpha) and Mc Donald's omega were calculated for each task. Rasch modeling techniques were used to analyze measurement accuracy, item fit, and grade-level differences in item functioning.

Teacher perception questionnaire used Jamovi (Version 2.6.45.0) software to compute means and standard deviations to summarize response patterns across the five-point Likert scale items. The survey data were analysed using descriptive statistics methods (percentages and averages). The results obtained were interpreted comparatively in order to comprehensively assess the quality of the test items.

Results and discussion

This section shows psychometric analysis of Unified National Test's English component. The overall mean score was 36.06, SD 6.90. The mean score for the whole test indicates a moderate level of performance with adequate level of variability as shown in Table 1. The UNT examination had a strong overall internal consistency (Cronbach's alpha 0.879).

Table 1 - Descriptive analysis of UNT's English component (Detailed item-level facility and discrimination indices are presented in Appendix A)

Section	Items	Mean	SD	Min	Max	Cronbach's α	McDonald's ω
Grammar	Q1–Q20	15.85	3.68	8	20	0.821	0.855
Multiple response	Q21–Q30	11.25	2.92	1	14	0.728	0.847
Reading	Q31–Q40	8.96	1.51	3	10	0.677	0.797
Total Test	Q1–Q40	36.06	6.90	18	43	0.879	0.898

The majority of reading tasks were located at medium logit ranges. Several tasks showed high degree of difficulty. This indicates that students found these activities to be reasonably accessible. Reading texts based on information therefore tasks do not require critical thinking skills. Overall, UNT's this reading part of measures students general reading comprehension and ability to identify specific information.

Table 2 - Rasch model statistics

Indicator	Value
Person Reliability	0.836
Person Separation Index	2.25
Estimated Ability Strata	≈ 3
Item Reliability	0.972
Item Separation Index	5.88

According to the Rasch Model Statistics, the Person Reliability index score was .836, which means the English section of the UNT test can consistently measure the level of students and reliably distinguish between students' abilities. In addition, Person Separation Index indicated 2.25 which means students can be divided into approximately 3 levels. It was observed that the Estimated Ability Strata can distinguish between three different groups of weak, medium, and high-ability students ≈ 3. Item Reliability showed = .972 the stability of the difficulty hierarchy of the questions, and Item Separation Index = 5.88 that the questions are at different levels of difficulty. Overall, the English section of the UNT test can divide students into levels and the structure of the questions is stable, as well as the measurement accuracy is sufficient.

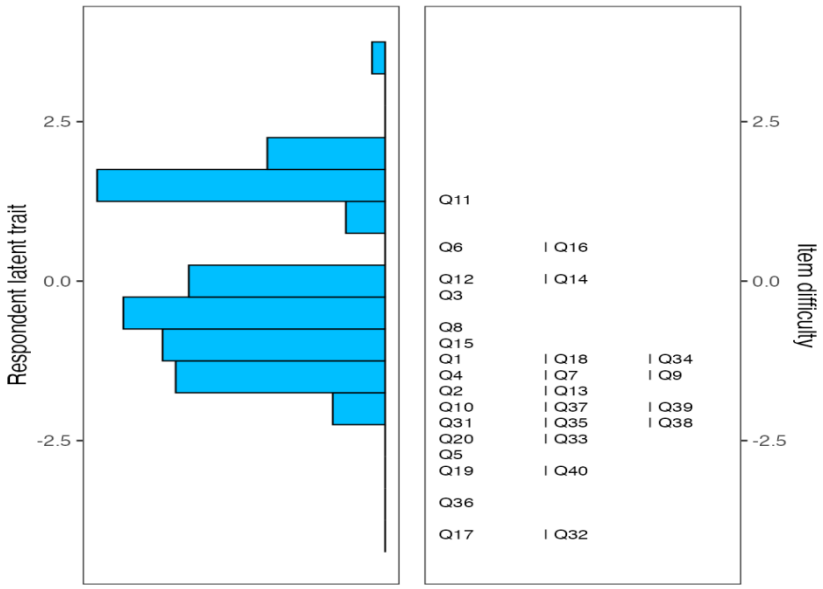


Figure 1 - Wright Map of Person ability and Item difficulty scale

Wright Map illustrates person ability and item difficulty scale logits most participants were varied between -1 and +1 logits, indicating pilot test results were low to moderate. Several items like Q11 and Q16 showed higher difficulty level whereas Q17 and Q37 were at lower logit levels showing easier tasks. According to results person mean was just above than item mean which means sample was easier for this group.

Table 3 - Rasch Item Fit Statistics

Fit Category	Number of Items	Interpretation
Infit 0.7–1.3	34	Acceptable fit
Outfit 0.7–1.3	33	Acceptable fit
Misfitting items	3	Require review
Overfit items	4	Very predictable responses

According to Rasch fit analysis majority of items were within the acceptable measurement scale. Moreover, 34 items met acceptable Infit criteria while 33 items met Outfit criteria (0.7–1.3) indicating satisfactory model fit data. Three items classified as misfitting which require further review while four items represented overfit which shows very predictable responses. Rash fit data indicated that most items functioned consistently inside the measuring framework, however, CFA results showed multidimensionality of the test.

Table 4 - Differential Item Functioning (Grade 10 vs Grade 11)

Item	Δ Difficulty	p-value	Effect Size	Interpretation
Q6	0.28	0.049	Small	Favors Grade 11
Q11	0.34	0.027	Small–Moderate	Favors Grade 11
Q26	0.39	0.013	Moderate	Favors Grade 11
Q34	0.45	0.006	Notable	Favors Grade 11
Other items	< 0.20	> 0.05	Negligible	No DIF

Table 4 illustrates that four questions Q6, Q11, Q26, Q34 showed significant difference between 10 and 11 graders. According to results test was slightly easier for 11 grade students, the level of difference does not exceed average limit. Overall, the test does not show any obvious unfairness between graders.

Table 5 - Descriptive Statistics of the Teacher Perception Questionnaire (N = 22)

Questions		M	SD	Min	Max
1	UNT reflects overall English proficiency	3.09	1.38	1	5
2	UNT focuses mainly on grammar	4.50	0.60	3	5
3	Test format limits language use	3.36	1.29	1	5
4	UNT reflects communicative competence	2.50	1.10	1	4
5	Test-only format restricts skills	3.41	0.96	1	5
6	Teachers can assess listening ability	4.05	1.00	1	5
7	Students understand spoken dialogue	3.50	1.06	1	5
8	Listening section would improve results	3.73	1.08	1	5
9	Listening is essential skill	4.09	1.15	1	5
10	UNT does not assess listening	4.14	0.85	2	5
11	Students can express ideas orally	3.36	1.09	1	5
12	Teacher speaks English in class	4.00	1.02	1	5
13	Speaking section would improve validity	3.77	1.11	1	5
14	UNT does not cover all language skills	3.36	1.09	1	5
15	Results do not fully reflect real ability	3.45	1.06	1	5
16	UNT results align with real ability	2.95	1.17	1	5
17	Listening section is necessary	3.23	1.23	1	5
18	Speaking tasks should be introduced	3.14	1.28	1	5
19	Results should align with international standards	3.14	1.13	1	5
20	Computer-based speaking is possible	3.50	1.19	1	5

According to questionnaire results teachers believe the English section of UNT is grammar-focused (M = 4.50, SD = 0.60). Teachers show strong agreement (M=4.14) that listening skills are not assessed in the English section of the test. However, there was less agreement regarding test's ability to fully reflect communicative competence (M = 2.50, SD = 1.10). Moreover, there is slight agreement regarding the limited alignment between UNT scores and students' real language ability (M=2.95). The results reveal that current test structure show an insufficiency in the assessment of productive skills and communicative competence.

Confirmatory factor analysis

Table 6 - Confirmatory factor analysis (Standardized factor loadings for all items are presented in Appendix B.)

Model	χ^2	df	CFI	TLI	RMSEA	90% CI
Full English Section (40 items)	7595	737	0.248	0.205	0.210	[0.205, 0.214]
Grammar Single Response (20 items)	1740	170	0.397	0.326	0.209	[0.200, 0.218]

Confirmatory factor analysis was conducted to examine the internal structure of the English section UNT. Model showed poor global fit $\chi^2(737) = 7595$, $p < .001$; CFI = .248; TLI = .205; RMSEA = .210 (90% CI [.205, .214]). The results indicate that the grammar section includes variety of linguistic elements such as: phonetics, syntax, conditional and transformational grammar also reading comprehension tasks. However, the results of the Rasch model showed that the measurement accuracy was good and the difficulty level of the tasks was stable. This variety suggests the test measures multiple partially independent subskills rather than a single latent competence dimension. Complete item-level facility and discrimination statistics are presented in Appendix A. Standardized factor loadings are reported in Appendix B. Items demonstrating negative discrimination coefficients (see Appendix A) were examined for potential multidimensionality.

Discussion. The findings provide mixed but theoretically meaningful evidence regarding the construct representation of the UNT English section. In addition to structural similarity, evidence of construct equivalency and consistency in score interpretation across contexts are necessary for large-scale examination alignment [26]. On the one hand, the internal consistency and Rasch reliability indicators prove that the test measures consistently. On the other hand, the CFA results show that the English section cannot be considered as a single measure of language ability.

The apparent disparity between CFA structural misfit and Rasch model fit raises the possibility that the underlying construct is not homogeneous even when items function consistently within a measuring scale. Therefore, the analysis shows that some of the tasks in the grammar and reading sections, especially according to the Rasch model results and Wright map, are relatively easy. Many students score high, but it is difficult to analyze their actual level in more depth. As for the reading section, the tasks are mainly focused on finding specific information in the text. The number of tasks that require deep analysis, critical thinking, or interpretation skills is limited. This may indicate that the test does not fully measure the high-level language competence of students.

The multifaceted pattern is consistent with Bachman's (1990) [18] theory that communicative language ability is not a single latent characteristic but rather a composite of textual, pragmatic, and grammatical competencies. The test included reading tasks that combined multiple internal processes, including information retrieval, inference, and lexical/grammatical processing, which may have influenced the task. Therefore, the results suggest that the tasks are not a weakness in the reading section, but rather an early diagnostic marker of multidimensionality.

According to Kane’s argument-based validation model, score use must be supported by evidence demonstrating construct comparability. In addition, applicants who have achieved a score of 6.0 on the international IELTS exam are automatically awarded the maximum score in the English section of the UNT. The six-point scale corresponds to approximately B2 level in the IELTS exam according to the CEFR system and comprehensively assesses four language skills (listening, reading, writing, pronunciation). The English section of the UNT, on the other hand, mainly measures grammar and reading skills. In this regard, the constructive structure of the two exams may not fully align. Therefore, the automatic scoring policy may require additional empirical justification in terms of CEFR compliance.

According to CFA a significant structural result was identified even grammar section was modeled separately, the fit remained poor. This indicates that Grammar section has variety language components as vocabulary recognition, phonological awareness, syntactic rule application etc.

Conclusion

A comprehensive psychometric assessment of the English section of the Unified National Testing (UNT) in Kazakhstan was conducted within the framework of structural validation based on Messick's unified validity theory and Kane's argument-based validation model. The findings indicate that the UNT English section demonstrates satisfactory internal consistency and acceptable Rasch-based measurement precision.

The data also revealed important limitations regarding the test structure. Topic-level analyses revealed several easy and two negative discriminant items, while the instructional component may reflect multidimensional processes rather than a unidimensional structure. In addition, teacher perceptions consistently emphasized the dominance of grammatical and receptive skills, along with an underrepresentation of productive and communicative competencies.

However, the findings raise concerns regarding alignment with CEFR criteria and the absence of speaking and listening section. IELTS 6.0 corresponds to approximately B2 level on the CEFR and comprehensively assesses four components of language competence. The UNT English section, on the other hand, measures only grammatical structures and text comprehension. Therefore, the automatic alignment of IELTS results with UNT results requires additional empirical justification in terms of construct validity.

In conclusion, from a psychometric perspective, the English section of the UNT consistently includes tasks of varying difficulty and distinguishes between weak, medium, and strong students. Future research is recommended to strengthen the evidence of external validity by comparing the results of the English section of the UNT with international test scores.

Limitations

The study was pilot-scale and used data from 212 participants for analysis. In addition, no direct comparison with IELTS results was made as an external criterion. Future studies are recommended to conduct correlational analyses with a larger national sample and international test results.

Furthemore, a significant model mismatch was found by confirmatory factor analysis, suggesting that the construct may be multidimensional. Future research should employ nationally representative samples, longitudinal validation designs, and cross-test linking methodologies to strengthen the evidentiary basis for score interpretation and policy decisions.

Appendix A

Item difficulty (facility) and discrimination indices

Item	Facility (p)	Corrected r_it
Q1	0.792	0.350
Q2	0.863	0.413
Q3	0.745	0.716
Q4	0.816	−0.048
Q5	0.939	0.171
Q6	0.717	0.489
Q7	0.820	0.506
Q8	0.910	0.287

Q9	0.660	0.561
Q10	0.835	-0.273
Q11	0.561	0.612
Q12	0.670	0.460
Q13	0.792	0.309
Q14	0.759	0.467
Q15	0.807	0.420
Q16	0.759	0.440
Q17	0.868	0.385
Q18	0.816	0.519
Q19	0.896	0.263
Q20	0.849	0.451
Q21	0.764	0.366
Q22	0.802	0.334
Q23	0.858	0.252
Q24	0.755	0.346
Q25	0.811	0.316
Q26	0.642	0.416
Q27	0.708	0.402
Q28	0.745	0.353
Q29	0.717	0.404
Q30	0.642	0.384
Q31	0.915	0.247
Q32	0.943	0.233
Q33	0.943	0.220
Q34	0.972	0.182
Q35	0.830	0.273
Q36	0.915	0.264
Q37	0.943	0.241
Q38	0.943	0.234
Q39	0.953	0.214
Q40	0.953	0.226

Appendix B

Factor loadings

Factor Loadings					
Factor	**Indicator**	**Estimate**	**SE**	**Z**	**p**
Grammar single response	**Q1**	0.20066	0.02709	7.406	<0.001
	Q2	0.14535	0.02342	6.206	<0.001
	Q3	0.35710	0.02511	14.223	<0.001
	Q4	0.01809	0.02788	0.649	0.517
	Q5	0.06800	0.01679	4.051	<0.001
	Q6	0.40583	0.02656	15.281	<0.001
	Q7	0.17340	0.02327	7.451	<0.001

	Q8	0.23450	0.02585	9.071	<0.001
	Q9	0.13767	0.02552	5.394	<0.001
	Q10	-0.16262	0.02986	-5.446	<0.001
	Q11	0.34956	0.03084	11.334	<0.001
	Q12	0.38445	0.02496	15.402	<0.001
	Q13	0.07684	0.02329	3.300	<0.001
	Q14	0.31216	0.02759	11.316	<0.001
	Q15	0.12881	0.02956	4.358	<0.001
	Q16	0.25540	0.03333	7.662	<0.001
	Q17	0.00982	0.01093	0.899	0.369
	Q18	0.18923	0.02454	7.710	<0.001
	Q19	0.03050	0.01771	1.722	0.085
	Q20	0.05062	0.01763	2.870	0.004
Grammar multiple response	**Q21**	0.30609	0.04252	7.198	<0.001
	Q22	0.29004	0.02660	10.904	<0.001
	Q23	0.16113	0.01861	8.657	<0.001
	Q24	0.90516	0.07635	11.855	<0.001
	Q25	0.22535	0.03618	6.229	<0.001
	Q26	0.02617	0.01204	2.173	0.030
	Q27	0.26643	0.03854	6.913	<0.001
	Q28	0.12138	0.02042	5.945	<0.001
	Q29	0.20410	0.02027	10.069	<0.001
	Q30	0.06413	0.00956	6.706	<0.001
Reading	**Q31**	0.13553	0.01880	7.211	<0.001
	Q32	0.07164	0.01003	7.141	<0.001
	Q33	0.23456	0.01251	18.745	<0.001
	Q34	0.11439	0.02977	3.843	<0.001
	Q35	0.12389	0.02323	5.332	<0.001
	Q36	-0.00743	0.01252	-0.594	0.553
	Q37	0.20400	0.02426	8.408	<0.001
	Q38	0.22189	0.01673	13.267	<0.001
	Q39	-0.02922	0.02638	-1.108	0.268
	Q40	0.14788	0.01201	12.313	<0.001

References

1. Brown H. D. Teaching by principles: An interactive approach to language pedagogy (2nd ed.). - Longman, 2004. - 491 p.
2. Prime Minister’s Office of the Republic of Kazakhstan. Quality and accessible education: Over 93,000 students received grants in 2025. - 2025. - https://primeminister.kz/en/news/quality-and-accessible-education-over-93000-students-received-grants-in-2025-30381
3. Weir C. J. Language testing and validation: An evidence-based approach. - Palgrave Macmillan, 2005. - 303 p.
4. McNamara T. Language testing. - Oxford University Press, 2000. - 152 p.
5. Advancing validation research in language assessment: Current issues and future directions / G. J. Ockey, A. J. Kunnan, A. Brown // Language Testing. - 2021. - Vol. 38, № 4. - P. 493–509. - https://doi.org/10.1177/02655322211016432

6. Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning // American Psychologist. - 1995. - Vol. 50, № 9. - P. 741–749. - https://doi.org/10.1037/0003-066X.50.9.741

7. Kane M. T. Validating the interpretations and uses of test scores // Journal of Educational Measurement. - 2013. - Vol. 50, № 1. - P. 1-73. - https://doi.org/10.1111/jedm.12000

8. Kane M. Validating the interpretation and use of test scores // Educational Measurement: Issues and Practice. - 2022. - Vol. 41, № 2. - P. 6-17. - https://doi.org/10.1111/emip.12464

9. Bachman L. F., Palmer A. S. Language assessment in practice. - Oxford University Press, 2010. - 512 p.

10. Council of Europe. Common European framework of reference for languages: Learning, teaching, assessment – Companion volume. - Council of Europe Publishing, 2018. - 235 p.

11. OECD. Strengthening national examinations: The experience of Kazakhstan. - OECD Publishing, 2020. - https://doi.org/10.1787/0f2b6d9a-en

12. Kirichok O., Amankeldiyeva S., Nussenov Z. Students' perceptions of unified national test scores, state scholarships, and academic performance // World Journal on Educational Technology: Current Issues. - 2025. - Vol. 17, № 1. - P. 1-13. - https://doi.org/10.18844/wjet.v17i1.7695

13. Mingisheva N. A. Development and challenges of standardized testing in Kazakhstan: Transition from national to international standards // Bulletin of Al-Farabi Kazakh National University. Pedagogical Sciences Series. - 2023. - Vol. 76, № 3. - P. 94–103. - https://doi.org/10.26577/JES.2023.v76.i3.08

14. Advancing digital item development in Asian assessment systems: Insights from Kazakhstan's Unified National Testing / N. Didarbekova, Y. Sarmurzin, S. Altybaeva [et al.] // Frontiers in Education. - 2025. - Vol. 10. - Article 1654674. - https://doi.org/10.3389/feduc.2025.1654674

15. Jumabayeva Z. The key drivers of the Unified National Test in Kazakhstan: A critical analysis of its impact on school leavers // NUGSE Research in Education. - 2016. - Vol. 1, № 2. - P. 16–20.

16. Turlanbekova Z. Applicants' or students' perception of Kazakhstani universities admission tests and exams // Sociology and Anthropology. - 2015. - Vol. 3, № 8. - P. 414-423. - https://doi.org/10.13189/sa.2015.030805

17. Validation of the Unified National Testing in Kazakhstan: Evidence-based framework for fair testing / B. S. Abdrasilov, L. E. Shinetova, A. E. Ivanova [et al.] // Educational Studies Moscow. - 2026. - https://doi.org/10.17323/vo-2026-27950

18. Bachman L. F. Fundamental considerations in language testing. - Oxford University Press, 1990. - 420 p.

19. Kunnan A. J. Test fairness // European language testing in a global context / Ed. by M. Milanovic, C. Weir. — Cambridge University Press, 2004. - P. 27–48.

20. McNamara T., Roever C. Language testing: The social dimension. - Blackwell Publishing, 2006. - 291 p.

21. Chapelle C. A. Validity in language assessment // Annual Review of Applied Linguistics. - 1999. - Vol. 19. - P. 254–272. - https://doi.org/10.1017/S0267190500001135

22. North B. The development of a common framework scale of language proficiency. - Peter Lang, 2000. - 380 p.

23. Alderson J. C. The CEFR and the need for more research // The Modern Language Journal. - 2007. - Vol. 91, № 4. - P. 659-663. - https://doi.org/10.1111/j.1540-4781.2007.00627_4.x

24. Relating examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR): A manual / N. Figueras, B. North, S. Takala [et al.] // Language Testing. - 2005. - Vol. 22, № 3. - P. 261–279. - https://doi.org/10.1191/0265532205lt309oa

25. Khalifa H., Weir C. J. Examining reading: Research and practice in assessing second language reading. - Cambridge University Press, 2009. - 346 p.

26. Xi X. How do we go about investigating test fairness? // Language Testing. - 2010. - Vol. 27, № 2. - P. 147–170. - https://doi.org/10.1177/0265532209349465

Ә.Ү. Серікбаева, Д.М. Акижанова, Б.С. Абдрасилов, Л.Е. Шинетова

ПСИХОМЕТРИКАЛЫҚ ТҰРҒЫДАН ЖОҒАРЫ МӘРТЕБЕЛІ АҒЫЛШЫН ТІЛІ ТЕСТІНІҢ ВАЛИДТІЛІГІН ДӘЛЕЛДЕУ: ҚАЗАҚСТАННЫҢ ҰЛТТЫҚ БІРЫҢҒАЙ ТЕСТІЛЕУІН РАШ МОДЕЛ

Ұлттық бірыңғай тестілеу (ҰБТ) жоғары оқу орындарына түсу үшін маңызды рөл атқаратын жоғары мәртебелі емтихан болып табылады. Жоғары мәртебелі тестілеуде нәтижелерді түсіндіруді негіздеу үшін психометрикалық дәлелдемелер қажет, алайда ҰБТ-ның психометрикалық қасиеттері әлі де жеткіліксіз зерттелген. Зерттеу ҰБТ-ның ағылшын тілі бөлімін психометрикалық тұрғыдан талдауға бағытталған, оның ішінде сенімділік, конструкттық валидтілік, тапсырмалардың қиындығы мен дискриминативтілік индекстері, сондай-ақ тестің оқу бағдарламасына және CEFR деңгейлеріне сәйкестігі қарастырылады. Бұл зерттеу тестің тілді меңгерудің нақты деңгейін қаншалықты өлшейтінін зерттейді. Ағылшын тілін таңдаған 10–11 сынып оқушылары арасында (n = 212) пилоттық сандық зерттеу жүргізілді. Классикалық тест теориясы (CTT) аясында дискриминативтілік индекстері есептелсе, сенімділік, сәйкестік статистикасы мен тапсырмалардың бағытталғандығы Раш моделінің көмегімен зерттелді. 10-шы және 11-ші сынып нәтижелерін салыстыру үшін тапсырмалардың дифференциалдық қызмет етуі (DIF) қолданылды. Психометрикалық нәтижелер ҰБТ-ның ағылшын тілі компонентінің жоғары ішкі сәйкестігін көрсетті (Кронбах α = .879; ω = .898). Тапсырмалардың көпшілігі қолайлы өлшеу көрсеткіштерін көрсетті (адам сенімділігі = .836; тапсырма сенімділігі = .972). Дегенмен, грамматика және оқу бөлімдерінде бірнеше тапсырма салыстырмалы түрде жеңіл болып шықты. Мұғалімдер сауалнамасы өнімді дағдылардың жеткіліксіз бағаланатынын және рецептивті дағдылардың айтарлықтай басым екенін көрсетті (M = 4.50). ҰБТ-ның ағылшын тілі бөлімі сенімділік пен құрылымдық тұрақтылық тұрғысынан жалпы қолайлы деңгейді көрсетеді. Алайда, коммуникативтік құзыреттілік пен өнімді дағдылардың шектеулі түрде ұсынылғаны байқалады. Сондықтан нәтижелер ҰБТ балдарын IELTS сияқты стандартталған тестілермен және CEFR деңгейлерімен сәйкестендіру қажеттілігін көрсетеді.

Түйін сөздер: Ұлттық бірыңғай тестілеу, ағылшын тілін бағалау, конструкттық валидтілік, Раш моделі.

Ә.Ү. Серікбаева, Д.М. Акижанова, Б.С. Абдрасилов, Л.Е. Шинетова

ПСИХОМЕТРИЧЕСКАЯ ВАЛИДИЗАЦИЯ ВЫСОКОСТАВОЧНОГО ТЕСТА ПО АНГЛИЙСКОМУ ЯЗЫКУ: АНАЛИЗ НА ОСНОВЕ МОДЕЛИ РАША (ЕНТ КАЗАХСТАНА)

Единое национальное тестирование (ЕНТ) является высокоставочным экзаменом, выполняющим роль ключевого критерия для поступления в вузы. Для обоснования интерпретации результатов высокоставочного тестирования необходимы психометрические доказательства, однако психометрические свойства ЕНТ остаются недостаточно изученными. Исследование направлено на анализ английского раздела ЕНТ с психометрической точки зрения, включая надежность, конструктную валидность, трудность и дискриминативность заданий, а также соответствие теста учебной программе и уровням CEFR. В работе изучается, насколько тест адекватно измеряет реальный уровень владения языком. Проведено пилотное количественное исследование с участием 212 учащихся 10–11 классов, выбравших английский язык в качестве профильного предмета. В рамках классической теории тестирования (CTT) были рассчитаны индексы дискриминативности, а надежность, показатели соответствия и нацеленность заданий изучались с использованием модели Раша. Для сравнения результатов 10-х и 11-х классов применялся анализ дифференциального функционирования заданий (DIF). Психометрические результаты показали высокую внутреннюю согласованность английского компонента ЕНТ (α Кронбаха = .879; ω = .898). Большинство заданий продемонстрировали приемлемые показатели измерения (надежность испытуемых = .836; надежность заданий = .972). Однако несколько заданий в разделах грамматики и чтения оказались относительно легкими. Опрос учителей выявил недостаточную оценку продуктивных навыков и сильное доминирование рецептивных навыков (M = 4.50). Английский раздел ЕНТ в целом демонстрирует приемлемый уровень надежности и структурной стабильности. Однако выявлена

ограниченная репрезентация коммуникативной компетенции и продуктивных навыков. Полученные результаты подчеркивают необходимость согласования баллов ЕНТ с результатами стандартизированных тестов, таких как IELTS, и привязки к уровням CEFR.

Ключевые слова: Единое национальное тестирование, оценка английского языка, конструктная валидность, модель Раша.

References

1. Brown, H. D. (2004). *Teaching by principles: An interactive approach to language pedagogy* (2nd ed.). Longman.
2. Prime Minister's Office of the Republic of Kazakhstan. (2025). *Quality and accessible education: Over 93,000 students received grants in 2025*. https://primeminister.kz/en/news/quality-and-accessible-education-over-93000-students-received-grants-in-2025-30381
3. Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.
4. McNamara, T. (2000). *Language testing*. Oxford University Press.
5. Ockey, G. J., Kunnan, A. J., & Brown, A. (2021). Advancing validation research in language assessment: Current issues and future directions. *Language Testing*, 38(4), 493–509. https://doi.org/10.1177/02655322211016432
6. Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741
7. Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000
8. Kane, M. (2022). Validating the interpretation and use of test scores. *Educational Measurement: Issues and Practice, 41*(2), 6–17. https://doi.org/10.1111/emip.12464
9. Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.
10. Council of Europe. (2018). *Common European framework of reference for languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing.
11. OECD. (2020). *Strengthening national examinations: The experience of Kazakhstan*. OECD Publishing. https://doi.org/10.1787/0f2b6d9a-en
12. Kirichok, O., Amankeldiyeva, S., & Nussenov, Z. (2025). Students' perceptions of unified national test scores, state scholarships, and academic performance. *World Journal on Educational Technology: Current Issues, 17*(1), 1–13. https://doi.org/10.18844/wjet.v17i1.7695
13. Mingisheva, N. A. (2023). Development and challenges of standardized testing in Kazakhstan: Transition from national to international standards. *Bulletin of Al-Farabi Kazakh National University. Pedagogical Sciences Series, 76*(3), 94–103. https://doi.org/10.26577/JES.2023.v76.i3.08
14. Didarbekova, N., Sarmurzin, Y., Altybaeva, S., Abdrasilov, B., & Karkenova, A. (2025). Advancing digital item development in Asian assessment systems: Insights from Kazakhstan's Unified National Testing. *Frontiers in Education, 10*, Article 1654674. https://doi.org/10.3389/feduc.2025.1654674
15. Jumabayeva, Z. (2016). The key drivers of the Unified National Test in Kazakhstan: A critical analysis of its impact on school leavers. *NUGSE Research in Education, 1*(2), 16–20.
16. Turlanbekova, Z. (2015). Applicants' or students' perception of Kazakhstani universities admission tests and exams. *Sociology and Anthropology, 3*(8), 414–423. https://doi.org/10.13189/sa.2015.030805
17. Abdrasilov, B. S., Shinetova, L. E., Ivanova, A. E., Kardanova, E. Y., Beisenova, Z. Z., Zhabayeva, G. K., Orlov, D., & Ashirbek, G. K. (2026). Validation of the Unified National Testing in Kazakhstan: Evidence-based framework for fair testing. *Educational Studies Moscow*. https://doi.org/10.17323/vo-2026-27950.
18. Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
19. Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context* (pp. 27–48). Cambridge University Press.
20. McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Blackwell Publishing.

21. Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254–272. https://doi.org/10.1017/S0267190500001135

22. North, B. (2000). *The development of a common framework scale of language proficiency*. Peter Lang.

23. Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal, 91*(4), 659–663. https://doi.org/10.1111/j.1540-4781.2007.00627_4.x

24. Figueras, N., North, B., Takala, S., Verhelst, N., & Van Avermaet, P. (2005). Relating examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR): A manual. *Language Testing, 22*(3), 261–279. https://doi.org/10.1191/0265532205lt309oa

25. Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.

26. Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147–170. https://doi.org/10.1177/0265532209349465

Information about authors:

Serikbayeva Assiya - doctoral student, Al-Farabi Kazakh National University, Almaty, Republic of Kazakhstan, e-mail: assiyaserikbayeva237@gmail.com

Akizhanova Diana Muratovna - PhD, Head of Translation Theory and Practice Department, acting Associate Professor, L.N.Gumilyov Eurasian National University, Astana, Republic of Kazakhstan, e-mail: diana.akizhanova@gmail.com

Shinetova Lyazzat Ermekovna - Master's degree, Headof the Laboratory of Scientific Researchand Psychometrics at the National Testing Center of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Astana, Republic of Kazakhstan, e-mail: shinetovalyazzat24@gmail.com

Abdrasilov Bolatbek Serikbayevich - Candidate of Physical and Mathematical Sciences, Doctor of Biological Sciences, Academician of the NAS RK, Director of the National Testing Centerof the Ministry of Science and Higher Education of the Republic of Kazakhstan, Astana, Republic of Kazakhstan, e-mail: uto@testcenter.kz

Авторлар туралы мәлімет:

Серікбаева Әсия Үмбетбайқызы - Түркітану және тіл теориясы кафедрасының 2-курс докторанты, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан Республикасы, e-mail: assiyaserikbayeva237@gmail.com

Акижанова Диана Муратовна - PhD, аударма теориясы мен практикасы кафедрасының меңгерушісі, доцент м.а., Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан Республикасы, e-mail: diana.akizhanova@gmail.com

Шинетова Ляззат Ермековна - магистр, ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК Жоғары, Ғылыми зерттеулер мен психометрика зертханасының меңгерушісі, Астана, Қазақстан Республикасы, e-mail: shinetovalyazzat24@gmail.com

Абдрасилов Болатбек Серикбаевич - физика-математика ғылымдарының кандидаты, биология ғылымдарының докторы, ҚР ҰҒА академигі, ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК директоры, Астана, Қазақстан Республикасы, e-mail: uto@testcenter.kz

Сведения об авторах:

Серікбаева Әсия Үмбетбайқызы - докторант Казахского национального университета имени аль-Фараби, Алматы, Республика Казахстан, e-mail: assiyaserikbayeva237@gmail.com

Акижанова Диана Муратовна - PhD, и.о.доцента, заведующий кафедрой теории и практики перевода, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Республика Казахстан, e-mail: diana.akizhanova@gmail.com

Шинетова Ляззат Ермековна - магистр, заведующий лабораторией научных исследований и психометрики РГП на ПХВ «Национальный центр тестирования» МНВО РК, Астана, Республика Казахстан, e-mail: shinetovalyazzat24@gmail.com

Абдрасилов Болатбек Серикбаевич - кандидат физико-математических наук, доктор биологических наук, академик НАН РК, директор РГП на ПХВ «Национальный центр тестирования» МНВО РК, Астана, Республика Казахстан, e-mail: uto@testcenter.kz