

ПЕДАГОГИКАЛЫҚ ӨЛШЕМДЕР

ҒЫЛЫМИ-ПРАКТИКАЛЫҚ ЖУРНАЛЫ

№1/2026



ПЕДАГОГИЧЕСКИЕ
ИЗМЕРЕНИЯ

НАУЧНО-ПРАКТИЧЕСКИЙ ЖУРНАЛ

PEDAGOGICAL
MEASUREMENTS

JOURNAL OF RESEARCH AND PRACTICE

Қазақстан Республикасының Ғылым және жоғары білім министрлігі «Ұлттық тестілеу орталығы»
шаруашылық жүргізу құқығындағы республикалық мемлекеттік кәсіпорны

«ПЕДАГОГИКАЛЫҚ ӨЛШЕМДЕР» ҒЫЛЫМИ-ПРАКТИКАЛЫҚ ЖУРНАЛЫ

Республиканское государственное предприятие на праве хозяйственного ведения «Национальный центр
тестирования» Министерства науки и высшего образования Республики Казахстан

НАУЧНО-ПРАКТИЧЕСКИЙ ЖУРНАЛ «ПЕДАГОГИЧЕСКИЕ ИЗМЕРЕНИЯ»

Republican state enterprise under the right of economic management «National Testing Center»
of the Ministry of Science and Higher Education of the Republic of Kazakhstan

JOURNAL OF RESEARCH AND PRACTICE «PEDAGOGICAL MEASUREMENTS»

№1(5) / 2026

*Жылына 4 рет шығарылады
2025 жылдан бастап шығады*

**Астана
2026**

ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК

«ПЕДАГОГИКАЛЫҚ ӨЛШЕМДЕР»
ғылыми-практикалық журналы*Бас редакторы:***Абдрасилов Б.С.**

ҚР ҒЖБМ Ұлттық тестілеу орталығының директоры, ҚР ҰҒА академигі

*Жауапты редактор:***Шинетова Л.Е.***Редакция алқасы:*

Matthew Courtney – PhD, Назарбаев Университет

Карданова Е. Ю. – ф.-м.ғ.к., доцент, «Жоғары экономика мектебі» Ғылыми-зерттеу университеті, Ресей

Иванова А.Е. – Білім беру ғылымдарының кандидаты, доцент, «Жоғары экономика мектебі» Ғылыми-зерттеу университеті, Ресей

Зулпыхар Ж.Е. – п.ғ.к., Л.Н. Гумилев атындағы Еуразия ұлттық университеті

Калкеева К.Р. – п.ғ.к., профессор, Л.Н. Гумилев атындағы Еуразия ұлттық университеті

Кошербаева А.Н. – п.ғ.д., профессор, Абай атындағы Қазақ ұлттық педагогикалық университеті

Танирбергенова А.Ш. – п.ғ.к., Ы.Алтынсарин атындағы Ұлттық білім академиясы

Абаева Г.А. – п.ғ.к., доцент, Абай атындағы Қазақ ұлттық педагогикалық университеті

Абсатова М.А. – п.ғ.к., профессор, Абай атындағы Қазақ ұлттық педагогикалық университеті

Акижанова Д.М. – PhD, Л.Н. Гумилев атындағы Еуразия ұлттық университеті

Бегимбетова Г.А. – Universitas Negeri Yogyakarta, Индонезия

Дидарбекова Н.А. – ф.ғ.к., Ұлттық тестілеу орталығы

Temur Nezir – PhD, профессор, Гази университеті, Түркия

*Техникалық редактор:***Бейсенова Ж.Ж.****«Педагогикалық өлшемдер» ғылыми-практикалық журналы. – 2026. – 5 т. – 73 б.**Редакцияның мекенжайы: 010000, Қазақстан Республикасы, Астана қ., Жеңіс даңғылы, 60, тел. 8(7172)799-889, і.н.157, e-mail: naukauto24@gmail.com, сайт: <https://testcenter-zhurnal.kz/>.

Қазақстан Республикасы Мәдениет және ақпарат министрлігімен тіркелген. №KZ54VPY00116912 есепке қою туралы куәлігі.

МАЗМҰНЫ / СОДЕРЖАНИЕ / CONTENT

БІЛІМ БЕРУ НӘТИЖЕЛЕРІН БАҒАЛАУДЫҢ ЗАМАНАУИ ТӘСІЛДЕРІ ЖӘНЕ БІЛІМ БЕРУ САПАСЫН БАСҚАРУ / СОВРЕМЕННЫЕ ПОДХОДЫ К ОЦЕНКЕ ОБРАЗОВАТЕЛЬНЫХ РЕЗУЛЬТАТОВ И УПРАВЛЕНИЕ КАЧЕСТВОМ ОБРАЗОВАНИЯ / MODERN APPROACHES TO EDUCATIONAL OUTCOMES ASSESSMENT AND EDUCATION QUALITY MANAGEMENT

Ш.Б. Алтыбаева, М.Ш. Тасбулатова, Г.Е. Нұрмұханова, Р.Б. Аубакирова ҰБТ-ДАҒЫ ҚАЗАҚ ТІЛІН ӨЛШЕУ МОДЕЛІНІҢ ВАЛИДТІЛІГІ: КОГНИТИВТІ ДЕҢГЕЙЛЕР МЕН СТАНДАРТТАРДЫ ВЕРИФИКАЦИЯЛАУ.....	5
A. Serikbayeva, D. Akizhanova, L. Shinetova, B. Abdrasilov PSYCHOMETRIC VALIDATION OF A HIGH-STAKES ENGLISH LANGUAGE TEST: A RASCH-BASED ANALYSIS OF KAZAKHSTAN'S UNIFIED NATIONAL TESTING.....	18
К.Т. Туралбаева, Г.С. Джарасова, И.Ж. Жуматова, Г.Т. Сагнаева, Б.К. Асимова МАТЕМАТИКА ПӘНІ БОЙЫНША ТЕСТТІҢ ПСИХОМЕТРИЯЛЫҚ ТАЛДАУЫ: РАШ МОДЕЛІ НЕГІЗІНДЕ БАҒАЛАУ.....	33
R. Ismailzhanova, G. Zhabayeva, Zh. Beisenova, D. Orlov INTERNATIONAL STUDENTS AT EMI UNIVERSITIES IN KAZAKHSTAN: EXPERIENCES, CHALLENGES, AND COPING STRATEGIES.....	43
А.Ж. Аuezханова, Т.С. Ерсултанова, А.Қ. Қуандық, С.Ә. Әлен 2025 ЖЫЛЫ ПББ АПРОБАЦИЯСЫ АЯСЫНДА ӨТКІЗІЛГЕН ТЕСТІЛЕУ НӘТИЖЕЛЕРІНІҢ ПСИХОМЕТРИЯЛЫҚ ТАЛДАУЫ.....	53

ЖАСАНДЫ ИНТЕЛЛЕКТ ЖӘНЕ БІЛІМ БЕРУДІ БАҒАЛАУДАҒЫ ИННОВАЦИЯ / ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ИННОВАЦИИ В ОЦЕНИВАНИИ ОБРАЗОВАНИЯ / ARTIFICIAL INTELLIGENCE AND INNOVATIONS IN EDUCATIONAL ASSESSMENT

К.В. Тарасова ПСИХОМЕТРИКА И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: НОВАЯ ПАРАДИГМА ОБРАЗОВАТЕЛЬНОГО ОЦЕНИВАНИЯ.....	63
---	----

**БІЛІМ БЕРУ НӘТИЖЕЛЕРІН БАҒАЛАУДЫҢ ЗАМАНАУИ ТӘСІЛДЕРІ ЖӘНЕ БІЛІМ
БЕРУ САПАСЫН БАСҚАРУ**

**СОВРЕМЕННЫЕ ПОДХОДЫ К ОЦЕНКЕ ОБРАЗОВАТЕЛЬНЫХ РЕЗУЛЬТАТОВ И
УПРАВЛЕНИЕ КАЧЕСТВОМ ОБРАЗОВАНИЯ**

**MODERN APPROACHES TO EDUCATIONAL OUTCOMES ASSESSMENT AND
EDUCATION QUALITY MANAGEMENT**

Ш.Б. Алтыбаева¹, М.Ш.Тасбулатова^{2*}, Г.Е. Нұрмұханова³, Р.Б.Аубакирова⁴

^{1,2,3}ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, Астана қ, Қазақстан Республикасы

⁴Астана қаласы әкімдігінің «№82 «Дарын» мамандандырылған мектеп-лицейі» КММ,

Астана қ, Қазақстан Республикасы

*e-mail: tasbulatova.makhabbat@mail.ru

¹ORCID 0000-0003-0306-861X, ²ORCID 0009-0009-2938-5307,

³ORCID 0009-0007-3665-0602, ⁴ORCID 0009-0008-7823-4946

ҰБТ-ДАҒЫ ҚАЗАҚ ТІЛІН ӨЛШЕУ МОДЕЛІНІҢ ВАЛИДТІЛІГІ: КОГНИТИВТІ ДЕҢГЕЙЛЕР МЕН СТАНДАРТТАРДЫ ВЕРИФИКАЦИЯЛАУ

Бұл зерттеу Ұлттық бірыңғай тестілеудің (ҰБТ) «Қазақ тілі» бейіндік пәні бойынша қолданылатын өлшеу моделінің валидтілігін талдауға бағытталған. Зерттеудің өзектілігі елімізде жоғары оқу орындарына түсу емтиханының білім сапасын басқарудағы стратегиялық рөлімен айқындалады. Зерттеу мақсаты тест мазмұнының оқу бағдарламасына сәйкестігін, когнитивтік деңгейлердің теңгерімін және психометриялық сапасын бағалау болып табылады. Мазмұндық салыстыру мен Блум таксономиясы бойынша когнитивтік верификация, сондай-ақ Раш моделіне негізделген психометриялық талдау (N=257) әдіснамалық негіз ретінде қолданылды. Нәтижелер тесттің мазмұндық тұрғыдан оқу бағдарламасына жалпы сәйкес келетінін және ішкі сенімділік көрсеткішінің жоғары екенін көрсетті (Reliability = 0,81; Item Separation = 7,58). Тапсырмалардың 70%-ы төменгі ретті ойлау дағдыларына (білу, түсіну) тиесілі болса, жоғары деңгейдегі талдау, бағалау және жинақтау дағдылары 10%-дан аз үлесті құрайды. Тапсырмалар қиындығы -2,89-дан +1,77 логитке дейінгі кең ауқымды қамтиды, алайда айнымалылар картасы жоғары қабілетті талапкерлер (+1,5 логиттен жоғары) үшін тапсырмалар тапшылығын айқын көрсетті. Қорытындылай келе тесттің конструкттық валидтілігін арттыру үшін жоғары когнитивтік деңгейдегі тапсырмалар үлесін ұлғайту, QTI стандарты негізінде интерактивті форматтарды енгізу және компьютерлік бейімделген тестілеу технологиясына кезең-кезеңімен көшу ұсынылады.

Түйін сөздер: ҰБТ, қазақ тілі, валидтілік, Раш моделі, Блум таксономиясы, QTI стандарты.

Кіріспе

Қазіргі білім беруде жоғары ставкалы емтихандар тек оқу жетістіктерін тіркеп қоймайды. Олар білім сапасын басқарудың, академиялық іріктеудің, әлеуметтік әділеттілікті қамтамасыз етудің маңызды құралына айналды. Осы тұрғыдан Қазақстандағы Ұлттық бірыңғай тестілеу (ҰБТ) стратегиялық мәнге ие – ол талапкерлердің жоғары оқу орнына түсуге дайындық деңгейін айқындайды.

Қазіргі психометрия ғылымы өлшеу құралының сапасы оның мақсатына, қолданыс аясына және өлшенетін конструкттың теориялық негізіне сәйкестігімен тікелей байланысты екенін көрсетіп отыр [1, 2]. Яғни, тест нәтижелерінің валидтілігі – бұл балдардың қандай когнитивтік қабілеттерді білдіретініне және сол нәтижелерге сүйеніп қабылданатын шешімдердің қаншалықты негізді екеніне тәуелді.

С. Мессик [1] валидтілікті тек статистикалық сипаттама ретінде емес, тестті қолдану мақсаты, нәтижелерді түсіндіру және олардың әлеуметтік салдары бірлігінде қарастыру керек деп есептейді. Ол мазмұндық, құрылымдық, сыртқы және салдарлық аспектілерді өлшеу құралын бағалаудың ажырамас стандарттары деп санайды. Ал Р. Мислеви [2] ұсынған Evidence-Centered Design тұжырымдамасы бағалау жүйесін білім алушының жасырын қабілеттері, бақыланатын әрекеттері және олардың арасындағы эмпирикалық дәлелдермен байланыстырады.

Экономикалық ынтымақтастық және даму ұйымының зерттеулерінде [3] ұлттық емтихандардың екі негізгі қызметі атап көрсетіледі: біріншісі – білім алушының білім деңгейін растау, екіншісі – одан әрі академиялық траекториясын анықтау үшін іріктеу. Бұл екі қызметті тиімді орындау үшін өлшеу моделі тест мазмұнына, құрылымына және бағалау шешімдерінің салдарына сай болуы керек.

ҰБТ нәтижелері талапкерлердің келесі оқу жолына да, мемлекеттік гранттардың бөлінуіне де тікелей әсер етеді. Сондықтан тесттің қалай құрастырылғаны – оның оқу бағдарламасындағы білім мен дағдыларды шынымен өлшейтіні, нәтижелердің әділдігі, қабылданатын шешімдердің орындылығы – басты назарда болуы тиіс. Бұл ҰБТ-дағы әр пән бойынша өлшеу моделін жүйелі түрде қайта қарастыру қажеттігін көрсетеді.

ҰБТ құрылымы үш міндетті және екі бейіндік пәнді қамтиды [4]. Тестілеуді үш тілде, оқу жылы ішінде бірнеше рет тапсыруға болады – бұл жүйеге белгілі бір икемділік береді.

ҰБТ пәндері бойынша өлшеу модельдерінің валидтілігін зерттеу – білім сапасын басқарудың маңызды бағыттарының бірі. Ұлттық тестілеу орталығы осы бағытта жұмыс істеп келеді. Мәселен, Алтыбаева [5] информатика пәні бойынша тест спецификациясында Блум таксономиясының «қолдану» және «жасау» деңгейлеріне басымдық берілгенін атап өтті. Бұл өзге де бейіндік пәндер бойынша валидтілік зерттеулерін жүргізу қажеттігін көрсетеді.

Біз бұл зерттеуде ҰБТ-дағы қазақ тілін бағалау моделінің мазмұндық және құрылымдық валидтілігін, оның заманауи технологиялық стандарттар негізінде жаңғырту мүмкіндіктерін қарастырдық. Тест тапсырмаларының оқу бағдарламасына сәйкестігін, әсіресе жоғары ретті ойлау дағдыларын (талдау, бағалау, шығармашылық) қаншалықты өлшейтінін анықтауды мақсат еттік [6, 7]. Сонымен қатар тест тапсырмаларының Блум таксономиясы бойынша когнитивтік деңгейлерде қалай таралғанын, психометриялық сапа көрсеткіштерін (сенімділік, дискриминативтілік, күрделілік) және QTI сияқты халықаралық стандарттарға сәйкес технологиялық трансформация мүмкіндіктерін талдадық [8]. Осы мәселелерді Мессиктің валидтілік тұжырымдамасы аясында зерделеп, ҰБТ-дағы қазақ тілін бағалауды теориялық және практикалық тұрғыдан жетілдіруге ғылыми негіз дайындауға тырыстық.

Зерттеу материалдары мен әдістері

Зерттеу барысында ҰБТ қазақ тілі бойынша өлшеу моделінің валидтілігі мен технологиялық трансформациясын бағалауды көздедік. Қойылған міндеттерді шешу үшін сандық психометриялық талдау мен тест мазмұнын сараптауды біріктіретін аралас әдіснаманы қолдандық [9, 10].

Алдымен нормативтік құжаттарға сапалық талдау жасадық. «Қазақ тілі» пәнінің типтік оқу бағдарламаларын, ондағы күтілетін нәтижелерді, олардың когнитивтік күрделілігін қарастырдық. Сонымен қатар 2026 жылға арналған тест спецификациясындағы тақырыптық ауқым, тапсырмалар пішімі және күрделілік деңгейлерінің үлесі (А – 50%, В – 30%, С – 20%) салыстырмалы түрде талданды.

Тест тапсырмаларының когнитивтік талаптарын бағалау үшін Блум таксономиясының классикалық нұсқасын қолдандық [11, 12]. Әр тапсырманы білу, түсіну, қолдану, талдау, синтез, бағалау деңгейлерінің қайсысына сәйкес келетінін анықтап, олардың таралу құрылымын бағаладық.

Негізгі материал ретінде ҰТО-ның ресми базасынан алынған «Қазақ тілі» тапсырмаларын пайдаландық. Үш форматтағы тапсырманы талдадық:

- бір немесе бірнеше дұрыс жауапты таңдау (MC) – грамматика, лексика, орфография;
- контекстік тапсырмалар (CONT) – мәтінді оқу, түсіну, контекстпен жұмыс;
- сәйкестікті анықтау (SOOTV) – логикалық, мағыналық байланыстар.

Деректер статистикалық өңдеуге жарамды форматта шығарылды, әр тапсырма мен тестіленушіге бірегей идентификатор берілді.

Психометриялық талдау үшін Раштың бірпараметрлік моделі мен Winsteps 5.2.3 бағдарламасын қолдандық [13]. Талдау 2025 жылғы ҰБТ апробациясының анонимделген деректеріне негізделді. Іріктеме кездейсоқ стратификациялық әдіспен құрастырылды, барлық өңірден 257 мектеп бітірушіні қамтыды. Іріктеуде гендерлік арақатынас (шамамен 55% әйелдер, 45% ерлер) және елді мекен түрі (қала/ауыл) ескерілді. Толық жауап берген бітірушілердің деректері ғана енгізілді.

Зерттеу барысында тесттің ішкі сәйкестік сенімділігін Кронбах альфа коэффициентімен (α) [14], тапсырмалардың күрделілік индексі (p -мәні) және дискриминациялық қабілетімен бағаладық. Раш моделіне сәйкестікті Infit/Outfit Mean-Square (MNSQ) статистикасымен тексердік. Қолайлы диапазон ретінде 0,5-1,5 аралығын алдық [15].

Психометриялық талдау ҰТО-дан алынған анонимді деректерге негізделген. Құжаттық талдау ашық дереккөздерге (ресми сайттар, жарияланған жинақтар) сүйенді. Нәтижелер тек жиынтық түрде ұсынылды, жеке талапкерлер немесе мекемелер туралы мәлімет ашылған жоқ.

Біздің қолданған әдістер тест мазмұнының оқу бағдарламасына сәйкестігін, когнитивтік деңгейлердің таралуын және психометриялық сапаны кешенді бағалауға, сондай-ақ қазақ тілін бағалау моделін технологиялық жаңғырту әлеуетін талдауға мүмкіндік берді.

Нәтижелер және талқылау

ҰБТ қазақ тілі пәнінің спецификациясы мен оқу бағдарламасын салыстырдық. Талдау көрсеткендей, тест мазмұны бағдарламаға жалпы сәйкес келеді. Спецификациядағы бес негізгі блок (орфография, лексика, грамматика, пунктуация, оқылым) бағдарламадағы тақырыптарды толық қамтиды. Оқылым бөлімінің функционалдық сауаттылыққа бағдарлануы да бағдарламадағы коммуникативтік және талдамалық құзыреттерді бағалауға мүмкіндік береді.

Дегенмен, бірқатар алшақтықтар да бар. Бағдарламада басымдық берілген жоғары деңгейлі когнитивтік дағдылар (мысалы, графикалық мәліметтерді салыстырмалы талдау, академиялық жанрларды талдау) тест спецификациясында міндетті тапсырма түрлері ретінде нақты көрсетілмеген. Бұл дағдылар бағалау практикасында шектеулі қамтылады.

Зерттеу барысында қазіргі тест моделінің күрделі коммуникативтік және талдамалық құзыреттерді бағалаудағы шектеулерін анықтадық. ҰБТ электронды форматта өткізілетіндіктен, QTI стандартын енгізу арқылы тапсырмаларды интерактивті етуге мүмкіндік бар.

Қазақ тілі пәні 40 тест тапсырмасынан тұрады: 25 бір дұрыс жауапты таңдау, 5 бір дұрыс жауапты таңдаудан тұратын 1 контекст, 5 бір немесе бірнеше дұрыс жауапты таңдау, 5 сәйкестікті анықтау [7]. Тапсырмаларды орындауға орта есеппен 2 минут беріледі.

Дәстүрлі көп таңдаулы тапсырмалардың (MCQ) мүмкіндігі шектеулі – олар негізінен жаттанды білімді тану немесе қарапайым қолдану деңгейінде тексереді. Ал спецификацияда жоғары когнитивтік деңгейдегі тапсырмалар күрделі білім мен дағдыларды, мәселелерді шешу үшін оларды біріктіруді, деректерді талдауды, пайымдау мен тұжырымдарды негіздеуді талап етеді. QTI сияқты ашық стандарттар интерактивті тапсырмаларды енгізуге жол ашады. Мәтіндік фрагменттерді ретке келтіру, графикалық деректермен жұмыс, элементтерді сәйкестендіру – мұндай тапсырмалар жоғары когнитивтік дағдыларды (талдау, синтез, пайымдау) бағалаудың валидтілігін арттыра алады. Ал компьютерлік бейімделген тестілеу әр талапкерге өз деңгейіне сәйкес тест ұсынып, бағалау дәлдігін арттырады.

Біздің байқауымызша, ҰБТ-ның қазіргі моделі кешенді құзыреттерді бағалауда шектеулі. Валидтілікті арттыру үшін тестті халықаралық цифрлық стандарттармен (QTI) үйлестіріп, адаптивтік технологияларды енгізу қажет.

Тест тапсырмаларын Блумның жаңартылған таксономиясы (Remembering, Understanding, Applying, Analyzing, Evaluating, Creating) негізінде талдадық. Мақсат – спецификацияда жарияланған бағалау мақсаттары мен нақты тапсырмалар арқылы өлшенетін когнитивтік үдерістердің сәйкестігін анықтау, сондай-ақ жоғары деңгейлі ойлау дағдыларының (HOTS) тестте қаншалықты қамтылғанын бағамдау.

Талдау нәтижелері бойынша, спецификацияда көрсетілген «пайдалану», «дәлелдер мен ұғымдарды пайдалану» сияқты әрекеттер жабық форматтағы тапсырмаларда негізінен төменгі деңгейлі когнитивтік әрекеттерге сәйкес келеді. Тестіленуші көбіне тілдік ережелерді тану, еске түсіру, қарапайым интерпретациялаумен шектеледі. Бұл тапсырмалар Remembering және Understanding деңгейлеріне жатады.

Спецификациядағы «талдау», «салыстыру», «жалпылау» сияқты етістіктер теориялық тұрғыдан Applying және Analyzing деңгейлеріне сәйкес келеді. Бірақ жабық типтегі тапсырмаларда бұл әрекеттер тестіленушінің өзіндік аналитикалық ойлауын емес, дайын нәтижені тануын талап етеді. Сондықтан мұндай тапсырмалардың нақты когнитивтік сұранысы Understanding немесе стандартты Applying деңгейінен аспайды.

Спецификацияда көрсетілген «біріктіру», «пайымдау», «негіздеу» сияқты күрделі әрекеттер Analyzing, Evaluating, Creating деңгейлеріне жатады. Алайда практикада олар көпнұсқалы немесе контексті тапсырмалар арқылы жүзеге асырылады, мұнда «пайымдау» дайын қорытындыны таңдауға, «негіздеу» дұрыс аргументті нұсқалар ішінен анықтауға сүйенеді.

Салыстырмалы талдау нәтижесінде оқыту мақсаттарында көзделген когнитивтік дағдылар мен тапсырмалар арқылы өлшенетін когнитивтік деңгейлер арасында айқын алшақтық бар екенін көрдік. Тест спецификациясында мәтінді талдау, ақпаратты синтездеу, аргументациялау, бағалау сияқты жоғары деңгейлі дағдылар көзделгенімен, тест құрылымы негізінен төменгі деңгейдегі ойлау дағдыларын (LOTS) тексеруге бағытталған.

Бұл теңгерімсіздік 1-кестеде анық көрінеді. Тапсырмалардың шамамен 70%-ы Remembering және Understanding деңгейлеріне тиесілі – бұл жабық форматтағы тапсырмалардың (40-тың 35-і) басым болуымен байланысты. Applying және Analyzing деңгейлері шектеулі көлемде қамтылған. Ал Evaluating және Creating деңгейлері тестте фрагментарлы, жанама түрде ғана көрініс табады.

1-кесте – Блум таксономиясы негізінде «Қазақ тілі» пәні бойынша тест тапсырмаларының когнитивтік деңгейлерінің бағалауыштық үлестірімі

Когнитивтік деңгей (Блум таксономиясы)	Негізгі тексерілетін әрекеттер	ҰБТ тестіндегі шамамен үлесі	Спецификациядағы күрделілік деңгейлеріне сәйкестігі
Remembering (Білу, қайта жаңғырту)	Терминдерді, ережелерді, фактілерді тану және есте сақтау	~50–60%	Негізінен базалық деңгей (А), ішінара В деңгейі
Understanding (Түсіну)	Интерпретациялау, қайта тұжырымдау, жіктеу, түсіндіру	~20–30%	Базалық (А) және орташа (В) деңгейлер
Applying (Қолдану)	Ережелерді, үлгілерді стандартты жағдаятта пайдалану	~15–20%	Көбіне орташа деңгей (В)
Analyzing (Талдау)	Байланыстарды анықтау, құрылымдық бөліктерді ажырату, аргументациялау	~5–10%	Орташа (В) және жоғары (С) деңгейдің жекелеген элементтері
Evaluating (Бағалау) Creating (Жасау)	Тексеру, сыни бағалау, көзқарасты негіздеу; Идеяларды генерациялау, жоспарлау, жаңа мәтін құрастыру	< 5%	Жоғары деңгейде (С) фрагментарлы түрде

Қазіргі тест моделі тілдік білімді репродуктивті түрде меңгеру деңгейін өлшеуге бағдарланған және тілді коммуникативтік әрі аналитикалық құрал ретінде қолдану қабілетін толыққанды бағаламайды. Анықталған когнитивтік дисбаланс тесттің конструкттық валидтілігінің шектеулі екенін көрсетеді және жоғары деңгейдегі ойлау дағдыларын өлшеу тұрғысынан тест мазмұнын қайта қарау қажеттігін айқындайды. Тест спецификациясындағы тапсырмалардың когнитивтік күрделілік деңгейлері Блум таксономиясының бес деңгейі бойынша салыстырылды. Оқу бағдарламасында мақсаттардың басым бөлігі «қолдану» (Applying), «талдау» (Analyzing) және «бағалау» (Evaluating) сияқты жоғары когнитивтік сатыларға бағытталған. Бұл оқушылардан тілдік білімдерін күрделі тапсырмаларды шешуге, сыни тұрғыдан талдау және бағалау дағдыларын пайдалануға талап етеді.

Психометриялық талдау талапкерлердің апробация бойынша алынған нәтижелілікке қатысты статистикалық деректер негізінде жүргізілді ($N = 257$).

Зерттеу аясында әрбір тапсырма элементінің өлшеу функциясын қаншалықты тиімді орындайтыны талданды. Талдау Winsteps бағдарламалық қамтамасыз етуінің (5.8.3.0 нұсқасы) көмегімен жүзеге асырылды. Қорытынды өңдеуге барлық қатысушылар мен тапсырмалар енгізілді, деректердің жетіспеушілігі немесе статистикалық шектен тыс мәндер анықталған жоқ.

Тесттің жинақталған психометриялық сипаттамаларына сәйкес (2-кесте), тапсырмалардың қиындығын бағалаудағы орташа стандарттық қате 0,32 логитті құрайды. Бұл көрсеткіш берілген іріктеме шегінде өлшеудің жеткілікті дәлдігін және алынған бағалардың сенімділігін айғақтайды.

2-кестеде ұсынылған деректер тесттің жинақталған психометриялық сипаттамаларының жалпы алғанда Раш моделінің талаптарына сәйкес келетінін көрсетеді. Нормаланған қалдықтардың (ZSTD) орташа мәндері Infit үшін 0,00, ал Outfit үшін 0,22 болып, рұқсат етілген -2,0 мен +2,0 аралығында орналасқан. ZSTD көрсеткіштері, әдетте, модельге сәйкестік индекстері (MNSQ) қабылданған шектерден асқан жағдайларда диагностикалық маңызға ие болатынын ескерсек, орташа ZSTD мәндерінің нөлге жуық болуы деректердің модельге жалпы сәйкестігін көрсетеді.

2-кесте - Тесттің жинақталған психометриялық сипаттамалары

Көрсеткіш	SCORE	COUNT	MEASURE	S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
MEAN	161,6	257,0	0,00	0,14	0,99	0,00	1,02	0,22
SEM	11,1	0,0	0,18	0,01	0,01	0,21	0,03	0,26
P.SD	69,1	0,0	1,10	0,03	0,08	1,32	0,18	1,60
S.SD	70,0	0,0	1,11	0,04	0,09	1,34	0,18	1,62
MAX	389,0	257,0	1,77	0,27	1,24	4,05	1,76	5,16
MIN	49,0	257,0	-2,89	0,09	0,85	-2,78	0,69	-3,04
Көрсеткіш	REAL RMSE	TRUE SD	Separation					
Item (Real)	0,14	1,09	7,58					
Item (Model)	0,14	1,09	7,66					

Нормаланған қалдықтар (ZSTD) Infit үшін 0,00, Outfit үшін 0,22 – рұқсат етілген -2,0 мен +2,0 аралығында. Бұл деректердің Раш моделіне жалпы сәйкестігін көрсетеді. Infit MNSQ (0,99) және Outfit MNSQ (1,02) рұқсат етілген 0,5–1,5 аралығында. Яғни, эмпирикалық деректер модельге жақсы үйлеседі.

Дегенмен, жеке тапсырмалар үшін максималды Infit ZSTD 4,05-ке, Outfit ZSTD 5,16-ға жетеді. Бұл кейбір тапсырмалардың модельден айтарлықтай ауытқуын көрсетеді. Оларды жеке сараптап, мазмұндық немесе құрылымдық түзету қажет. Бірақ бұл ауытқулар жалпы тест деңгейіндегі психометриялық сапаға елеулі әсер етпейді.

Бөлу коэффициенттері (Item Separation = 7,58–7,66) тесттің тапсырмаларды күрделілігі бойынша айқын ажырата алатынын көрсетеді. Нақты және модельдік орташа квадраттық қате (RMSE = 0,14) төмен, шынайы стандартты ауытқу 1,09 логит. Бұл өлшеу дәлдігінің жоғары екенін білдіреді.

3-кесте - Тапсырмалар параметрлерінің психометриялық сипаттамалары

№	Item	саны	Measure	S.E.	Infit MNS Q	Infit ZSTD	Outfit MNS Q	Outfit ZSTD	PTM EA Corr	Obs %	Exp %
1	MC-1	116	0,38	0,13	1,01	0,23	1,03	0,5	0,32	60,7	64,1
2	MC-2	105	0,58	0,13	1	0,1	1,02	0,34	0,32	68,1	65,3
3	MC-3	162	-0,44	0,14	0,99	-0,27	1,01	0,16	0,34	71,6	67,7
4	MC-4	185	-0,9	0,15	0,94	-0,88	0,88	-1,3	0,4	73,9	74
5	MC-5	223	-1,93	0,19	0,94	-0,4	0,99	0,01	0,32	86,8	86,8
6	MC-6	201	-1,27	0,16	0,94	-0,64	0,95	-0,4	0,37	80,2	79
7	MC-7	209	-1,49	0,17	0,9	-1,05	0,8	-1,46	0,43	82,5	81,8
8	MC-8	108	0,52	0,13	1,09	2,06	1,14	1,24	0,21	59,9	64,9
9	MC-9	80	1,05	0,14	1,09	1,45	1,12	1,38	0,33	68,9	71
10	MC-10	125	0,22	0,13	1,17	4,05	1,19	3,35	0,13	53,7	63,8
11	MC-11	173	-0,66	0,14	0,85	-2,78	0,78	-3,04	0,53	73,5	70,4
12	MC-12	124	0,24	0,13	1,02	0,6	1,05	0,91	0,3	64,2	63,8
13	MC-13	49	1,77	0,17	1,08	0,88	1,29	1,7	0,34	79	81,2
14	MC-14	133	0,08	0,13	1,04	0,98	1,07	1,36	0,28	62,3	63,9
15	MC-15	189	-0,99	0,15	0,93	-1,02	0,88	-1,17	0,41	78,6	75,2
16	MC-16	131	0,12	0,13	0,99	-0,19	0,99	-0,25	0,35	63,8	63,8
17	MC-17	211	-1,55	0,17	0,98	-0,17	0,87	-0,88	0,33	82,5	82,5
18	MC-18	79	1,07	0,14	0,98	-0,31	1,01	0,18	0,32	71,6	71,3
19	MC-19	185	-0,9	0,15	0,93	-0,97	0,88	-1,29	0,41	75,5	74
20	MC-20	165	-0,5	0,14	0,99	-0,09	0,97	-0,41	0,34	68,9	68,4
21	MC-21	163	-0,46	0,14	0,96	-0,84	1	0,06	0,37	72,8	67,9
22	MC-22	142	-0,08	0,13	1,04	1	1,07	1,32	0,28	61,9	64,5
23	MC-23	66	1,35	0,15	1,02	0,27	1,08	0,76	0,36	75,1	75,3
24	MC-24	64	1,39	0,15	0,93	-0,93	1	0,01	0,36	77,8	76
25	MC-25	97	0,72	0,14	0,91	-1,94	0,9	-1,55	0,43	75,1	66,6
26	CONT-26	228	-2,13	0,2	0,92	-0,5	0,87	-0,56	0,33	88,7	88,7
27	CONT-27	242	-2,89	0,27	0,92	-0,28	0,69	-1	0,31	94,2	94,2
28	CONT-28	187	-0,95	0,15	0,95	-0,77	0,93	-0,72	0,39	75,5	74,6
29	CONT-	75	1,15	0,14	1,08	1,31	1,27	1,74	0,16	74,7	72,5

	29										
30	CONT-30	51	1,71	0,16	0,97	-0,29	1,06	0,48	0,29	81,3	80,4
31	SOOTV-31	187	0,69	0,09	1,09	1,25	1,06	0,69	0,4	44	48,2
32	SOOTV-32	389	-0,98	0,1	1	0,04	1,18	1,56	0,42	66,5	60,8
33	SOOTV-33	124	1,25	0,1	1,24	1,76	5,16	2,16	0,36	51,4	59,7
34	SOOTV-34	288	-0,12	0,09	0,97	-0,34	0,97	-0,44	0,47	50,2	50,7
35	SOOTV-35	219	0,44	0,09	1,19	1,77	1,28	2,15	0,32	38,9	44,2
36	MC-36	237	0,32	0,09	0,95	-0,71	0,96	-0,51	0,5	44,7	45,9
37	MC-37	132	1,28	0,1	0,96	-0,41	0,95	-0,42	0,43	57,2	58,5
38	MC-38	216	0,53	0,1	0,9	-1,45	0,89	-1,48	0,53	54,9	51,7
39	MC-39	209	0,58	0,09	0,86	-2,07	0,84	-2,21	0,57	57,2	50,1
40	MC-40	193	0,8	0,1	0,96	-0,5	0,97	-0,42	0,45	56,4	55,1
Орташа мән (MEAN)		161,6	257	0	0,14	0,99	0	1,02	0,22	68,1	67,9
Стандартты ауытқу (P.SD)		69,1	0	1,1	0,03	0,08	1,32	0,18	1,6	12,7	11,7

3-кестеде көрсетілгендей, тапсырмалардың көпшілігі Раш моделінің талаптарына қанағаттанарлық сәйкес келеді. Күрделілік деңгейі кең ауқымды қамтиды (–2,89-дан +1,77 логитке дейін). Ең оңай тапсырма – CONT-27 (–2,89 логит), оны 94,2% қатысушы дұрыс орындаған. Ең қиыны – MC-13 (1,77 логит), дұрыс орындағандар 79%.

Модельге сәйкессіздік көрсеткен тапсырмалар да бар. SOOTV-33 (Outfit MNSQ = 1,76; ZSTD = 5,16) және MC-10 (Outfit MNSQ = 1,19; ZSTD = 3,35) ерекшеленеді. Бұл тапсырмалардың жауаптарында күтілмеген үлгілер кездесетінін білдіреді – мұндай тапсырмалардың құрылымын немесе жауап нұсқаларын қайта қарау қажет.

Тапсырма түрлері бойынша:

- MC тапсырмалары тұрақты психометриялық сипаттамаларға ие. Жоғары дұрыс жауап пайызы (75–87%) олардың салыстырмалы жеңілдігін көрсетеді.

- CONT тапсырмалары жоғары орындалу пайызымен (74–94%) және тұрақты Infit/Outfit мәндерімен ерекшеленеді.

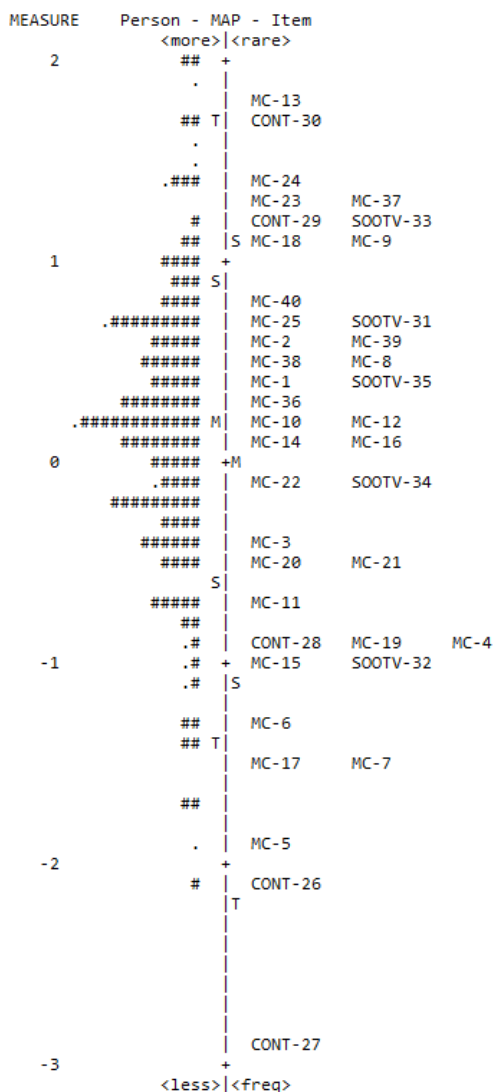
- SOOTV тапсырмаларының кейбірінде проблемалар бар. SOOTV-33 сәйкессіздігі жоғары, SOOTV-35 (38,9%) және SOOTV-31 (44%) төмен орындалу пайызы олардың қиындығын көрсетеді.

Жалпы тест сапасы:

- Item Separation = 7,58 – тапсырмалар күрделілігі бойынша жақсы ажыратылған.
- Reliability = 0,81 – ішкі сәйкестік жоғары.
- RMSE = 0,14 – өлшеу дәлдігі жоғары.

Айнымалылар картасы (Сурет 1) бойынша респонденттердің қабілеттері –2,5-тен +2,0 логит аралығында, негізгі массасы 0,0-ден +1,0 логит аралығында шоғырланған. Ең көп адам шоғыры +0,5 логит шамасында. Орташа қабілет 0,0 логит – бұл қатысушылардың орташа дайындық деңгейінің басым екенін және тапсырмалардың орташа күрделілікке сәйкес келетінін көрсетеді.

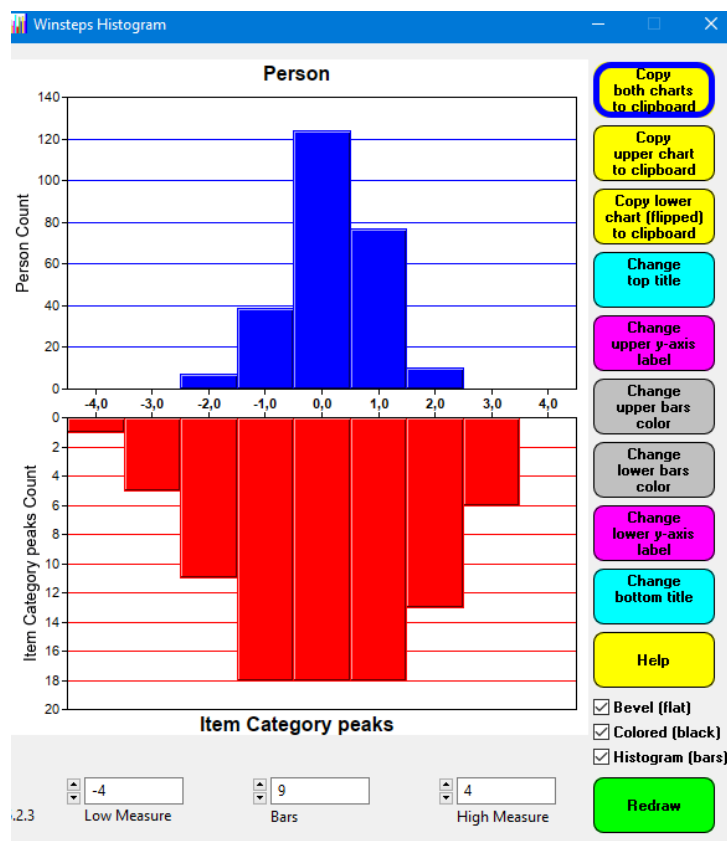
INPUT: 257 Person 40 Item REPORTED: 257 Person 40 Item 90 CATS WINSTEPS 5.8.3.0



Сурет 1 - Айнымалылар картасы: Респонденттер – Тапсырмалар

Тапсырмалар қиындығы $-2,89$ -дан $+1,77$ логитке дейінгі аралықты қамтиды. Ең сәйкес диапазон $-1,5$ -тен $+1,5$ логит аралығы. Бірақ $+1,5$ логиттен жоғары аймақта тапсырмалар саны азаяды (MC-13, MC-24, CONT-30, MC-23, MC-37). Бұл жоғары қабілетті қатысушылар үшін өлшеу дәлдігін төмендетеді. Мұны түзету үшін «Талдау» және «Бағалау» деңгейлеріне сәйкес қосымша тапсырмалар әзірлеу керек.

Жалпы, тест қабілеттер континуумын жақсы қамтиды, төмен және орта деңгейдегі қатысушыларды сенімді бағалауға мүмкіндік береді. Негізгі тест нұсқасын жасауда тапсырмаларды қиындық деңгейіне сәйкес қалыпқа келтіру принципі қолданылды.



Сурет 2 - Вraith диаграммасы

Гистограмма көрсеткендей, қатысушылар қабілеті мен тапсырмалар қиындығы арасындағы сәйкестік негізінен қанағаттанарлық. Негізгі қатысушылар (0,0–1,0 логит) үшін тапсырмалар жеткілікті. Бірақ жоғары қабілеттілер (+1,5 логиттен жоғары) үшін тапсырмалар саны аз – бұл «төбе» эффектісіне әкеледі. Төмен қабілеттілер (–1,5 логиттен төмен) үшін де өте оңай тапсырмалар саны шектеулі

Талқылау. Зерттеу барысында 2026 жылдан бастап қолданылатын «Қазақ тілі» пәні бойынша ҰБТ спецификациясының мазмұны мен құрылымы талданды. Нәтижелер тест мазмұнының мемлекеттік жалпыға міндетті білім беру стандарты талаптарына жалпы сәйкес келетінін көрсетті. Спецификация тілдік нормалардың негізгі компоненттерін (фонетика, лексика, морфология, синтаксис, пунктуация) және оқылым дағдыларын қамтып, мазмұндық тұтастықты қамтамасыз етеді.

Тапсырмалар қиындығының үш деңгейде (А – 50%, В – 30%, С – 20%) ұйымдастырылуы қатысушыларды саралап бағалауға мүмкіндік береді. Алайда когнитивтік сұраныс тұрғысынан деңгейлер арасындағы айырмашылық жеткілікті дәрежеде айқын емес. Жоғары деңгейге (С) жатқызылған тапсырмалар көбінесе мазмұндық немесе форматтық күрделілікке (мысалы, бірнеше дұрыс жауапты таңдау, сәйкестікті анықтауға арналған тест тапсырмасы) негізделеді. Зерттеу нәтижелері бұл тапсырмаларда жоғары ретті ойлау дағдыларының (HOTS), атап айтқанда синтездеу және пайымдаудың шектеулі түрде көрініс табатынын көрсетті. Соның салдарынан тест Блум таксономиясының төменгі сатыларын – білу, түсіну және қолдануды (Remembering, Understanding, Applying) басым өлшейді, бұл оның конструкттық валидтілігін шектейді.

Психометриялық тұрғыдан тесттің сенімділік көрсеткіші (Reliability = 0,81) жақсы деңгейде

және нәтижелердің тұрақтылығын қамтамасыз етеді. Алайда сенімділік пен валидтілік ұғымдарының мазмұны әртүрлі: сенімділік өлшеудің тұрақтылығын сипаттаса, валидтілік тесттің нақты қандай конструктты өлшейтінін анықтайды. Жүргізілген талдау тесттің негізінен реттеуші білімді (ережелерді білу және қолдану) сенімді бағалайтынын, бірақ коммуникативтік және аналитикалық құзыреттіліктерді толыққанды қамтымайтынын көрсетті.

Раш моделінің көрсеткіштері (Item Separation = 7,58) тапсырмалардың күрделілік деңгейі бойынша жақсы ажыратылғанын дәлелдейді. Дегенмен, бұл ажырату көбіне төмен және орта деңгейдегі когнитивтік тапсырмалар шеңберінде жүзеге асады. Айнымалылар картасы (Wright Map) бұл қорытындыны растайды: тапсырмалар күрделілік континуумында кең таралғанымен, қатысушылардың басым бөлігі (шамамен +0,5 логит деңгейінде) негізінен орташа және салыстырмалы жеңіл тапсырмалар арқылы өлшенеді. +1,5 логиттен жоғары аймақта тапсырмалар санының шектеулі болуы жоғары қабілетті талапкерлерді дәл саралау мүмкіндігін төмендетеді.

Аталған нәтижелерді OECD-нің «Future of Education and Skills 2030» [16] тұжырымдамасы аясында қарастырсақ, белгілі бір дисбаланс байқалады. OECD құжаты XXI ғасыр оқушылары үшін тек пәндік білімді ғана емес, сонымен қатар сыни ойлау, проблеманы шешу, рефлексия, жауапкершілік, құндылықтарға негізделген шешім қабылдау сияқты кешенді құзыреттіліктерді дамыту қажеттігін айқындайды. Бұл парадигма білімді репродукциялаудан гөрі оны жаңа жағдайларда трансферлеу мен интеграциялауға басымдық береді.

Зерттеу шектеулері. Алынған нәтижелер 257 қатысушының деректеріне негізделген бір тест нұсқасын талдаумен шектеледі. Раш моделі бойынша стандартты қателік (0,32 логит) параметрлердің бағалану сенімділігін көрсеткенімен, іріктемені кеңейту мен тесттің басқа нұсқаларын салыстыру тұжырымдардың жалпылану дәрежесін арттырады. Дегенмен, бұл шектеулер тест құрылымындағы когнитивтік дисбаланс пен оны жаңғырту бағыттарын айқындауға кедергі келтірмейді.

Осы жұмысты дайындау барысында авторлар мәтінді тілдік редакциялау және құрылымын жетілдіру үшін ChatGPT құралын пайдаланды. Аталған құралды қолданғаннан кейін мазмұнды қайта қарап, қажетті түзетулер енгізілді. Авторлар жарияланымның мазмұны үшін толық жауапкершілікті өз мойнына алады.

Әдебиеттер тізімі

1. Messick, S. Validity // Educational Measurement / ed. by R. L. Linn. - 3rd ed. - New York: American Council on Education ; Macmillan, 1989. - P. 13-103.
2. Mislevy, R. J. Evidence and Inference in Educational Assessment. - Princeton, NJ : Educational Testing Service, 1994. - 45 p.
3. Strengthening national examinations in Kazakhstan to achieve national goals / OECD // OECD Education Policy Perspectives. - 2020. - No. 24. - Paris : OECD Publishing. - <https://doi.org/10.1787/0bf8662b-en>
4. Ұлттық бірыңғай тестілеуді өткізу қағидалары : Қазақстан Республикасы Білім және ғылым министрінің 2017 жылғы 2 мамырдағы № 204 бұйрығы. - <https://adilet.zan.kz/kaz/docs/V1700015173>
5. Алтыбаева, Ш. Б. Валидация содержания теста по информатике в рамках ЕНТ: экспертная оценка и психометрический анализ / Ш. Б. Алтыбаева, И. Сагиндикова, Г. Жабаева, Ш. Шакралиева // Қ. Алтынсарин атындағы Ұлттық білім академиясының «Білім-Образование» ғылыми-педагогикалық журналы. - 2025. - Т. 114, № 3. - С. 68-82. - <https://doi.org/10.59941/2960-0642-2025-3-68-82>
6. Мемлекеттік жалпыға міндетті білім беру стандарттары : Қазақстан Республикасы Оқу-ағарту министрінің 2022 жылғы 3 тамығындағы № 348 бұйрығы. - <https://adilet.zan.kz/kaz/docs/V2200029031>
7. «Қазақ тілі» пәні бойынша Ұлттық бірыңғай тестілеуге арналған тест спецификациясы / Ұлттық тестілеу орталығы. - 2025. - <https://testcenter.kz/wp->

content/uploads/2025/10/01_%D2%9A%D0%B0%D0%B7%D0%B0%D2%9B-%D1%82%D1%96%D0%BB%D1%96_%D0%BA%D0%B0%D0%B7.pdf

8. Question & Test Interoperability (QTI) Specification Documents / 1EdTech Consortium. - 2024. - <https://www.1edtech.org/standards/qti>
9. Yin, R. K. Case Study Research and Applications: Design and Methods. - 6th ed. - Los Angeles : SAGE Publications, 2018. - 352 p.
10. Creswell, J. W. Qualitative Inquiry and Research Design: Choosing Among Five Approaches / J. W. Creswell, C. N. Poth. - 4th ed. - Los Angeles : SAGE Publications, 2018. - 488p.
11. Bloom, B. S. Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain / B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, D. R. Krathwohl. - New York : Longmans, Green, 1956. - 207 p.
12. Krathwohl, D. R. A Revision of Bloom's Taxonomy: An Overview // Theory Into Practice. - 2002. - Vol. 41, No. 4. - P. 212-218. - https://doi.org/10.1207/s15430421tip4104_2
13. Linacre, J. M. Winsteps® Rasch Measurement Computer Program. - Beaverton, Oregon : Winsteps.com, 2024. - <https://www.winsteps.com>
14. Cronbach, L. J. Coefficient Alpha and the Internal Structure of Tests // Psychometrika. - 1951. - Vol. 16, No. 3. - P. 297-334. - <https://doi.org/10.1007/BF02310555>
15. Bond, T. G. Applying the Rasch Model: Fundamental Measurement in the Human Sciences / T. G. Bond, C. M. Fox. - 3rd ed. - New York : Routledge, 2015. - 406p. - <https://doi.org/10.4324/9781315814698>
16. Future of Education and Skills 2030 / OECD. - 2022. - <https://www.oecd.org/en/about/projects/future-of-education-and-skills-2030.htm>

Ш.Б. Алтыбаева, М.Ш. Тасбулатова, Г.Е. Нұрмұханова, Р.Б. Аубакирова

ВАЛИДНОСТЬ ИЗМЕРИТЕЛЬНОЙ МОДЕЛИ ПО КАЗАХСКОМУ ЯЗЫКУ В ЕНТ: ВЕРИФИКАЦИЯ КОГНИТИВНЫХ УРОВНЕЙ И СТАНДАРТОВ

Данное исследование направлено на анализ валидности измерительной модели, применяемой в рамках Единого национального тестирования (ЕНТ) по профильному предмету «Казахский язык». Актуальность исследования определяется стратегической ролью вступительного экзамена в высшие учебные заведения в управлении качеством образования в стране. Целью исследования является оценка соответствия содержания теста учебной программе, баланса когнитивных уровней, а также его психометрического качества. В качестве методологической основы использованы содержательное сопоставление и когнитивная верификация на основе таксономии Блума, а также психометрический анализ, основанный на модели Раша ($N=257$). Результаты показали, что тест в целом соответствует учебной программе с содержательной точки зрения и характеризуется высоким уровнем внутренней надежности ($Reliability = 0,81$; $Item Separation = 7,58$). При этом 70% заданий относятся к навыкам низшего порядка (знание, понимание), тогда как задания, направленные на развитие навыков высокого уровня – анализа, оценки и синтеза – составляют менее 10%. Диапазон сложности заданий варьируется от -2,89 до +1,77 логита, однако карта переменных выявила дефицит заданий для абитуриентов с высоким уровнем подготовки (выше +1,5 логита). В заключение, для повышения конструктивной валидности теста рекомендуется увеличить долю заданий высокого когнитивного уровня, внедрить интерактивные форматы на основе стандарта QTI, а также поэтапно перейти к технологии компьютерного адаптивного тестирования.

Ключевые слова: ЕНТ, казахский язык, валидность, модель Раша, таксономия Блума, стандарт QTI.

Sh.B. Altybayeva, M.Sh. Tasbulatova, G.E. Nurmukhanova, R.B. Aubakirova

VALIDITY OF THE KAZAKH LANGUAGE MEASUREMENT MODEL IN THE UNIFIED NATIONAL TESTING: VERIFICATION OF COGNITIVE LEVELS AND STANDARDS

This study analyzes the validity of the measurement model for the «Kazakh Language» subject within the Unified National Testing (UNT) and explores its potential for modernization based on contemporary technological standards. The relevance of the research is determined by the strategic role of high-stakes exams in academic selection and educational quality management. The aim of the study is to assess the alignment of the test content with the curriculum, the balance of cognitive levels, and its psychometric quality. A mixed-method approach was employed as a methodological framework: content audit and cognitive verification using Bloom's Taxonomy, along with psychometric analysis based on the Rasch model (N=257). The results indicate that the test generally aligns with the curriculum content and demonstrates high internal reliability (Reliability = 0,81; Item Separation = 7,58). 70% of the items target Lower Order Thinking Skills (LOTS: remembering, understanding), while Higher Order Thinking Skills – analyzing, evaluating, and creating (HOTS) – account for less than 10%. Item difficulty covers a wide range (from -2,89 to +1,77 logits); however, the variable map clearly shows a shortage of items for high-ability test-takers (above +1,5 logits). In conclusion, to enhance the test's construct validity, it is recommended to increase the proportion of high-cognitive-level items, introduce interactive formats based on the QTI standard, and gradually transition to computerized adaptive testing technology.

Keywords: UNT, Kazakh language, validity, Rasch model, Bloom's Taxonomy, QTI standard.

References

1. Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). American Council on Education; Macmillan.
2. Mislevy, R. J. (1994). *Evidence and inference in educational assessment*. Educational Testing Service.
3. OECD. (2020). *Strengthening national examinations in Kazakhstan to achieve national goals*. OECD Education Policy Perspectives, (24). Paris: OECD Publishing. <https://doi.org/10.1787/0bf8662b-en>
4. Ultyyq biryngai testileudi otkizu qagidalary: Kazakhstan Respublikasy Bilim jane gylym ministrinin 2017 jylgy 2 mamyrday No. 204 buirygy [Rules for conducting the Unified National Testing: Order of the Minister of Education and Science of the Republic of Kazakhstan dated May 2, 2017, No. 204]. (2017). <https://adilet.zan.kz/kaz/docs/V1700015173>
5. Altybayeva, S. B., Sagindikov, I., Zhabayeva, G., Shakraliyeva, S. (2025). Validatsiya soderzhaniya testa po informatike v ramkakh ENT: ekspertnaya otsenka i psikhometricheskiy analiz [Validation of the content of the ICT test within the framework of the UNT: Expert assessment and psychometric analysis]. *Q. Altynsarin atyndagy Ultyyq bilim akademiynynyn «Bilim-Obrazovanie» gylymi-pedagogikalyq jurnaly* [Scientific-Pedagogical Journal «Bilim-Obrazovanie» of the National Academy of Education named after Y. Altynsarin], 114(3), 68-82. <https://doi.org/10.59941/2960-0642-2025-3-68-82>
6. Memlekettyk jalpyga mindetty bilim beru standarttary: Kazakhstan Respublikasy Oqu-agartu ministrinin 2022 jylgy 3 tamyzyndagy No. 348 buirygy [State compulsory standards of education: Order of the Minister of Education of the Republic of Kazakhstan dated August 3, 2022, No. 348]. (2022). <https://adilet.zan.kz/kaz/docs/V2200029031>
7. «Kazakh tili» pani boiynsha Ultyyq biryngai testileuge arnalgan test spetsifikatsiyasy [Test specification for the Unified National Testing in the subject «Kazakh Language»]. (2025). Ultyyq testileu ortalygy. https://testcenter.kz/wp-content/uploads/2025/10/01_%D2%9A%D0%B0%D0%B7%D0%B0%D2%9B-%D1%82%D1%96%D0%BB%D1%96_%D0%BA%D0%B0%D0%B7.pdf
8. 1EdTech Consortium. (2024). *Question & test interoperability (QTI) specification documents*. <https://www.1edtech.org/standards/qti>
9. Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE Publications.

10. Creswell, J. W., Poth, C. N. (2018). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). SAGE Publications.
11. Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. Longmans, Green.
12. Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, 41(4), 212-218. https://doi.org/10.1207/s15430421tip4104_2
13. Linacre, J. M. (2024). *Winsteps® Rasch measurement computer program*. <https://www.winsteps.com>
14. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334. <https://doi.org/10.1007/BF02310555>
15. Bond, T. G., Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge. <https://doi.org/10.4324/9781315814698>
16. OECD. (2022). *Future of education and skills 2030*. <https://www.oecd.org/en/about/projects/future-of-education-and-skills-2030.htm>

Авторлар туралы мәлімет:

Алтыбаева Шугыла Болатовна - ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК Орта және жоғары білім берудегі тест тапсырмаларын қалыптастыру басқармасының басшысы, Астана, Қазақстан, e-mail: shugyla.altybayeva@gmail.com

Тасбулатова Махаббат Шалабаевна (автор-корреспондент) - ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, бас сарапшы, Астана, Қазақстан, e-mail: tasbulatova.makhabbat@mail.ru

Нұрмұханова Гүлнар Ерғалиқызы - ҚР ҒЖБМ Ұлттық тестілеу орталығы» ШЖҚ РМК, бас сарапшы, Астана, Қазақстан, e-mail: Gulnarae@bk.ru

Аубакирова Риза Бекмұхамедовна - Астана қаласы әкімдігінің «№82 «Дарын» мамандандырылған мектеп-лицейі» КММ, қазақ тілі мен әдебиеті пәні мұғалімі, Астана, Қазақстан, e-mail: riza211989@mail.ru

Сведения об авторах:

Алтыбаева Шугыла Болатовна - руководитель Управления по формированию тестовых заданий для высшего и среднего образования РГП на ПХВ «Национальный центр тестирования» МНВО РК, Астана, Казахстан, e-mail: shugyla.altybayeva@gmail.com

Тасбулатова Махаббат Шалабаевна (автор-корреспондент) - РГП на ПХВ «Национальный центр тестирования» МНВО РК, заведующий лабораторией, Астана, Казахстан, e-mail: tasbulatova.makhabbat@mail.ru

Нұрмұханова Гүлнар Ерғалиқызы - РГП на ПХВ «Национальный центр тестирования» МНВО РК, главный эксперт, Астана, Казахстан, e-mail: Gulnarae@bk.ru

Аубакирова Риза Бекмұхамедовна - учитель казахского языка и литературы КГУ «Специализированный школа-лицей №82 «Дарын» акимата города Астаны, Астана, Казахстан, e-mail: riza211989@mail.ru

Information about authors:

Altybayeva Shugyla Bolatovna - Head of the Office for the Formation of Test Items for Higher and Secondary Education «National Testing Center» Ministry of Science and Higher Education of the Republic of Kazakhstan, Astana, Kazakhstan, e-mail: shugyla.altybayeva@gmail.com

Tasbulatova Makhabbat Shalabaevna (corresponding author) - Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Chief expert, Astana, Kazakhstan, e-mail: tasbulatova.makhabbat@mail.ru

Nurmukhanova Gulnar Ergalikyzy - Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Chief expert, Astana, Kazakhstan, e-mail: Gulnarae@bk.ru

Aubakirova Riza Bekmukhamedovna - teacher of Kazakh language and literature KSU «Specialized school- lueum №82 «Daryn» of the akimat of Astana city, Astana, Kazakhstan, e-mail: riza211989@mail.ru

A. Serikbayeva^{1*}, D. Akizhanova², L. Shinetova³, B. Abdrasilov⁴

¹Al-Farabi Kazakh National University, Almaty, Republic of Kazakhstan

²Eurasian National University, Astana, Republic of Kazakhstan

^{3,4}National Testing Center, Astana, Republic of Kazakhstan

*email: assiyaserikbayeva237@gmail.com

¹ORCID 0009-0002-7516-482X, ²ORCID 0000-0002-1300-0339

³ORCID 0000-0003-4280-7999, ⁴ORCID 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 10th and 11th 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 whether 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.

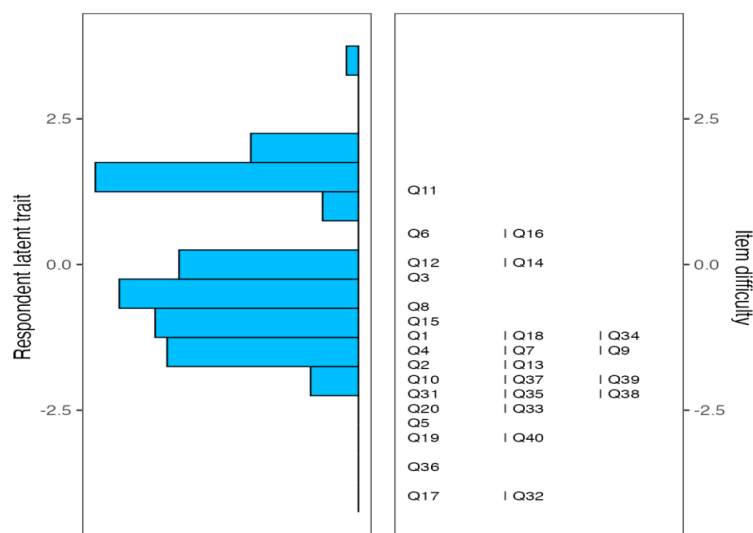


Figure 1 - Wright Map of Person ability and Item difficulty scale

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.

Furthermore, 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/educ.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 (СТТ) 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 $\alpha = .879$; $\omega = .898$). 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, конструктивную 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 (СТТ) 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; $\omega = .898$). 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, конструктивная 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, Head of the Laboratory of Scientific Research and 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 Center of 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

**К.Т. Туралбаева*¹, Г.С. Джарасова², И.Ж. Жуматова³, Г.Т. Сагнаева⁴,
Б.К. Асеева⁵**

^{1,3,4,5}ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, Астана қ., Қазақстан Республикасы

²ҚР ҒЖБМ «Жоғары және жоғары оқу орнынан кейінгі білім комитеті», Астана қ., Қазақстан Республикасы

*e-mail: kengebahit@mail.ru

¹ORCID 0009-0004-6640-0522, ²ORCID 0000-0001-9736-0499, ³ORCID 0009-0001-5382-4519,

⁴ORCID 0009-0009-7651-508X, ⁵ORCID 0009-0003-4857-4679

МАТЕМАТИКА ПӘНІ БОЙЫНША ТЕСТТІҢ ПСИХОМЕТРИЯЛЫҚ ТАЛДАУЫ: РАШ МОДЕЛІ НЕГІЗІНДЕ БАҒАЛАУ

Мақалада Ұлттық бірыңғай тестілеу (ҰБТ) шеңберінде математика пәні бойынша апробациядан өткен тест тапсырмаларына жүргізілген психометриялық талдау нәтижелері ұсынылады. Зерттеудің негізгі мақсаты – тесттің өлшеу сипаттамаларын бағалау, жеке тапсырмалардың жұмыс істеу ерекшеліктерін анықтау және тест нәтижелерінің сенімділігі мен валидтілігін қамтамасыз ету. Талдау барысында бір параметрлі Раш моделі (Rasch model) қолданылды. Бұл тәсіл тапсырмалардың күрделілігі мен қатысушылардың дайындық деңгейін бір шкалада қарастыруға мүмкіндік береді, сонымен қатар тесттің сапасын объективті және дәл бағалауға жағдай жасайды. Раш моделі тесттің әрбір элементінің қатысушылар арасындағы айырмашылықты қаншалықты дұрыс көрсететінін анықтауға көмектеседі. Зерттеу деректері ретінде математика пәні бойынша бір тест нұсқасының апробациясына қатысқан 326 тестіленушінің нәтижелері пайдаланылды. Мәліметтер WINSTEPS бағдарламасында (5.8.3.0 нұсқасы) өңделіп, тапсырмалардың күрделілігі, қатысушылардың қабілет деңгейі, модельге сәйкестік көрсеткіштері (Infit/Outfit MNSQ, ZSTD), дискриминативтілігі және тесттің сенімділігі анықталды. Нәтижелер тесттің жалпы алғанда сенімді екенін көрсетті. Сонымен қатар, жекелеген тапсырмалар бойынша мазмұндық және құрылымдық түзетулер қажет екендігі анықталды. Мұндай жетілдірулер тест нәтижелерінің дәлдігін арттыруға, әртүрлі дайындық деңгейіндегі қатысушыларды неғұрлым дұрыс саралауға және тестті қолдану тиімділігін жоғарылатуға мүмкіндік береді.

Түйін сөздер: ҰБТ, Раш моделі, психометриялық талдау, тапсырмалардың қиындығы, тесттің сенімділігі, валидтілік.

Кіріспе

ҰБТ тапсырмалары білім алушылардың тек пәндік білімін ғана емес, сонымен қатар олардың функционалдық сауаттылығын бағалауға бағытталған. Осыған байланысты тест мазмұны мен құрылымына қойылатын талаптар жыл сайын нақтыланып, жетілдіріліп отырады.

Қазіргі уақытта тестілеу жүйелері педагогикалық өлшеулер теориясына сүйенеді және білім сапасын объективті бағалаудың маңызды құралдарының бірі ретінде қарастырылады [1]. ҰБТ форматында қолданылатын тапсырмалар әртүрлі формада ұсынылады: бір дұрыс жауабы бар сұрақтар, бірнеше дұрыс жауабы бар тапсырмалар, контекстік тапсырмалар және сәйкестікті анықтауға арналған тапсырмалар. Тапсырмалар базасын әзірлеу көпсатылы үдеріс болып табылады. Ол тапсырмаларды құрастырудан бастап, сараптамадан өткізу, түзету және апробациялау кезеңдерін қамтиды. Алайда тек сарапшылар пікірімен шектелу жеткіліксіз. Тест сапасын жан-жақты бағалау үшін нақты деректерге сүйенген талдау қажет [2].

Соңғы жылдары классикалық тест теориясымен қатар заманауи психометриялық тәсілдер, соның ішінде Раш моделі кеңінен қолданыла бастады. Бұл модель тестіленушінің қабілеті мен тапсырманың қиындығы арасындағы байланысты бағалауға мүмкіндік береді және тест нәтижелерін интерпретациялаудың дәлдігін арттырады [2, 3].

Зерттеу материалдары мен әдістері

Зерттеу барысында Қазақстанның әртүрлі өңірлерінен қатысқан 326 тестіленушінің нәтижелері қарастырылды. Талдау математика пәні бойынша 40 тапсырмадан тұратын бір тест нұсқасының деректері негізінде жүргізілді.

Бір шкала бойынша сұрақтың қиындығы мен қатысушылардың шеберлік деңгейін бағалауға мүмкіндік беретін негізгі құрал ретінде Rasch моделі пайдаланылды.

Зерттеу объектісі мен контингенті:

- Республиканың әр түрлі өңірлерінен таңдалған 326 қатысушы.

Қолданылған құралдар мен материалдар:

- Математика пәні бойынша 40 тапсырмадан тұратын тест;

- Деректерді өңдеу үшін WINSTEPS бағдарламасы (5.8.3.0 нұсқасы) [3].

Статистикалық және психометриялық әдістер:

- Бір параметрлі Раш моделі (Rasch model) және оның политомды нұсқасы (тапсырмалар 0, 1, 2 балл бойынша бағаланған жағдайда) [4, 5];

- Тест элементтерінің күрделілігін, қатысушылар қабілетін бағалау;

- Модельге сәйкестік индекстері (Infit/Outfit MNSQ және ZSTD);

- Тапсырмалардың дискриминативтілігі (Pt-Measure Corr.);

- Тесттің сенімділігі (Reliability) және жіктелуі (Separation) [6, 7];

Деректерді өңдеу:

- Психометриялық деректерді компьютерлік бағдарлама арқылы талдау;

- Тесттің әр элементінің қалай жұмыс істейтінін бағалау, коррекция қажет тапсырмаларды анықтау [8, 9].

Нәтижелер және талқылау

Контингент пен шкаланың жалпы сипаттамасы. Қатысушылардың қабілет деңгейлерін және тапсырмалардың қиындығын таралуын талдау (1 кестесіндегі «Person-Item Map» картасын талдау) тесттің тепе-теңдігін анықтауға алғашқы маңызды қорытындылар жасауға мүмкіндік береді.

- Тапсырмалардың қиындық диапазоны (Item Measure): Тапсырмалар қиындықтың кең ауқымын қамтиды – (-1,86) мен (+1,71) логит аралығында. Бұл жақсы көрсеткіш, яғни тест өте жеңіл және өте қиын тапсырмаларды қамтиды.

- Қатысушылардың қабілет деңгейінің диапазоны (Person Measure): Сынақтан өткендер (-2,33) мен (+1,84) логит аралығында таралған, бұл топтың дайындық деңгейі бойынша әртүрлілігін айқындайды.

- Шкаланы ортаға келтіру: Қатысушылардың қабілет деңгейінің орташа көрсеткіші (Mean Person) (-0,36) логитті құрайды, ал тапсырмалардың орташа қиындығы (Mean Item) бастапқыда 0,00 деп белгіленген.

Осыған байланысты қатысушылардың таралуы сәл төмен қабілеттер жағына ығысқан. Бұл көрсеткіш берілген контингент үшін тесттің орташа деңгейде қиын болғанын білдіреді. Топ, жалпы алғанда, шкала бағытындағы «орташа гипотетикалық» студентке қарағанда тестті біршама нашар орындады.

- Шкала бойынша қамту: Тапсырмалар жалпы алғанда қатысушылардың көпшілігінің қабілет диапазонын жақсы қамтиды. Дегенмен, картаның визуалды талдауы жоғары қабілетті деңгейлерде (+1,5 логиттен жоғары) тапсырмалардың жеткіліксіз екенін көрсетеді. Ең мықты оқушылар үшін тест олардың нәтижелерін дәл ажыратуға жеткілікті мүмкіндік бермей тұр.

Сенімділік пен жіктелу бойынша талдау. Тесттің сенімділігін бағалау педагогикалық өлшеулердің негізгі талаптарының бірі болып табылады. Сенімділік көрсеткіштері өлшеу нәтижелерінің тұрақтылығын және тесттің әртүрлі топтарда қолданылған жағдайда ұқсас нәтижелер беру қабілетін сипаттайды [10].

Кронбах альфа коэффициенті тесттің ішкі үйлесімділігін бағалау үшін кеңінен қолданылатын статистикалық көрсеткіштердің бірі болып табылады [11].

1 кестесінен алынған тест сапасының негізгі көрсеткіштері төменде 1-кестеде берілген:

1-кесте - Тест сапасының негізгі көрсеткіштері

Көрсеткіш	Мәні	Интерпретация
Қатысушылардың сенімділігі (Person Reliability)	0,83 (Model)	Жақсы. Тест пәндерді білім деңгейі бойынша сенімді түрде ажыратады. Мұнда 0,8-ден жоғары мән топтық және жеке шешімдер қабылдау үшін қолайлы.
Тапсырмалардың сенімділігі (Item Reliability)	0,98 (Model)	Өте жоғары. Тапсырманың күрделілігінің иерархиясы өте тұрақты. Егер бұл тест ұқсас үлгіде қайталанса, тапсырманың күрделілігінің реті сақталады.
Қатысушылардың жіктелуі (Person Separation)	2,19	Орташа. Дайындық деңгейіне қарай шамамен 3-4 статистикалық тұрғыда ерекшеленетін қатысушылар тобын анықтауға мүмкіндік береді.
Тапсырмалардың жіктелуі (Item Separation)	7,74	Жоғары. Тапсырмалар қиындығы бойынша өте жақсы ерекшеленеді. Көрсеткіш тапсырмалардың қиындықтардың кең ауқымын қамтитынын растайды.
Кронбах коэффициенті (Cronbach's Alpha)	0,83	Жақсы ішкі үйлесімділік. Қатысушылардың сенімділік көрсеткішімен корреляцияланады.
Орташа балл (Raw Score)	27 / 50	54%. Бұл қорытынды тесттің таңдалған қатысушылар үшін біршама қиын болғанын растайды (нәтиже жалпы баллдың ортасынан асады)

Тапсырмалар жіктелуінің жоғары көрсеткіші (7,74) және өте жоғары сенімділік көрсеткіші (0,98) бұл құралдың қиындық шкаласы бойынша өте жақсы реттелгенін білдіреді, бұл тесттің күшті жақтарының бірі болып саналады.

Модельдің сәйкестігін тексеру (Infit/Outfit). Модельдің сәйкестігі Infit және Outfit статистикалары (MNSQ және ZSTD) арқылы тексерілді. Сапалы тестте MNSQ орташа мәндері 1,0-ге, ал ZSTD мәндері 0-ге жақын болуы тиіс.

- *Қатысушылар (Persons):* Infit/Outfit MNSQ орташа мәндері 1,01 және 1,02, ал ZSTD мәндері нөлге жақын. Бұл қатысушылардың модельге сәйкес әрекет ететінін көрсетеді. Дегенмен, Outfit MNSQ максималды мәндерінің 2,52-ге жетуі кейбір қатысушылардың болжанбаған жауап бергендігін білдіреді (мысалы, қабілетті оқушылар жеңіл сұрақтарда қателесуі мүмкін немесе әлсіз оқушылар қиын сұрақтарды кездейсоқ дұрыс жауабымен тапсыруы мүмкін).

- *Тапсырмалар (Items):* Орташа көрсеткіштер де идеалға жақын: Infit MNSQ = 1,00, Outfit MNSQ = 1,02. Орташа ZSTD мәндері (-0,23 және -0,09) «underfit» тенденциясының аздап бар екенін көрсетеді (яғни, шу көлемі жоғары, болжамдылығы төмен), бұл нақты деректер үшін жалпы қабылданады. Дегенмен, кейбір тапсырмалардың ZSTD мәндерінің 3,85-ке жетуі олардың жан-жақты талдануын қажет етеді.

Тапсырмалардың жұмыс жасау қабілетін талдау. Тапсырмалардың сапасын бағалау кезінде олардың қиындық деңгейі мен дискриминациялық қабілеті маңызды рөл атқарады. Бұл көрсеткіштер тест тапсырмаларының жоғары және төмен дайындық деңгейіндегі қатысушыларды ажырату қабілетін анықтайды [12].

Тапсырмалардың статистикалық көрсеткіштерін талдау классикалық тест теориясы мен

заманауи психометриялық әдістердің үйлесімді қолданылуын талап етеді [13].

Талдаудың ең маңызды бөлігі – қорытынды бағалауды бұрмалайтын, күткеніміздей жұмыс жасамайтын тапсырмаларды анықтау. Тапсырмалар проблемалардың маңызды деңгейіне қарай Infit/Outfit, ZSTD мәндері және нүктелік-бисериалдық корреляция (Pt-measure) көрсеткіштері негізінде топтарға бөлінді. Тапсырмалар проблемалық деңгейіне қарай 2,3-кестелерде ұсынылады.

1) Тапсырмаларды проблемалық дәрежесіне қарай жіктеу

А ТОБЫ: КРИТИКАЛЫҚ МӘСЕЛЕЛЕР (Тез арада тексеруді немесе жоюды қажет етеді).

Бұл тапсырмалар тесттің жалпы тестімен бірдей айнаымалыны өлшемейді немесе тапсырма шартында екіойлылық бар болуы мүмкін.

2-кесте - Проблемалық тапсырмалардың А-тобы

Тапсырма	Мәселе	Ұсыныс
MC-9	Outfit MNSQ = 1,53 (жоғары шу), ZSTD = 3,85, Pt-measure corr. = -0,01 (теріс!). Тапсырма «босқа» немесе зиянды жұмыс істейді: қабілеті жоғары оқушылар оны әлсіздерге қарағанда нашар орындайды.	Тапсырманы жан-жақты тексеру. Критикалық ақау.
MC-24	Infit ZSTD = -4,08, Outfit ZSTD = -4,11 (күшті overfit). Өте болжамды тапсырма. Немесе мәтінде дұрыс жауапты меңзейтін мәлімет бар.	Тапсырманы тексеру. Күрделендіру немесе ауыстыру мүмкіндігін қарастыру.
SOOTV-31	Infit ZSTD = 3,58, Outfit ZSTD = 3,48 (күшті underfit). Жоғары шу, болжамдылық төмен. Exact match (дәл сәйкес келу көрсеткіші) – 40,2%, күтілген деңгейден төмен (48,0%).	Тапсырма шартының анықтылығын тексеру.

В ТОБЫ: ОРТАША МӘСЕЛЕЛЕР (тексеру ұсынылады)

Тапсырмалар күткендегідей жақсы жұмыс істемейді, бірақ бұл критикалық емес.

3-кесте - Проблемалық тапсырмалардың В-тобы

Тапсырма	Мәселе	Ұсыныс
MC-14	Infit ZSTD = 2,34, Outfit ZSTD = 2,63, Pt-measure corr. = 0,16 (төмен).	Мүмкіндігінше дистракторларды тексеру.
MC-21	Infit ZSTD = 2,79, Pt-measure corr. = 0,19 (төмен).	Мүмкіндігінше тапсырма шарты тексеру.
CONT-30	Infit ZSTD = 2,00, Outfit ZSTD = 2,55 (шектелген жағдай).	Мүмкіндігінше мазмұнына талдау жүргізу.

2) Жоғары психометриялық қасиеттерге ие тапсырмалар

Тест тапсырмаларының көпшілігі (80%-дан астам) тамаша психометриялық қасиеттерді көрсетеді. Олардың ішінде: MC-1, MC-2, MC-3, MC-7, MC-10, MC-12, MC-15, MC-16, MC-17, MC-22, MC-23, MC-25, CONT-27, CONT-28, CONT-29, MC-39, MC-40 және т.б. Олар қабылданатын Infit/Outfit көрсеткіштеріне ие, жалпы баллмен жоғары корреляция көрсетеді (>0.3) және категориялардың шекті мәндері дұрыс реттелген.

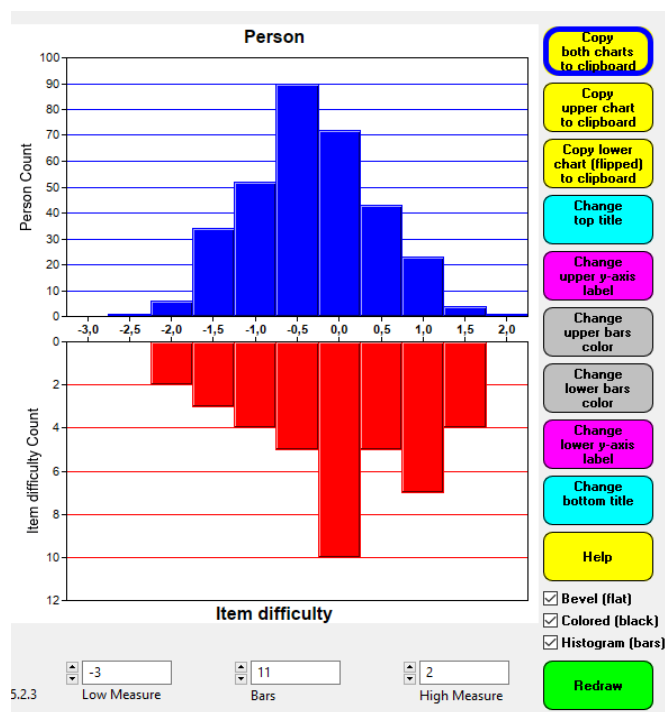
Дифференциалдық функционалдауды талдау (DIF және DPF). DIF (Differential Item Functioning) талдауы тест тапсырмаларының әртүрлі топтар үшін бірдей жұмыс істейтінін анықтауға мүмкіндік береді және бағалаудың әділдігін қамтамасыз етуде маңызды рөл атқарады [14].

DIF талдауы көрсеткендей, көптеген тапсырмаларда белгілі бір топқа артықшылық беретін жүйелі ауытқулар байқалмайды. Яғни, тест осы белгі бойынша әртүрлі оқушы топтары үшін

әділ деп қарастыруға болады.

DPF (Differential Person Functioning): жауаптары ерекше экстремалды болған жеке оқушылар анықталды (мысалы, ID 320, 322). Белгілі бір типтегі тапсырмалар бойынша олардың нәтижелері (мысалы, C типі) жалпы модельден айтарлықтай ерекшеленеді, бұл көшіріп жазу, немқұрайлылық немесе білімдегі нақты олқылықтарға байланысты болуы мүмкін.

Раш моделі бойынша гистограмманы сипаттау (Person-Item Histogram). Раш моделі аясында алынған қатысушылар мен тапсырмалардың таралу гистограммасы (1-сурет) көрсеткендей, негізгі тапсырмалар массасы тестке қатысушылардың дайындық деңгейіне қарағанда төмен қиындық деңгейінде орналасқан. Көптеген қатысушылар латенттік айнымалының орташа мәндері аймағында шоғырланған, ал тапсырмалар айтарлықтай теріс логиттер аймағына ығысқан, бұл жеңіл тапсырмалардың көлемі көбірек екенін көрсетеді.



1-сурет: Раш моделі бойынша гистограмма

Мұндай көрсеткіш тесттің төмен және орта деңгейдегі дайындықты тиімді өлшейтінін білдіреді, дегенмен жоғары дайындық деңгейі бар қатысушыларды айыра білуде шектеулі мүмкіндік бар екендігін анықтайды.

Ұсыныстар. Нәтижелерді түсіндіру кезінде тек статистикалық көрсеткіштерге емес, тапсырмалардың мазмұнына да назар аудару қажет. Бұл тәсіл тесттің валидтілігі мен сенімділігін арттыруға мүмкіндік береді.

Математика пәні бойынша тестті Раш моделі арқылы психометриялық талдау жүргізу нәтижесінде бірқатар маңызды ұсыныстар айқындалды.

- «Проблемалық» тапсырмаларды қайта қарау: MC-9 тапсырмасын қайта өңдеу (теріс корреляция көрсеткен);

- «Жоғары underfit» көрсеткіштері бар тапсырмаларға мазмұндық талдау жүргізу: SOOTV-31, MC-14, MC-21.

- «Күшті overfit» көрсеткіші бар тапсырмаларда (MC-24) болжамдылықтың бар-жоғын тексеру.

Психометриялық талдау нәтижелері бойынша тестті жетілдіру.

Математика пәні бойынша тестті Раш моделі арқылы психометриялық талдау жүргізу нәтижесінде қорытындылар жасалды, сонымен қатар кемшіліктерді қалпына келтіру бойынша жұмыстар жүргізілді.

1. **Тесттің жалпы сапасы:** Тест қатысушылар (0,83) мен тапсырмалардың (0,98) сенімділік көрсеткіштері бойынша жақсы әрі жоғары нәтижелер көрсетеді. Құрал жалпы математика білімін өлшеу және оқушыларды дайындық деңгейі бойынша дифференциациялау үшін жарамды.

2. **Модельге сәйкестік:** Деректер негізінен Раш моделіне сәйкес келеді, бұл Infit/Outfit орташа мәндерінің 1,0-ге жақын болуымен расталады.

3. **Шкала балансы:** Тест қатысушылардың қабілеттерінің орташа ауқымын жақсы камтиды, бірақ өте жақсы оқитын оқушыларға арналған қиын тапсырмалардың аздығы байқалады және бұл қатысушылар тобының аздап жоғары қиындығы бар тапсырмаларға ығысуы байқалады.

Психометриялық талдау нәтижелеріне сүйене отырып, тесттегі анықталған кемшіліктерді жою және өлшеу сапасын арттыру бойынша кешенді жұмыс жүргізілді. Ең алдымен, MC-9 тапсырмасы бойынша күтілетін психометриялық көрсеткіштерге қатысты ауытқулар тіркелді, сондықтан ол қосымша тексеру мен нақтылауға жататын тапсырмалар тізіміне енгізілді. Айқын underfit белгілері бар бірнеше тапсырмалар (SOOTV-31, MC-14, MC-21) мазмұндық сараптамадан өтті, нәтижесінде тапсырма шартының құрылымы, нұсқаулықтар және дистракторларды құру логикасы нақтыланды, екіойлылықты жою және жауаптардағы кездейсоқ шуды азайту мақсатында өзгерістер енгізілді.

Overfit белгілері бар тапсырмалар (MC-24) үшін жасырын болжам мен артық қарапайымдыққатексеру жүргізілді. Нәтижесінде кейбір тапсырмалар күрделендірілді, ал кейбірі балама шарттармен алмастырылды, бұл өзгерістер өлшенетін құрылымды сақтау арқылы жасалды.

Қорытынды

Жүргізілген талдау нәтижелері қарастырылған тест нұсқасының жалпы алғанда жеткілікті деңгейде сенімді екенін көрсетті. Тапсырмалар мен қатысушылар көрсеткіштерінің үйлесімділігі тесттің құрылымы дұрыс қалыптастырылғанын аңғартады.

Жалпы алғанда, алынған деректер Раш моделін тест сапасын бағалау мен жетілдіруде тиімді қолдануға болатынын айқындайды. Алынған нәтижелер сынақтың жалпы жақсы өлшеу қасиеттеріне ие екенін және үміткерлердің дайындық деңгейін бағалау үшін пайдалануға болатынын көрсетеді.

Дегенмен, одан әрі нақтылауды қажет ететін кейбір міндеттер анықталды. Бұларды реттеу, әсіресе дайындық деңгейі жоғары қатысушылар үшін өлшемдердің дәлдігін жақсартады.

Әдебиеттер тізімі

1. Магомедов А. Г. Применение тестирования как инструмента педагогических измерений: возможности и ограничения // Современное педагогическое образование. - 2022. - № 9. - С. 62-67.
2. Crocker L., Algina J. Introduction to Classical and Modern Test Theory. - Mason, OH : Cengage Learning, 2008. - 540 p.
3. Linacre M. Practical Rasch Measurement. Tutorial 1. Analysis of dichotomous (true-false, multiple-choice) // Winsteps. - 2011. - July. - <https://www.winsteps.com/tutorials.htm>
4. Bond T. G., Fox C. M. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. -

4th Edition. - New York : Routledge, 2021. - 520 p.

5. Wright B. D., Masters G. N. Rating Scale Analysis: Rasch Measurement. - Chicago : MESA Press, 1982. - 240 p.

6. Анализ влияния первичных баллов заданий на уровни подготовленности респондентов в дихотомической модели Раша на основе взвешенного метода максимального правдоподобия / Е. Б. Белов, М. В. Алексеев, Н. П. Китаев, А. И. Кучумов // Вестник Российского экономического университета имени Г. В. Плеханова. - 2019. - № 2 (104). - С. 169-180. - <http://dx.doi.org/10.21686/2500-3925-2019-2-89-98>

7. Engelhard G. Jr. Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences. - New York : Routledge, 2013. - 480 p.

8. Boone W. J., Staver J. R., Yale M. S. Rasch Analysis in the Human Sciences. - Dordrecht : Springer, 2014. - 270 p.

9. Bond T., Yan Z., Heene M. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. - Routledge, 2020. - 540 p.

10. Корнилов С. А. Проблема надежности психологических шкал и ее значение для психодиагностики в научных и практических целях // Экспериментальная психология в России: традиции и перспективы / под ред. В. А. Барабанщикова, А. Л. Журавлева. - М. : Изд-во «Институт психологии РАН», 2010. - С. 101-105.

11. Cronbach L. J. Coefficient alpha and the internal structure of tests // Psychometrika. - 1951. - Vol. 16, № 3. - P. 297-334. - <https://doi.org/10.1007/BF02310555>

12. Evaluating Item Difficulty and Discrimination in Knowledge Tests // Food Safety Institute. - 2024. - Jan. 4. - <https://foodsafety.institute/research-methodology/evaluating-item-difficulty-discrimination-tests/>

13. Thompson N. Classical Statistics and their Interpretation // Assess. - <https://assesshelp.zendesk.com/hc/en-us/articles/360037560611-Classical-Statistics-and-their-Interpretation>

14. Карданова Е. Ю., Иванова А. Е. Психометрические исследования: современные методы и новые возможности для образования // Вопросы образования. - 2023. - № 3. - С. 8-19. - <https://doi.org/10.17323/vo-2023-17951>

15. Wright B. D., Stone M. H. Best Test Design. - Chicago : MESA Press, 1979. - 224 p.

К.Т. Туралбаева, Г.С. Джарасова, И.Ж. Жуматова, Г.Т. Сагнаева, Б.К. Асимова

ПСИХОМЕТРИЧЕСКИЙ АНАЛИЗ ТЕСТА ПО МАТЕМАТИКЕ: ОЦЕНКА НА ОСНОВЕ МОДЕЛИ РАША

В статье представлены результаты комплексного психометрического анализа заданий по предмету «Математика», прошедших апробацию в рамках Единого национального тестирования (ЕНТ). Цель исследования заключалась в оценке измерительных свойств теста, анализе функционирования 40 заданий, а также проверке степени соответствия эмпирических данных теоретической модели и надежности теста. Для проведения анализа использовалась однопараметрическая модель Раша (Rasch model), которая позволяет учитывать сложность заданий и уровень подготовки участников на единой шкале, обеспечивая объективную оценку качества теста. В исследовании были использованы данные 326 участников, которые проходили апробацию одного варианта теста по математике. Психометрическая обработка данных осуществлялась с помощью специализированной программы WINSTEPS (версия 5.8.3.0), что позволило вычислить основные метрики теста, включая показатели сложности заданий, способности участников, соответствие модели (Infit/Outfit MNSQ и ZSTD), дискриминативность заданий и надежность инструмента. Результаты анализа показали высокий уровень общей надежности теста, что подтверждает его пригодность для оценки функциональной грамотности абитуриентов. Вместе с тем, выявлены отдельные задания, требующие содержательной корректировки и совершенствования, особенно для более подготовленных участников. Улучшение этих заданий позволит повысить точность измерений и эффективность тестирования на всех уровнях подготовки.

Ключевые слова: ЕНТ, модель Раша, психометрический анализ, трудность заданий, надежность теста, валидность.

K. Turalbaeva, G. Jarassova, I. Zhumatova, G. Sagnaeva, B. Assemova

PSYCHOMETRIC ANALYSIS OF A MATHEMATICS TEST: EVALUATION BASED ON THE RASCH MODEL

The article presents the results of a comprehensive psychometric analysis of tasks in the subject «Mathematics», which were piloted within the framework of the Unified National Testing (UNT). The purpose of the study was to evaluate the measurement properties of the test, analyze the functioning of 40 tasks, and assess the degree to which empirical data correspond to the theoretical model, as well as the reliability of the test. The analysis was conducted using the one-parameter Rasch model, which allows for accounting for task difficulty and participant ability on a single scale, providing an objective assessment of test quality. The study utilized data from 326 participants who completed the pilot of a single version of the mathematics test. Psychometric data processing was carried out using the WINSTEPS software (version 5.8.3.0), which enabled calculation of key test metrics, including task difficulty, participant ability, model fit (Infit/Outfit MNSQ and ZSTD), task discriminative power, and instrument reliability. The analysis results demonstrated a high overall level of test reliability, confirming its suitability for assessing the functional literacy of students. At the same time, several tasks were identified that require content refinement and improvement, particularly for participants with higher levels of preparation. Enhancing these tasks will increase measurement precision and improve the effectiveness of testing across all levels of participant ability.

Keywords: UNT, Rasch model, psychometric analysis, item difficulty, test reliability, validity.

References

1. Magomedov, A. G. (2022). *Primenenie testirovaniya kak instrumenta pedagogicheskikh izmereniy: vozmozhnosti i ogranicheniya* [Application of testing as a tool for pedagogical measurements: opportunities and limitations]. *Sovremennoe pedagogicheskoe obrazovanie*, (9), 62–67.
2. Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.
3. Linacre, M. (2011, July). *Practical Rasch measurement. Tutorial 1. Analysis of dichotomous (true-false, multiple-choice)*. Winsteps. <https://www.winsteps.com/tutorials.htm>
4. Bond, T. G., Fox, C. M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge.
5. Wright, B. D., Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. MESA Press.
6. Belov, E. B., Alekseev, M. V., Kitaev, N. P., & Kuchumov, A. I. (2019). *Analiz vliyaniya pervichnykh balloov zadaniy na upobni podgotovlennosti respondentov v dikhotomicheskoy modeli Rasha na osnove vzveshennogo metoda maksimal'nogo pravdopodobiya* [Analysis of the influence of primary item scores on respondents' ability levels in the dichotomous Rasch model based on the weighted maximum likelihood method]. *Vestnik Rossiyskogo ekonomicheskogo universiteta imeni G. V. Plekhanova*, (2), 169–180. <http://dx.doi.org/10.21686/2500-3925-2019-2-89-98>
7. Engelhard, G., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.
8. Boone, W. J., Staver, J. R., Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.
9. Bond, T., Yan, Z., Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.
10. Kornilov, S. A. (2010). *Problema nadezhnosti psikhologicheskikh shkal i ee znachenie dlya psikhodiagnostiki v nauchnykh i prakticheskikh tselyakh* [The problem of reliability of psychological scales and its significance for psychodiagnosics for scientific and practical purposes]. In V. A. Barabanshchikov & A. L. Zhuravlev (Eds.), *Eksperimental'naya psikhologiya v Rossii: traditsii i perspektivy* (pp. 101-105). Institute of Psychology RAS.
11. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>

12. Food Safety Institute. (2024, January 4). *Evaluating item difficulty and discrimination in knowledge tests*. <https://foodsafety.institute/research-methodology/evaluating-item-difficulty-discrimination-tests/>
13. Thompson, N. (n.d.). *Classical statistics and their interpretation*. Assess. <https://assesshelp.zendesk.com/hc/en-us/articles/360037560611-Classical-Statistics-and-their-Interpretation>
14. Kardanova, E. Y., & Ivanova, A. E. (2023). Psikhometricheskie issledovaniya: sovremennyye metody i novyye vozmozhnosti dlya obrazovaniya [Psychometric research: Modern methods and new opportunities for education]. *Voprosy obrazovaniya / Educational Studies Moscow*, (3), 8–19. <https://doi.org/10.17323/vo-2023-17951>
15. Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.

Сведения об авторах:

Туралбаева Кенжекул Турганбаевна (автор-корреспондент) – заведующий лабораторией, РГП на ПХВ «Национальный центр тестирования» МНВО РК, Астана, Казахстан, e-mail: kengebahit@mail.ru

Джарасова Гульжан Сагидуллаевна – кандидат педагогических наук, председатель, Комитет высшего и послевузовского образования, Астана, Казахстан, e-mail: yulzhan@mail.ru

Жуматова Индира Жаналыковна – заведующий лабораторией, РГП на ПХВ «Национальный центр тестирования» МНВО РК, Астана, Казахстан, e-mail: i.zh.86@bk.ru

Сагнаева Гульнар Туяковна – главный эксперт, РГП на ПХВ «Национальный центр тестирования» МНВО РК, Астана, Казахстан, e-mail: cgulnar_t@mail.ru

Асемова Бибіфатима Кәкімжанқызы – ведущий эксперт РГП на ПХВ «Национальный центр тестирования» МНВО РК, Астана, Казахстан, e-mail: assemova00@mail.ru

Авторлар туралы мәлімет:

Туралбаева Кенжекул Турганбаевна (автор-корреспондент) – зертхана меңгерушісі, ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, Астана, Қазақстан, e-mail: kengebahit@mail.ru

Джарасова Гульжан Сагидуллаевна – педагогика ғылымдарының кандидаты, Жоғары және жоғары оқу орнынан кейінгі білім комитетінің төрағасы, Астана, Қазақстан, e-mail: yulzhan@mail.ru

Жуматова Индира Жаналыковна – зертхана меңгерушісі, ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, Астана, Қазақстан, e-mail: i.zh.86@bk.ru

Сагнаева Гульнар Туяковна - бас сарапшы, ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, Астана, Қазақстан, e-mail: cgulnar_t@mail.ru

Асемова Бибіфатима Кәкімжанқызы – жетекші сарапшы, ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, Астана, Қазақстан, assemova00@mail.ru

Information about authors:

Turalbayeva Kenzhekul Turganbayevna (corresponding author) – Head of laboratory, Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Astana, Kazakhstan, e-mail: kengebahit@mail.ru

Jarassova Gulzhan Sagidullaevna – Candidate of Pedagogical Sciences, Chairman of the Committee of Higher and Postgraduate Education, Astana, Kazakhstan, e-mail: yulzhan@mail.ru

Zhumatova Indira Zhanalykovna – Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Head of laboratory, Astana, Kazakhstan, e-mail: i.zh.86@bk.ru

Sagnaeva Gulnara Tuyakovna – Chief Expert, Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Astana, Kazakhstan, e-mail: cgulnar_t@mail.ru

Assemova Bibifatima Kakimzhanovna – Senior expert, Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Astana, Kazakhstan, e-mail: assemova00@mail.ru

R. Ismailzhanova^{1*}, G. Zhabayeva², Zh. Beisenova³, D. Orlov⁴

¹Nazarbayev University Graduate School of Education, Astana, Republic of Kazakhstan,

^{2,3,4}National Testing Center, Astana, Republic of Kazakhstan,

*email: rozaoy.ismailzhanova@nu.edu.kz

¹ORCID 0009-0008-1704-1542, ²ORCID 0009-0006-1003-6316

³ORCID 0000-0002-7347-1538, ⁴ORCID 0000-0002-7699-5345

INTERNATIONAL STUDENTS AT EMI UNIVERSITIES IN KAZAKHSTAN: EXPERIENCES, CHALLENGES, AND COPING STRATEGIES

In the context of higher education globalization and Kazakhstan's active internationalization policies, attracting international students has become a strategic priority. However, research on the experiences, challenges, and coping strategies of international students at English-Medium Instruction (EMI) universities in Kazakhstan remains limited. To explore the positive experiences, challenges, and coping strategies employed by undergraduate international students at EMI universities in Kazakhstan. A qualitative phenomenological study was conducted using semi-structured interviews with seven second- and third-year international students from two EMI universities in Kazakhstan. Data were analyzed using thematic analysis. The study identified three categories of challenges: academic adaptation to new learning environments and workload, linguistic barriers (insufficient proficiency in English, Kazakh, and Russian), and socio-cultural challenges including feelings of isolation, loneliness, and culture shock. Students employed four primary coping strategies: utilizing university resources (academic advisors, international offices), socialization (participation in student clubs and events), learning local languages, and seeking personal space for emotional regulation. The findings indicate the need for a comprehensive support system for international students, including language training, cultural orientation programs, and enhanced interaction between local and international students. The results can inform the improvement of internationalization policies in Kazakhstani universities.

Keywords: international students, Kazakhstan, EMI universities, academic adaptation, linguistic barriers.

Introduction

The globalization of higher education has fundamentally transformed the landscape of academic mobility worldwide. In recent decades, the number of students pursuing education outside their home countries has increased dramatically, rising from 1.6 million in 2000 to over 6.4 million by 2022 [1]. This phenomenon reflects broader trends in internationalization, where nations seek to enhance their global competitiveness, foster cross-cultural understanding, and diversify their educational environments [2]. Kazakhstan, since gaining independence in 1991, has actively embraced internationalization as a central component of its higher education reform agenda [3].

The internationalization of Kazakhstan's higher education system has been driven by several strategic initiatives. The State Program for Education and Science Development 2020–2025 explicitly identifies the attraction of international students as a key objective, alongside the establishment of branches of leading foreign universities and the creation of a unified Central Asian educational space [4]. The Concept for the Development of Higher Education and Science for 2023–2029 further elaborates these goals, setting targets for increasing the international student population and enhancing the global visibility of Kazakhstani universities [5]. By the 2022–2023 academic year, 26,080 international students (4.5% of total enrollment) were studying in Kazakhstani higher education institutions, representing a steady increase from previous years [6].

Central to these internationalization efforts has been the development of English-Medium Instruction (EMI) universities. EMI is defined as the use of English to teach academic subjects in contexts where the majority of the population does not speak English as a first language [7]. In Kazakhstan, EMI universities have emerged as primary destinations for international students, offering programs taught entirely in English and providing a familiar linguistic environment for students from diverse backgrounds [8]. Nazarbayev University, established in 2010 as a flagship institution, exemplifies this model, with a strategic goal of enrolling 10% international students by 2025 [9].

Despite the growing presence of international students in Kazakhstani EMI universities, research on their experiences remains notably scarce. Existing studies have examined various aspects of internationalization, including policy frameworks [10], institutional strategies [11], and the general challenges faced by international students [12]. Some scholars have investigated international students' motivation to study in Kazakhstan, identifying push-pull factors such as educational costs, scholarship opportunities, and geopolitical considerations [13]. Others have explored academic and sociocultural adjustment, highlighting the critical role of language proficiency in facilitating adaptation [14]. Psychological adaptation and mental health among international students have also received attention [15].

However, a significant gap persists in understanding the lived experiences of undergraduate international students specifically within the EMI context in Kazakhstan. Previous research has not systematically examined the coping strategies these students employ to navigate academic, linguistic, and socio-cultural challenges. This gap is particularly important given the unique characteristics of Kazakhstan as a host country: a multilingual environment (Kazakh and Russian), a post-Soviet educational tradition, and a distinct cultural context that may present unfamiliar challenges for students from diverse backgrounds [16]. Understanding these experiences is essential for developing effective support systems, enhancing student retention, and achieving the internationalization goals articulated in national policy documents.

This study addresses this gap by investigating the following research questions:

1. What are the positive experiences of international students at EMI universities in Kazakhstan?
2. What challenges do undergraduate international students face while pursuing their studies at EMI universities in Kazakhstan?
3. What coping strategies do international students employ to effectively address and surmount these challenges?

The study is guided by Lysgaard's U-Curve theory of cultural adjustment [17], which provides a conceptual framework for understanding the adaptation process of sojourners in new cultural environments. According to this theory, individuals experience four stages of cultural adjustment: the initial "honeymoon" phase characterized by excitement and fascination, followed by a "disillusionment" or "culture shock" phase marked by frustration and difficulty, then an "adjustment" phase where gradual adaptation occurs, and finally a "mastery" phase where individuals function effectively in the new culture. This framework offers a lens for interpreting international students' experiences and the strategies they employ to navigate the challenges of studying abroad.

Materials and research methods

Research Approach and Design. This study employed a qualitative research approach with a phenomenological design. Phenomenology is particularly suited for exploring the lived experiences of individuals and understanding the essence of a phenomenon as it is perceived by those who have experienced it [18]. This approach allows researchers to go beyond surface-level observations and capture the subjective meanings that participants attach to their experiences [19]. The choice of

phenomenology aligns with the study's aim to understand international students' experiences in depth, focusing on their personal narratives and the meanings they construct from their time in Kazakhstan.

Research Site. The research was conducted at two EMI universities in Kazakhstan. To protect the confidentiality of participants and institutions, the universities are referred to by pseudonyms: AA University and XY University. Both institutions are recognized for their significant populations of undergraduate international students and for their commitment to English-medium instruction. AA University is characterized by a highly internationalized campus environment with dedicated support services for international students, while XY University represents a more traditional Kazakhstani university with a growing international student population. This selection allowed for a comparative exploration of how institutional contexts shape international student experiences.

Participants. Participants were recruited using criterion sampling under the purposive sampling method [20]. Criterion sampling involves selecting individuals who meet predetermined criteria relevant to the research phenomenon. For this study, the inclusion criteria were: (a) being an international student (i.e., not holding Kazakhstani citizenship), (b) enrolled in the second or third year of undergraduate studies, and (c) having completed at least one academic year in Kazakhstan. Second- and third-year students were specifically targeted because they have sufficient experience in the host country to reflect on their adaptation process, unlike first-year students who may still be in the initial adjustment phase, and they can more readily recall their challenges compared to fourth-year students who may have fully adapted.

Seven participants were recruited, a number consistent with phenomenological studies, which typically involve 5 to 25 participants to achieve depth of analysis without overwhelming data.

Table 1 - Demographic Information of Participants

Participant Code	Gender	Home Country	University	Year of Study	Duration in Kazakhstan	Ethnicity
P1_UZ_XY	Male	Uzbekistan	XY University	2nd year	2 years	Ethnic Kazakh
P2_FR_AA	Female	France	AA University	2nd year	2 years	Algerian
P3_FR_AA	Male	France	AA University	2nd year	2 years	French
P4_UZ_XY	Female	Uzbekistan	XY University	3rd year	3 years	Ethnic Kazakh
P5_ID_AA	Male	Indonesia	AA University	2nd year	2 years	Indonesian
P6_AM_XY	Female	Armenia	XY University	3rd year	8 years	Armenian
P7_MY_XY	Female	Malaysia	XY University	2nd year	2 years	Pakistani

The participants represented diverse geographic origins, including: Uzbekistan, France, Indonesia, Armenia, and Malaysia. This diversity was intentional, as it allowed for exploration of how experiences vary across different cultural backgrounds, linguistic competencies, and lengths of stay in Kazakhstan.

Data Collection. Data were collected through semi-structured interviews, a method that balances structure with flexibility. Semi-structured interviews allow researchers to follow a predetermined set of questions while remaining responsive to participants' emerging insights and allowing for elaboration on topics of particular relevance to individual experiences. This format was particularly appropriate for this study, as it enabled in-depth exploration of participants' personal experiences while ensuring consistency across interviews.

The interview protocol was developed in three languages (English, Kazakh, and Russian) to accommodate participants' linguistic preferences. The protocol consisted of open-ended questions organized around three main themes corresponding to the research questions: positive experiences, challenges, and coping strategies. Given that many participants returned to their home countries during the winter holiday period, interviews were conducted online using the Zoom platform. Each interview lasted between 30 and 45 minutes and was audio-recorded with explicit participant consent. Interviews were conducted between November 27 and December 15, 2023.

Data Analysis. Data analysis followed the thematic analysis procedures outlined by Creswell. The analysis proceeded through several iterative stages. First, interviews were transcribed manually to ensure accuracy and preserve nuances in participants' responses. Transcriptions were then read multiple times to achieve familiarization with the data. Next, open coding was applied to identify meaningful segments of text. Codes were generated inductively from the data, capturing both anticipated themes related to the research questions and unexpected themes that emerged from participants' narratives. Through iterative review, initial codes were grouped into broader categories, which were then refined into final themes.

Ethical Considerations. Ethical approval was obtained from the Graduate School of Education Institutional Research Ethics Committee at Nazarbayev University prior to data collection. All participants were provided with detailed information about the study's purpose, procedures, potential risks, and their rights. Informed consent was obtained in writing before interviews commenced. Participants were assured of their right to withdraw from the study at any time without penalty. Confidentiality was maintained through the use of pseudonyms and anonymization of university names. Audio recordings and transcripts were stored on a password-protected computer accessible only to the researcher.

Results and discussion

The findings are organized according to the three research questions, with each section presenting thematic analyses of participants' experiences. The results are discussed in relation to Lysgaard's U-Curve theory and existing literature on international student adaptation.

Positive Experiences of International Students. The analysis revealed that international students' positive experiences varied according to the duration of their stay in Kazakhstan and their proficiency in local languages. Participants consistently highlighted the welcoming atmosphere of their universities and local communities. A participant from France described her experience:

"It's an incredible experience because I think everything here is easy for international students, especially at AA University. When I first came, people were very welcoming and kind." (P2_FR_AA)

This sentiment was echoed by a participant from Indonesia:

"The university has been supportive, offering resources that aid adaptation. Despite cultural differences, the local community has been welcoming, providing a sense of belonging. People are open and ready to help in different situations, which I am grateful for." (P5_ID_AA)

These findings align with the first phase of Lysgaard's U-Curve theory, the "honeymoon" stage, characterized by initial excitement and fascination with the new environment [17]. The positive reception described by participants appears to have contributed to this initial positive adjustment.

For some participants, studying in Kazakhstan represented a significant life transition, described as a "new chapter." A participant from Algeria who had previously experienced discrimination in France explained:

"I faced racism and discrimination in France. I decided to stop my studies in France because of this issue, and so after stopping my studies, I didn't know what I would do and was just thinking about restarting a new university chapter. So, my parents introduced me to Kazakhstan... I didn't face issues I had encountered in France." (P2_FR_AA)

This narrative illustrates the transformative potential of international study, where a new environment can offer opportunities for personal renewal and escape from prior difficulties. This finding resonates with research by Wang and Hannes, who highlight the role of cultural and educational immersion in fostering a sense of renewal and personal development among international students.

Challenges of Adaptation to New Academic Environment. Despite positive experiences, participants faced significant academic challenges that reflected the "disillusionment" phase of the U-Curve theory, where initial enthusiasm gives way to frustration as the realities of daily life in a new culture become apparent.

The most frequently reported academic challenge was adapting to the new academic environment and curriculum. Participants described the study load as "massive" and the expectations as substantially different from their previous educational experiences. A participant from Uzbekistan explained:

"After the education system in Uzbekistan, it was challenging for me to study at university here. There, I face a different workload. It was a massive workload for the first time in Kazakhstan, but now I know how to study and do assignments." (P4_UZ_XY)

This challenge was compounded by participants' initial expectations about the academic rigor of their universities. Several participants had anticipated that their studies would be relatively easy, given that their university was not highly ranked in global league tables. This expectation was quickly contradicted by the demanding academic requirements they encountered. The discrepancy between expectations and reality created additional stress, as participants had to quickly revise their assumptions about academic life in Kazakhstan. This finding aligns with Zhou et al., who found that unrealistic expectations about host country experiences can exacerbate adjustment difficulties.

The academic challenges identified in this study resonate with findings from other contexts. Research on international students in the United Kingdom and the United States has similarly documented difficulties related to differing educational systems, unfamiliar pedagogical approaches, and demanding academic workloads.

Linguistic Challenges. Linguistic barriers emerged as a critical factor affecting both academic performance and social integration. Participants who were not proficient in Russian or Kazakh reported feeling excluded when their peers shifted to local languages during group work and informal interactions. A participant from Algeria described a particularly poignant experience:

"I was in the lab, and everyone spoke Kazakh and Russian. And I was the only one there. And I felt very left out, you know, I was like, where am I? Am I in the right place?" (P2_FR_AA)

This sense of exclusion extended beyond the classroom to everyday interactions in the local community. Participants who did not speak Russian or Kazakh found it difficult to communicate with shopkeepers, taxi drivers, and other community members, limiting their ability to navigate daily life independently. This finding is consistent with research by Spencer-Oatey and Xiong, who found that inadequate language abilities constrain international students' ability to form friendships with locals.

In contrast, participants who were proficient in Russian or Kazakh reported significantly smoother adaptation. A participant from Uzbekistan noted:

"I speak Russian and Kazakh languages, therefore I easily interact with local people and made friends quickly." (P1_UZ_XY)

For participants studying in EMI environments, insufficient English proficiency also posed challenges. Some participants found it difficult to understand lectures, analyze readings, and express themselves in academic writing. Research by McMahon similarly found that English proficiency serves as a "gateway" to academic and social success for international students, with those lacking proficiency facing substantial barriers.

Socio-Cultural Challenges and Culture Shock. Participants from XY University reported socio-cultural challenges that significantly impacted their experiences. These challenges included feelings of isolation, loneliness, and difficulty forming meaningful relationships with local students. A participant from Uzbekistan described a particularly difficult living situation:

"Even with my roommate, we do not talk, and sometimes I feel lonely and miss my country... my parents ask me, 'Who's your roommate?' But I didn't know what to say because I didn't know with whom I lived." (P1_UZ_XY)

This finding resonates with research by Hendrickson et al., who found that limited interaction with local students and peers contributes significantly to feelings of loneliness among international students.

Another participant from Uzbekistan described a traumatic experience that highlighted the absence of expected social support:

"My roommate, she was not friendly with me, did not offer any help. It's just she was on her own and did not even ask how I was doing. I tried talking to her, but she always ignored me." (P4_UZ_XY)

This participant had broken her leg and expected, based on her cultural background, that her roommate would offer assistance and emotional support. The absence of this support compounded her physical difficulties with emotional distress, illustrating how cultural differences in expectations about interpersonal relationships can create significant adjustment challenges.

Beyond interpersonal difficulties, participants reported experiencing culture shock related to broader cultural differences. A participant from Malaysia described what she called a "no smile culture" in Kazakhstan:

"There's no smile culture here at all, right? So if you go to Malaysia, people are just smiling at you... people here are not that approachable, helpful, you know, hard to talk to." (P7_MY_XY)

This observation reflects a common aspect of culture shock: the disorientation that occurs when familiar cultural cues are absent or replaced by unfamiliar ones. The participant's experience highlights how subtle cultural differences can significantly affect international students' sense of belonging and comfort.

Coping Strategies. Utilizing University Support and Resources. The most frequently reported coping strategy was seeking help from university resources, particularly academic advisors, professors, and international student offices. A participant from France explained:

"Whenever I have an issue concerning my studies or something, I ask my advisor, and she's always ready to help me and has always resolved my problems." (P2_FR_AA)

Research by Akanwa similarly found that the availability of support services significantly influences international students' success and well-being.

Socialization as a Coping Strategy. Participants actively engaged in social activities as a means of overcoming challenges. Joining student clubs was identified as particularly beneficial. A participant from Indonesia described:

"I joined international student groups and participated in cultural events, fostering a supportive network." (P5_ID_AA)

These findings align with research by Zhou et al., who found that extracurricular activities help students forge meaningful connections and facilitate cultural exchange.

Learning Local Languages. Another strategy participants employed was actively learning local languages. A participant from France described his approach:

"My challenges were related to my poor language skills, and I did my best to try to hear what people were saying, learn it, and adapt it to myself to use it. That's how I managed to go shopping." (P3_FR_AA)

This participant also emphasized the symbolic importance of learning local languages as a sign of respect for the host country. This finding aligns with Kim's assertion that language competence is crucial for effective cultural adjustment.

Seeking Privacy as a Coping Strategy. A unique coping strategy identified was seeking privacy for emotional regulation. One participant described:

"Whenever I'm having a bad day, I would go out and treat myself... go for junk food, have coffee when facing challenges." (P7_MY_XY)

This finding reflects the diversity of coping strategies that individuals employ and aligns with Lazarus and Folkman's stress and coping theory, which recognizes the importance of emotion-focused strategies in managing psychological distress.

Conclusion

This phenomenological study provides a comprehensive understanding of the experiences, challenges, and coping strategies of undergraduate international students at EMI universities in Kazakhstan. The findings reveal that international students' experiences are shaped by multiple factors, including the duration of their stay, proficiency in local languages, and the institutional context of their university.

The study identified three primary categories of challenges: (1) academic adaptation to new environments and curricula; (2) linguistic barriers; and (3) socio-cultural challenges, including feelings of isolation, loneliness, and culture shock. In response, students employed four main coping strategies: (1) utilizing university resources; (2) socialization through student clubs and events; (3) learning local languages; and (4) seeking privacy for emotional regulation.

The findings have significant implications for policy and practice. Universities should strengthen orientation programs, provide targeted language support, promote interaction between local and international students, and ensure counseling services are culturally sensitive and accessible. As Kazakhstan continues to pursue its internationalization goals, understanding and addressing the needs of international students will be essential for attracting and retaining this important population.

References

1. UNESCO Institute for Statistics. Global flow of tertiary-level students. - 2023. - <http://uis.unesco.org>
2. De Wit H. Globalisation and internationalisation of higher education // Revista de Universidad y Sociedad del Conocimiento (RUSC). - 2011. - Vol. 8, № 2. - P. 241-248.
3. Kuzhabekova A. Internationalization as a mechanism of higher education modernization in Kazakhstan // The Bloomsbury handbook of internationalization of higher education in the global south / Ed. by F. Huang. - London : Bloomsbury Publishing, 2020. - P. 100-115. <https://doi.org/10.5040/9781350139275.ch-008>
4. State program for education and science development in Kazakhstan for 2020-2025 / Ministry of Education and Science of the Republic of Kazakhstan. - Astana, 2019. - 64 p.

5. Concept for the development of higher education and science for 2023-2029 / Ministry of Education and Science of the Republic of Kazakhstan. - Astana, 2022. - 58 p.
6. National Information Center on Academic Recognition and Mobility. Sistema vysshego obrazovaniya v Kazahstane. - 2023. - <https://enic-kazakhstan.edu.kz>
7. Dearden J. English as a medium of instruction: A growing global phenomenon. - London : British Council, 2015. - 62 p.
8. Karabassova L. Understanding trilingual education reform in Kazakhstan: Why is it stalled? // Education, equity, economy: Education in Central Asia a kaleidoscope of challenges and opportunities / Ed. by D. Egea. - Cham : Springer, 2020. - P. 37-50. https://doi.org/10.1007/978-3-030-50127-3_3
9. Strategy 2018-2030 / Nazarbayev University. - Astana, 2018. - 42 p.
10. Jumakulov Z., Ashirbekov A. Higher education internationalization: Insights from Kazakhstan // Hungarian Educational Research Journal. - 2016. - № 6. - P. 38–58. DOI:10.14413/HERJ.2016.01.03
11. Internationalization of higher education in Kazakhstan: From political will to implementation / A. Akkari, A. Seidikenova, A. Bakitov, G. Minazheva // Ensaio: Avaliação e Políticas Públicas em Educação. - 2023. - Vol. 31, № 119. - P. 1-26. <https://doi.org/10.1590/S0104-40362023003103730>
12. Kim T. Academic and sociocultural adjustment experiences of international students studying in Kazakhstan : unpublished master's thesis / Nazarbayev University. - Astana, 2021. - 112 p.
13. Оспанова З. Социально-психологические особенности мотивации к учебной деятельности иностранных студентов медицинских специальностей // Гуманитарные науки: гуманитарные исследования. - 2018. - № 1 (31). - С. 56-61. <https://doi.org/10.17238/issn1998-5320.2018.31.56>
14. Mukhamajanova D. International students in Kazakhstan: A narrative inquiry of human agency in the process of adaptation // International Journal of Comparative Education and Development. - 2019. - Vol. 21, № 2. - P. 82-96. <https://doi.org/10.1108/IJCED-07-2018-0024>
15. Depression, anxiety and somatic distress in domestic and international undergraduate medical students in Kazakhstan / Z. Ibrayeva, D. Aldyngurov, A. Myssayev [et al.] // Iranian Journal of Public Health. - 2018. - Vol. 47, № 6. - P. 919-921.
16. Kuzhabekova A., Lee J. Internationalization and local research capacity strengthening: Factors affecting knowledge sharing between international and local faculty in Kazakhstan // European Education. - 2020. - Vol. 52, № 3. - P. 215-230. <https://doi.org/10.1080/10564934.2020.1723422>
17. Lysgaard S. Adjustment in a foreign society: Norwegian Fulbright grantees visiting the United States // International Social Science Bulletin. - 1955. - № 7. - P. 45-51.
18. Creswell J. W. Qualitative inquiry and research design: Choosing among five approaches. - 3rd ed. - Thousand Oaks : Sage Publications, 2013. - 448 p.
19. Merriam S. B., Tisdell E. J. Qualitative research: A guide to design and implementation. - 4th ed. - San Francisco : Jossey-Bass, 2016. - 368 p.
20. Patton M. Q. Qualitative research and evaluation methods. - 4th ed. - Thousand Oaks : Sage Publications, 2015. - 832 p.

Р. Исмаилжанова, Г.К. Жабаева, Ж.Ж. Бейсенова, Д.М. Орлов

ҚАЗАҚСТАНДАҒЫ АҒЫЛШЫН ТІЛІНДЕ ОҚЫТАТЫН ЖОҒАРЫ ОҚУ ОРЫНДАРЫНДАҒЫ ШЕТЕЛДІК СТУДЕНТТЕР: ТӘЖІРИБЕСІ, ҚИЫНДЫҚТАРЫ ЖӘНЕ ЕҢСЕРУ СТРАТЕГИЯЛАРЫ

Жоғары білімнің жаһандануы және Қазақстанның интернационалдандыру саясаты аясында шетелдік білім алушыларды тарту стратегиялық басымдыққа айналды. Алайда, қазақстандық ағылшын тілінде білім беретін университеттеріндегі шетелдік білім алушылардың тәжірибесі, проблемалары және қиындықтарды жеңу стратегиялары туралы зерттеулер шектеулі. Қазақстандағы ағылшын тілінде білім беретін университеттерінде оқитын бакалавриат шетелдік білім алушыларының оң тәжірибелерін, қиындықтарын және оларды жеңу стратегияларын зерттеу. Феноменологиялық сапалық зерттеу шеңберінде Қазақстанның екі ағылшын тілінде білім беретін университетінде оқитын екінші және

үшінші курс жеті шетелдік білім алушымен жартылай құрылымдалған сұхбат жүргізілді. Деректерді талдау тақырыптық талдау әдісімен жүзеге асырылды. Зерттеу қиындықтардың үш тобын анықтады: жаңа академиялық ортаға және оқу жүктемесіне бейімделу, тілдік кедергілер (ағылшын, қазақ және орыс тілдерін жеткіліксіз меңгеру), әлеуметтік-мәдени қиындықтар, соның ішінде оқшаулану, жалғыздық сезімі және мәдени шок. Білім алушылар қиындықтарды жеңу үшін төрт негізгі стратегияны қолданды: университет ресурстарына жүгіну (академиялық кеңесшілер, халықаралық кеңселер), әлеуметтену (студенттік клубтар мен іс-шараларға қатысу), жергілікті тілдерді үйрену және эмоционалды реттеу үшін жеке кеңістік іздеу. Нәтижелер шетелдік білім алушыларға тілдік дайындықты, мәдени-бағдарлау бағдарламаларын және жергілікті және шетелдік білім алушылар арасындағы өзара әрекеттестікті күшейтуді қамтитын кешенді қолдау жүйесін құру қажеттілігін көрсетеді. Алынған мәліметтер қазақстандық университеттердегі интернационалдандыру саясатын жетілдіру үшін пайдаланылуы мүмкін.

Түйін сөздер: шетелдік білім алушылар, Қазақстан, ЕМІ университеттері, академиялық бейімделу, тілдік кедергілер.

Р. Исмаилжанова, Г.К. Жабаева, Ж.Ж. Бейсенова, Д.М. Орлов

ИНОСТРАННЫЕ СТУДЕНТЫ В КАЗАХСТАНСКИХ ВУЗАХ С АНГЛИЙСКИМ ЯЗЫКОМ ОБУЧЕНИЯ: ОПЫТ, ТРУДНОСТИ И СТРАТЕГИИ ПРЕОДОЛЕНИЯ

В условиях глобализации высшего образования и активной политики интернационализации Казахстана привлечение иностранных обучающихся становится стратегическим приоритетом. Однако исследования, посвященные опыту, проблемам и стратегиям преодоления трудностей иностранных обучающихся в казахстанских университетах с английским языком обучения (ЕМІ), остаются ограниченными. Изучить позитивный опыт, трудности и стратегии преодоления, которые используют иностранные обучающиеся бакалавриата в ЕМІ университетах Казахстана. В рамках феноменологического качественного исследования проведены полуструктурированные интервью с семью иностранными обучающимися второго и третьего курсов из двух ЕМІ университетов Казахстана. Анализ данных осуществлялся методом тематического анализа. Исследование выявило три группы трудностей: академическая адаптация к новой учебной среде и учебной нагрузке, языковые барьеры (недостаточное владение английским, казахским и русским языками), социокультурные вызовы, включая чувство изоляции, одиночества и культурный шок. Обучающиеся использовали четыре основные стратегии преодоления: обращение к университетским ресурсам (академические консультанты, международные офисы), социализацию (участие в студенческих клубах и мероприятиях), изучение местных языков и поиск личного пространства для эмоциональной регуляции. Результаты свидетельствуют о необходимости создания комплексной системы поддержки иностранных обучающихся, включающей языковую подготовку, культурно-ориентационные программы и усиление взаимодействия между местными и иностранными обучающимися. Полученные данные могут быть использованы для совершенствования политики интернационализации в казахстанских университетах.

Ключевые слова: иностранные обучающиеся, Казахстан, ЕМІ университеты, академическая адаптация, языковые барьеры.

References

1. UNESCO Institute for Statistics. (2023). *Global flow of tertiary-level students*.
2. De Wit, H. (2011). Globalisation and internationalisation of higher education. *Revista de Universidad y Sociedad del Conocimiento (RUSC)*, 8(2), 241–248.
3. Kuzhabekova, A. (2020). Internationalization as a mechanism of higher education modernization in Kazakhstan. In F. Huang (Ed.), *The Bloomsbury handbook of internationalization of higher education in the global south* (pp. 100–115). Bloomsbury Publishing. <https://doi.org/10.5040/9781350139275.ch-008>

4. Ministry of Education and Science of the Republic of Kazakhstan. (2019). *State program for education and science development in Kazakhstan for 2020-2025*.
5. Ministry of Education and Science of the Republic of Kazakhstan. (2022). *Concept for the development of higher education and science for 2023-2029*.
6. National Information Center on Academic Recognition and Mobility. (2023). *Sistema vysshego obrazovaniya v Kazahstane*.
7. Dearden, J. (2015). *English as a medium of instruction: A growing global phenomenon*. British Council.
8. Karabassova, L. (2020). Understanding trilingual education reform in Kazakhstan: Why is it stalled? In D. Egea (Ed.), *Education, equity, economy: Education in Central Asia a kaleidoscope of challenges and opportunities* (pp. 37–50). Springer. https://doi.org/10.1007/978-3-030-50127-3_3
9. Nazarbayev University. (2018). *Strategy 2018-2030*.
10. Jumakulov, Z., Ashirbekov, A. (2016). Higher education internationalization: Insights from Kazakhstan. *Hungarian Educational Research Journal*, (6), 38–58. DOI:10.14413/HERJ.2016.01.03
11. Akkari, A., Seidikenova, A., Bakitov, A., & Minazheva, G. (2023). Internationalization of higher education in Kazakhstan: From political will to implementation. *Ensaio: Avaliação e Políticas Públicas em Educação*, 31(119), 1–26. <https://doi.org/10.1590/S0104-40362023003103730>
12. Kim, T. (2021). *Academic and sociocultural adjustment experiences of international students studying in Kazakhstan* [Unpublished master's thesis]. Nazarbayev University.
13. Ospanova, Z. (2018). Sotsialno-psikhologicheskie osobennosti motivatsii k uchebnoy deyatel'nosti inostrannykh studentov meditsinskikh spetsialnostey [Socio-psychological characteristics of motivation for academic activity of foreign medical students]. *Gumanitarnye nauki: gumanitarnye issledovaniya*, 1(31), 56–61. <https://doi.org/10.17238/issn1998-5320.2018.31.56>
14. Mukhamajanova, D. (2019). International students in Kazakhstan: A narrative inquiry of human agency in the process of adaptation. *International Journal of Comparative Education and Development*, 21(2), 82–96. <https://doi.org/10.1108/IJCED-07-2018-0024>
15. Ibrayeva, Z., Aldyngurov, D., Myssayev, A., et al. (2018). Depression, anxiety and somatic distress in domestic and international undergraduate medical students in Kazakhstan. *Iranian Journal of Public Health*, 47(6), 919–921.
16. Kuzhabekova, A., Lee, J. (2020). Internationalization and local research capacity strengthening: Factors affecting knowledge sharing between international and local faculty in Kazakhstan. *European Education*, 52(3), 215–230. <https://doi.org/10.1080/10564934.2020.1723422>
17. Lysgaard, S. (1955). Adjustment in a foreign society: Norwegian Fulbright grantees visiting the United States. *International Social Science Bulletin*, 7, 45–51.
18. Creswell, J. W. (2013). *Qualitative inquiry and research design: Choosing among five approaches* (3rd ed.). Sage Publications.
19. Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative research: A guide to design and implementation* (4th ed.). Jossey-Bass.
20. Patton, M. Q. (2015). *Qualitative research and evaluation methods* (4th ed.). Sage Publications.

Information about the authors:

Ismailzhanova Rozaoy (corresponding author) - master of Arts in Multilingual Education, Graduate School of Education, Nazarbayev University, Astana, Republic of Kazakhstan. email: rozaoy.ismailzhanova@nu.edu.kz.

Zhabaeva Guldana Kuatovna – master, «National Testing Center», Leading Researcher, Astana, Republic of Kazakhstan, e-mail: guldanzhabaeva@gmail.com.

Beisenova Zhanar Zhumagazinovna - master, «National Testing Center», Leading Researcher, Astana, Republic of Kazakhstan, e-mail: zhanar-beisenova@mail.ru.

Orlov David – PhD, «National Testing Center», Chief Researcher, Astana, Republic of Kazakhstan, e-mail: dorlov@alumni.nu.edu.kz.

Автор туралы мәліметтер:

Исмаилжанова Розаой (автор-корреспондент) – «Көптілді білім беру» мамандығы бойынша білім магистрі, Жоғары білім беру мектебі, Назарбаев университеті, Астана қ., Қазақстан Республикасы, email: rozaoy.ismailzhanova@nu.edu.kz.

Жабаева Гульдана Куатовна – магистр, «Ұлттық тестілеу орталығы», жетекші ғылыми қызметкер, Астана қ., Қазақстан Республикасы, e-mail: guldanzhabayeva@gmail.com.

Бейсенова Жанар Жумагазиновна – магистр, «Ұлттық тестілеу орталығы», жетекші ғылыми қызметкер, Астана қ., Қазақстан Республикасы, e-mail: zhanar-beisenova@mail.ru.

Орлов Давид Михайлович – PhD докторы, «Ұлттық тестілеу орталығы», бас ғылыми қызметкер, Астана қ., Қазақстан Республикасы, e-mail: dorlov@alumni.nu.edu.kz.

Сведения об авторах:

Исмаилжанова Розаой (автор-корреспондент) - магистр образования по специальности «Мультиязычное образование», Высшая школа образования, Назарбаев университет, г. Астана, Республика Казахстан, e-mail: rozaoy.ismailzhanova@nu.edu.kz.

Жабаева Гульдана Куатовна - магистр, «Национальный центр тестирования», ведущий научный сотрудник, г. Астана, Республика Казахстан, e-mail: guldanzhabayeva@gmail.com.

Бейсенова Жанар Жумагазиновна - магистр, «Национальный центр тестирования», ведущий научный сотрудник, г. Астана, Республика Казахстан, e-mail: zhanar-beisenova@mail.ru.

Орлов Давид Михайлович - PhD доктор, «Национальный центр тестирования», главный научный сотрудник, г. Астана, Республика Казахстан, e-mail: dorlov@alumni.nu.edu.kz.

Ауезханова А.Ж.^{1*}, Ерсұлтанова Т.С.², Қуандық А.Қ.³, Әлен С.Ә.⁴
^{1,2,3,4}ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, Астана қ., Қазақстан Республикасы
*email: auezkhanova72@gmail.com

¹ORCID 0009-0005-9335-9593, ²ORCID 0000-0003-3286-9096,

³ORCID 0009-0001-4370-4636, ⁴ORCID 0009-0009-2079-4331

2025 ЖЫЛЫ ПББ АПРОБАЦИЯСЫ АЯСЫНДА ӨТКІЗІЛГЕН ТЕСТІЛЕУ НӘТИЖЕЛЕРІНІҢ ПСИХОМЕТРИЯЛЫҚ ТАЛДАУЫ

Бұл зерттеу Қазақстан Республикасында педагогтердің білімін бағалауда қолданылатын физика пәні тест тапсырмаларының психометриялық сипаттамаларына талдауға арналған. Зерттеу 2025 жылы өткізілген педагогтердің білімін бағалау (ПББ) апробациясының нәтижелеріне негізделген. Талдау барысында Раш өлшеу моделі қолданып, тесттің сенімділігі, валидтілігі және дифференциациялық қабілеті бағаланды. Жалпы апробацияға 6000-нан астам педагог қатысқанымен, талдау үшін 118 қатысушыдан тұратын репрезентативті іріктеме пайдаланылды. Зерттеу нәтижелері тесттің жоғары сенімділігін көрсетті (Item Reliability = 0,99), тапсырмалардың 92%-ы валидтілік критерийлеріне сәйкес келді (Infit/Outfit MNSQ 0,5–1,5 аралығында). Сонымен қатар, тапсырмалардың қиындық деңгейі логит шкаласында бағаланып, дистракторлық талдау жүргізілді. Кейбір тапсырмаларда сәйкессіздік байқалып, қайта қарау ұсынылады. Тестіленушілер мен тапсырмалардың қабілеті мен күрделілігі арасындағы сәйкестік тесттің білім деңгейін кең ауқымда бағалауға жарамды екенін көрсетті. 2022 жылдан бастап енгізілген тақырыптық талдау педагогтерге өз нәтижелерін терең түсінуге мүмкіндік береді. Зерттеу нәтижелері қолданылған тесттің педагогтердің пәндік білім деңгейін объективті бағалауға жарамды екенін дәлелдейді және тест тапсырмаларының сапасын одан әрі жетілдіруге бағытталған ұсыныстар ұсынады.

Түйін сөздер: педагогтердің білімін бағалау, психометриялық талдау, тест, Раш моделі, сенімділік, валидтілік.

Кіріспе

ПББ тарихы және зерттеудің өзектілігі. Қазақстан Республикасында педагогтердің білімін бағалау жүйесі кезең-кезеңімен жетілдіріліп, нормативтік-құқықтық және технологиялық тұрғыдан жүйелі жаңғыртудан өтті. Бұл өзгерістер бағалау рәсімдерінің ашықтығын, объективтілігін және сенімділігін қамтамасыз етуге бағытталды.

2010 жылдан бастап педагогтердің білімін бағалау мазмұндық тұрғыдан жаңартылған кезеңі басталды. Осы кезеңде біліктілікті растаудың негізгі құралы ретінде педагог қызметкерлерді кешенді тестілеу енгізілді. Бұл жүйе Қазақстан Республикасы Білім және ғылым министрінің 2010 жылғы 22 қаңтардағы №16 «Педагог қызметкерлерді аттестаттау қағидаларын бекіту туралы» бұйрығы [1] негізінде жүзеге асырылды. Аталған құжат мектепке дейінгі тәрбие мен оқыту, бастауыш, негізгі орта және жалпы орта білім беру, техникалық және кәсіптік, орта білімнен кейінгі, қосымша білім беру бағдарламаларын іске асыратын білім беру ұйымдарының педагогтеріне қатысты болды.

2018 жылдан бастап педагог қызметкерлердің білімін бағалау мазмұны мен құрылымы қайта қарастырылып, тестілеудің жаңа форматы – Ұлттық біліктілік тестілеуі енгізілді. Осы жылы тестілеу қағаз форматында өткізілсе, 2019 жылдан бастап тестілеу электрондық форматқа көшірілді. Аталған өзгеріс бағалау үдерісінің ашықтығын, жеделдігін және объективтілігін арттыруға мүмкіндік берді, сондай-ақ адами фактордың ықпалын төмендетуге ықпал етті.

2020 жылдан бастап бүгінгі күнге дейін Ұлттық біліктілік тестілеуді арнайы бағдарламалық қамтамасыз ету арқылы жүргізіліп келеді. Жаңа жүйе тестілеу рәсімін толық автоматтандыруға, нәтижелердің дәлдігі мен сенімділігін қамтамасыз етуге бағытталған.

2022 жылдан бастап Педагогтердің білімін бағалау (ПББ) форматы енгізілді [2]. Тестілеу педагогтердің пәндік білімдерін бағалауға бағытталған. Максималды балл саны - 50. Тапсырмалар қиындық деңгейі бойынша үш деңгейге бөлінеді: базалық (26%), орташа (60%) және жоғары (14%). Тест 50 тапсырмадан тұрады. Тестілеуден өту үшін педагог кемінде барлық тапсырманың 50 %-на дұрыс жауап беруі қажет.

Педагогтердің біліктілік санатына сәйкес шекті балл деңгейлері белгіленген [2]:

«педагог» / «педагог-тағылымдамашы» – кемінде 50%;

«педагог-модератор» біліктілік санаты – кемінде 60%;

«педагог-сарапшы» біліктілік санаты – кемінде 70%;

«педагог-зерттеуші» біліктілік санаты – кемінде 80%;

«педагог-шебер» біліктілік санаты – кемінде 90%.

Осыған байланысты тест тапсырмаларының сапасын ғылыми негізде бағалау өзекті болып табылады. Зерттеудің мақсаты – физика пәні бойынша тест тапсырмаларының сапасын Раш моделі негізінде психометриялық талдау.

Халықаралық тәжірибеде педагогтердің білімін бағалаудың білім сапасына тікелей әсер ететінін көрсетеді [3]. АҚШ-та педагогтердің білімін бағалау үшін Praxis емтихандары пәндік білім мен педагогикалық құзыреттілікті кешенді түрде өлшейді [4], ал edTPA моделі мұғалімнің практикалық оқыту дағдыларын бағалауға бағытталған [5].

Ұлыбританияда мұғалімдер Qualified Teacher Status (QTS) алу үшін кәсіби стандарттарға сәйкестігін дәлелдеуі тиіс [5]. Оңтүстік Корея мен Қытайда педагогтерді іріктеу орталықтандырылған емтихандар арқылы жүзеге асырылады [3].

Финляндияда формалды тестілеу жүйесі қолданылмайды, алайда мұғалімдерді даярлау үдерісі жоғары талаптары мен қатаң іріктеуге негізделген [5]. TALIS халықаралық зерттеуі педагогтердің кәсіби дамуы мен оқыту тәжірибесін салыстырмалы түрде талдауға мүмкіндік береді [3]. Бұл тәжірибелер педагогтерді бағалау жүйесі ғылыми негізделген, сенімді және көпдеңгейлі болуы қажет екенін көрсетеді [6, 7].

Зерттеу материалдары мен әдістері

Зерттеу дизайны. Бұл зерттеу сандық әдіснамаға негізделген және тест тапсырмаларының психометриялық сипаттамаларын талдауға бағытталған. Негізгі талдау құралы ретінде Раш өлшеу моделі (кесте 1) қолданылды [6, 8]. Бұл модель тест тапсырмаларының қиындық деңгейін және тестіленушілердің қабілетін бір логиттік шкалада бағалауға мүмкіндік береді [8]. Тест тапсырмаларының базасын қалыптастыру үдерісі бірнеше кезеңнен тұрады: әзірлеу, бірінші және екінші сараптама, сондай-ақ бірінші және екінші түзету кезеңдері [9]. Тапсырмаларды әзірлеушілер мен сарапшылар ретінде колледждер, орта мектептер және жоғары және (немесе) жоғары оқу орнынан кейінгі білім беру ұйымдары (ЖЖОКБҰ) ұсынған тәжірибелі педагогтері мен оқытушылары іріктелді. Бұл құрылым тесттің ғылыми негізделген және әділ болуын қамтамасыз етеді [9, 10].

Деректер және құралдар. Зерттеу 2025 жылғы ПББ апробациясы нәтижелеріне негізделген [9]. Жалпы қатысушылар саны 6000-нан асқанымен, осы зерттеуде 188 педагогтің деректері талданды.

Тест 50 тапсырмадан тұрады:

- 40 – бір дұрыс жауапты таңдау;
- 10 – контекстке негізделген тапсырмалар.

Максималды балл – 50. Тапсырмалар дихотомиялық форматта ұсынылған, бұл классикалық тест теориясы мен Item Response Theory талаптарына сәйкес келеді [7].

Деректерді талдау. Деректер Winsteps бағдарламасы арқылы өңделді. Келесі көрсеткіштер талданды:

- сенімділік;
- валидтілік (Infit, Outfit MNSQ);
- қиындық деңгейі (логит);
- корреляция;
- дифференциациялық қабілет.

Валидтілік Infit және Outfit MNSQ статистикалары арқылы бағаланды (қалыпты диапазон: 0,5–1,5) [6, 11]. Сонымен қатар, дистракторлық талдау жүргізіліп, жауап нұсқаларының тиімділігі бағаланды [10].

Әр тапсырма бір дұрыс жауаптан және бірнеше дистрактордан тұрады, тапсырмалар күрделілік деңгейі мен когнитивтік дағдыларға сәйкес әзірленген [10]. Сонымен қатар, тапсырмалар мен жалпы балл арасындағы корреляция есептелді, бұл тесттің педагогтердің қабілетін объективті бағалау қабілетін арттырды [12].

Нәтижелер және талқылау

Психометриялық талдау нәтижелері тесттің жоғары сапасын көрсетті. Тапсырмалардың орташа стандартты қатесі 0,23 логитті құрады, бұл өлшеу дәлдігінің жоғары екенін көрсетеді [6]. Тесттің ішкі сенімділігі Cronbach $\alpha = 0,83$, стандартталған сенімділік 0,84, SEM = 2,98, яғни тест педагогтердің қабілетін сенімді бағалауға қабілетті [10, 12]. Тапсырмалардың 92%-ы валидті, қалғандарында сәйкессіздік байқалды. Тапсырмалардың қиындық деңгейі кең диапазонда таралған, бұл тесттің әртүрлі деңгейдегі педагогтерді ажыратуға қабілетті екенін көрсетеді [7, 8].

Дистракторлық талдау нәтижесінде жауап нұсқаларының тиімділігін көрсетті.

Кесте 1 - Раш моделіндегі қорытынды статистика

Көрсеткіш	Орташа мән (Mean)	Стандарттық қате (SEM)	Стандарттық ауытқу (SD)	Минимум	Максимум
Жалпы ұпай (Score)	23.0	0.7	7.2	9.0	47.0
Өлшем (Measure, логит)	-0.14	0.08	0.85	-1.98	3.32
Қате (S.E.)	0.34	0.00	0.04	0.32	0.61
Infit MNSQ	1.00	0.02	0.19	0.65	1.52
Outfit MNSQ	1.03	0.03	0.36	0.55	2.18
Infit ZSTD	-0.11	0.11	1.20	-2.95	2.85
Outfit ZSTD	0.00	0.11	1.16	-2.30	3.13

Кесте 1-де көрсетілгендей, 118 педагогтің орташа ұпайы 23,0 (максималды 50 ұпайдан), қабілет деңгейі (Measure) -0,14 логит, яғни тест тапсырмаларының орташа күрделілік деңгейі тестіленушілердің қабілет деңгейінен сәл жоғары екенін білдіреді [8]. Тестіленушілердің қабілет деңгейінің стандартты ауытқуы 0,85 логитті құрады, бұл іріктеме ішінде білім деңгейінің жеткілікті деңгейде әртүрлі екенін көрсетеді. Модельге сәйкестік көрсеткіштері қалыпты аралықта (Infit MNSQ 1,00; Outfit MNSQ 1,03), ZSTD нөлге жақын, бұл деректердің Раш моделіне жақсы сәйкес келетінін көрсетеді [6, 11].

Сенімділік көрсеткіші (Person Reliability) 0,82–0,84 аралығында болды, бұл тесттің педагогтердің қабілет деңгейін сенімді түрде ажыратуға қабілетті екенін көрсетеді [11]. Жалпы алғанда, Кесте 1 нәтижелері тесттің педагогтердің білім деңгейін бағалауда тиімді құрал екенін көрсетеді [6, 8], алайда тест құрылымын одан әрі жетілдіру мақсатында жекелеген көрсеткіштерге назар аудару ұсынылады.

Кесте 2 - Тапсырмалар деңгейіндегі психометриялық көрсеткіштер (Раш моделі)

TABLE 14.1 118-1-428174				ZOU296v5.TXT Dec 01 2025 16:18									
INPUT: 118 Person 50 Item				REPORTED: 118 Person 50 Item 2 CATS WINSTEPS 5.8.3.0									
Person: REAL SEP.: 2.15 REL.: .82				Item: REAL SEP.: 5.36 REL.: .97									
Item STATISTICS: ENTRY ORDER													
ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	EMLE MEASURE	MODEL S.E.	MNSQ ZSTD	INFI T	OUTFI T	PTMEASUR-AL CORR.	EXP.	EXACT MATCH	Item		
1	93	118	-1.64	.24	.85	-1.17	.73	-1.30	.45	.27	78.8	79.1	10001
2	23	118	1.45	.25	.90	-.66	.98	-.04	.42	.34	84.7	82.0	10002
3	44	118	.43	.20	.91	-1.20	.85	-1.41	.46	.35	72.0	68.1	10003
4	80	118	-1.01	.21	.95	-.64	.87	-.88	.38	.31	72.0	70.0	10004
5	52	118	.10	.20	.98	-.28	.98	-.21	.37	.35	66.1	65.0	10005
6	34	118	.87	.22	1.28	2.57	1.49	3.07	.01	.35	65.3	74.0	10006
7	36	118	.77	.21	1.11	1.10	1.15	1.13	.23	.35	72.0	72.6	10007
8	109	118	-2.90	.35	1.04	.25	1.04	.24	.12	.18	92.4	92.4	10008
9	40	118	.60	.21	1.07	.81	1.12	1.01	.27	.35	69.5	70.1	10009
10	93	118	-1.64	.24	.93	-.54	1.45	1.87	.31	.27	78.8	79.1	10010
11	81	118	-1.05	.21	.87	-1.53	.82	-1.18	.45	.31	78.0	70.6	10011
12	98	118	-1.94	.25	.90	-.59	.77	-.87	.37	.25	84.7	83.1	10012
13	26	118	1.28	.24	1.18	1.32	1.33	1.66	.12	.34	78.0	79.8	10013
14	22	118	1.52	.25	.99	-.03	1.20	.95	.30	.33	83.9	82.8	10014
15	35	118	.82	.22	1.15	1.45	1.16	1.17	.19	.35	71.2	73.3	10015
16	58	118	-.13	.20	.79	-3.64	.75	-2.89	.58	.35	81.4	64.2	10016
17	68	118	-.52	.20	.93	-1.02	.90	-.92	.41	.33	69.5	65.6	10017
18	47	118	.30	.20	1.20	2.72	1.35	3.18	.10	.35	61.0	66.7	10018
19	111	118	-3.18	.40	.98	.04	.65	-.66	.23	.16	94.1	94.1	10019
20	89	118	-1.43	.22	.80	-1.92	.66	-1.92	.53	.28	76.3	75.9	10020
21	16	118	1.95	.29	1.03	.22	1.12	.50	.28	.32	87.3	87.3	10021
22	54	118	.03	.20	1.08	1.23	1.09	.93	.26	.35	64.4	64.6	10022
23	17	118	1.87	.28	1.07	.43	1.56	1.89	.16	.32	86.4	86.5	10023
24	98	118	-1.94	.25	.90	-.63	.72	-1.00	.39	.25	83.1	83.1	10024
25	22	118	1.52	.25	.99	.00	1.00	.07	.34	.33	82.2	82.8	10025
26	49	118	.22	.20	1.10	1.49	1.17	1.71	.23	.35	59.3	66.0	10026
27	46	118	.34	.20	.93	-.97	1.03	.30	.41	.35	70.3	67.1	10027
28	24	118	1.39	.24	1.00	.03	1.29	1.41	.28	.34	84.7	81.3	10028
29	55	118	-.01	.20	.94	-1.03	.90	-1.13	.43	.35	63.6	64.4	10029
30	49	118	.22	.20	1.01	.17	1.03	.36	.34	.35	69.5	66.0	10030
31	64	118	-.36	.20	.85	-2.57	.81	-1.97	.51	.34	74.6	64.8	10031
32	79	118	-.97	.21	.93	-.87	.91	-.57	.39	.31	71.2	69.5	10032
33	36	118	.77	.21	.98	-.23	.93	-.48	.39	.35	70.3	72.6	10033
34	30	118	1.06	.23	.95	-.41	.86	-.86	.42	.35	77.1	76.9	10034
35	56	118	-.05	.20	1.03	.53	1.00	.00	.32	.35	59.3	64.3	10035
36	43	118	.47	.20	1.01	.12	1.01	.14	.34	.35	69.5	68.6	10036
37	32	118	.96	.22	1.07	.69	1.30	1.85	.23	.35	75.4	75.5	10037
38	112	118	-3.34	.42	.95	-.01	.65	-.59	.24	.15	94.9	94.9	10038
39	18	118	1.79	.27	1.08	.50	1.36	1.37	.19	.32	84.7	85.7	10039
40	54	118	.03	.20	1.05	.85	1.07	.77	.29	.35	64.4	64.6	10040
41	101	118	-2.14	.27	.99	.02	.96	-.05	.24	.23	85.6	85.6	10041
42	60	118	-.21	.20	.89	-1.92	.86	-1.56	.47	.34	72.0	64.3	10042
43	51	118	.14	.20	.95	-.79	.92	-.79	.41	.35	68.6	65.3	10043
44	41	118	.55	.21	1.04	.50	1.02	.24	.31	.35	68.6	69.6	10044
45	12	118	2.31	.32	.90	-.36	1.19	.65	.34	.30	91.5	90.4	10045
46	48	118	.26	.20	1.15	2.13	1.35	3.19	.15	.35	58.5	66.3	10046
47	78	118	-.92	.21	.99	-.08	.96	-.22	.33	.32	66.1	68.9	10047
48	50	118	.18	.20	.94	-.83	.90	-1.02	.42	.35	67.8	65.6	10048
49	39	118	.64	.21	.99	-.13	1.00	.00	.36	.35	72.9	70.7	10049
50	42	118	.51	.21	1.04	.57	1.06	.52	.30	.35	66.1	69.1	10050
MEAN	54.3	118.0	.00	.23	.99	-.09	1.03	.15			74.8	74.2	
P.S.D	27.3	.0	1.32	.05	.10	1.18	.22	1.33			9.4	9.0	
TABLE 14.3 118-1-428174				ZOU296v5.TXT Dec 01 2025 16:18									
INPUT: 118 Person 50 Item				REPORTED: 118 Person 50 Item 2 CATS WINSTEPS 5.8.3.0									

Раш моделі негізінде жүргізілген талдау тест тапсырмаларының сапасын жан-жақты бағалауға мүмкіндік берді [6, 8]. Кесте 2-де көрсетілгендей, тапсырмалардың орташа қиындық деңгейі 0.00 логит, диапазон -3.34-тен 2.31-ге дейін. Бұл диапазон тест құрамында өте жеңіл және өте күрделі тапсырмалардың бар екенін және оның тестіленушілердің әртүрлі қабілет деңгейлерін ажыратуға мүмкіндік беретінін дәлелдейді. № 6, 8, 13, 15, 18, 46, 23, 39

тапсырмалар корреляция көрсеткіштері төмен ($<0,20$), қайта қарауды қажет етеді. Ал, № 11, 17, 27, 43, 29, 48, 34 тапсырмалар модельге жақсы сәйкес келеді.

Корреляция коэффициенті төмен ($<0,20$) болған тапсырмаларға сәйкес келетін тақырыптар педагогтердің білімін бағалау сапасына әсер ететін маңызды мазмұндық аймақтарды көрсетеді. Жоғарыда баяндалғандай, 2022 жылдан бастап енгізілген тақырыптық талдау педагогтерге өз нәтижелерін терең түсінуге мүмкіндік береді. Талдау нәтижелері бойынша төмен корреляция байқалған негізгі тақырыптар төмендегідей:

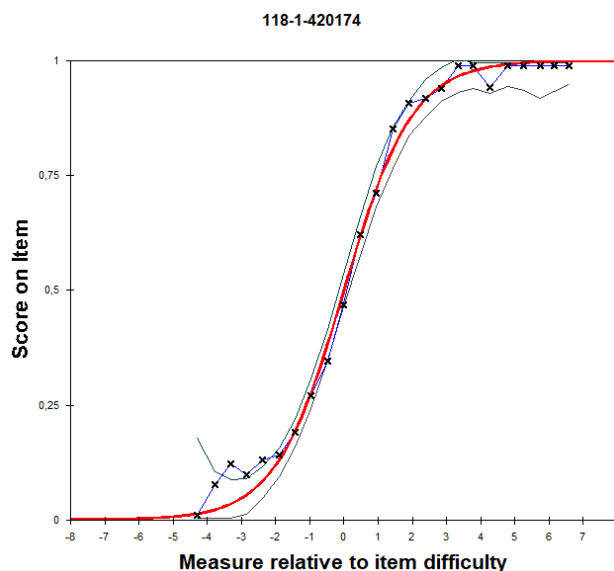
- Динамика (№6)
- Қысым (№8)
- Сақталу заңдары (№13)
- Тербелістер мен толқындар (№15)
- Жылу құбылыстары және заттың агрегаттық күйлері (№18)
- Термодинамика негіздері (№23)
- Атом ядросының физикасы (№39)

Тапсырмалардың жалпы тест нәтижесімен әлсіз байланысы олардың күрделілік деңгейінің сәйкес келмеуінен (тым жеңіл немесе тым күрделі), мазмұнның кешенділігінен және бірнеше дағдыны қатар талап етуінен туындауы мүмкін. Жоғары білім деңгейіндегі педагогтердің өзі тым жеңіл тапсырмаларға дұрыс жауап бере алмауы мүмкін, бұл олардың дискриминациялық қабілетін төмендетеді. Сондықтан аталған тақырыптардағы тапсырмалар мазмұндық, құрылымдық және психометриялық тұрғыдан қайта қарауды қажет етеді.

Ал, №11 (Статика), №17 (Сұйықтар мен газдар механикасы), №27 (Электростатика негіздері), №29 (Тұрақты электр тогы), №34 (Толқындық оптика), №43 және №48 (Мәнмәтін) тапсырмалар Раш моделіне жақсы сәйкес келеді. Олар нақты дағдыны өлшейді, мазмұны айқын, дистракторлары сапалы және тесттің сенімділігі мен валидтілігін арттырады.

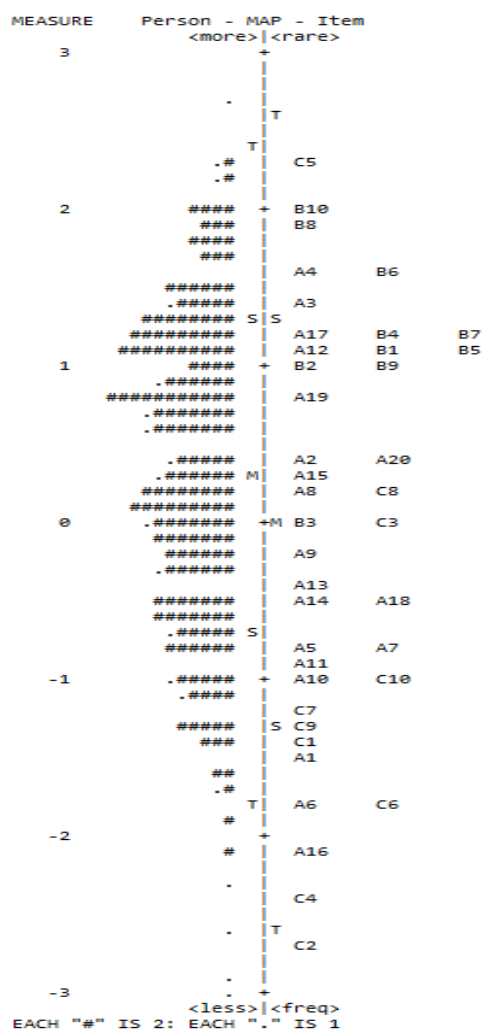
1-суретте тест тапсырмасының (Item) күрделілік деңгейіне (difficulty) байланысты тестіленушінің дұрыс жауап беру ықтималдығын көрсетеді.

Дегенмен, 1-сурет тест тапсырмасының күрделілік деңгейіне байланысты тестіленушінің дұрыс жауап беру ықтималдығы сызығы тесттің педагогтердің білім деңгейіне сәйкес келетінін дәлелдейді [8, 13].



1 сурет - Тестіленушілер мен тапсырмалардың қабілеті және күрделілік деңгейлерінің таралу сызығы

2-суретте көрсетілген тестіленушілердің қабілет деңгейі мен тапсырмалардың күрделілік деңгейі арасындағы сәйкестікті береді. Картада тестіленушілердің қабілет деңгейі (сол жақта) және тапсырмалардың күрделілік деңгейі (оң жақта) шамамен –3-тен +3 логит аралығында таралғаны байқалады.



2 сурет - Қатысушылар мен тапсырмалар картасын талдау

Егер тестіленушілердің қабілеті мен тапсырмалардың күрделілігі шкалада шамамен бірдей таралса, онда тест педагогтердің барлық деңгейін тиімді бағалайды деп айтуға болады. Мысалы, тапсырма С5 шкалада жоғарырақ орналасқан, бұл қиындық деңгейі «А» тапсырма, дегенмен ол педагогтер үшін күрделі тапсырма болғанын көрсетеді. Ал тестіленушілердің басым бөлігі шкаланың ортаңғы бөлігінде шоғырланған, яғни орташа қабілет деңгейі бар [14, 15].

Бұл карта тесттің деңгейлік балансын бағалауға және қажет болған жағдайда тапсырмаларды қайта қарауға мүмкіндік береді.

Талқылау. Раш моделі тапсырмаларды дәл бағалауға мүмкіндік береді. Кейбір тапсырмалар талапқа сай емес, сондықтан тест базасын жетілдіру қажет. 2022 жылдан бастап енгізілген тақырыптық талдау педагогтердің кері байланысын жақсартып, кәсіби дамуға септігін тигізеді.

Қорытынды

Физика пәні бойынша тест педагогтердің білімін бағалауда сенімді және валидті құрал екенін көрсетті. Зерттеу нәтижелері тесттердің жоғары сенімділігін (Item Reliability = 0,99) және тапсырмалардың 92%-ының валидтілік критерийлеріне сәйкес келетінін растады (Infit/Outfit MNSQ 0,5–1,5 аралығында). Сондай-ақ тапсырмалардың қиындық деңгейі логит шкаласында бағаланып, дистракторлық талдау жүргізілді. Бұл нәтижелер педагогтердің білімін объективті бағалауға мүмкіндік беретінін және тесттің білім деңгейін кең ауқымда бағалауға жарамдылығын дәлелдейді.

Әдебиеттер тізімі

1. Педагог қызметкерлерді аттестаттау қағидаларын бекіту туралы : Қазақстан Республикасы Білім және ғылым министрінің 2010 жылғы 22 қаңтардағы № 16 бұйрығы. – Астана, 2010.
2. Педагогтерді аттестаттаудан өткізу қағидалары мен шарттарын бекіту туралы : Қазақстан Республикасы Оқу-ағарту министрінің 2016 жылғы 27 қаңтардағы № 83 бұйрығы. – Астана, 2016.
3. OECD. TALIS 2018 Results (Volume I): Teachers and School Leaders as Lifelong Learners. – Paris: OECD Publishing, 2019. – P. 55–110. - <https://doi.org/10.1787/1d0bc92a-en>
4. ETS. Praxis Test Technical Manual. – Princeton : Educational Testing Service, 2022. – P. 15–60. - DOI: 10.1080/02619768.2017.1315399
5. Darling-Hammond L. Teacher Education around the World: What Can We Learn? / L. Darling-Hammond // European Journal of Teacher Education. – 2017. – Vol. 40, № 3. – P. 291–309. - <https://doi.org/10.1080/02619768.2017.1315399>
6. Bond T. G. Applying the Rasch Model: Fundamental Measurement in the Human Sciences / T. G. Bond, C. M. Fox. – 3rd ed. – New York : Routledge, 2015. – 230 p. - DOI:10.4324/9781315814698
7. Hambleton R. K. Fundamentals of Item Response Theory / R. K. Hambleton, H. Swaminathan, H. J. Rogers. – Newbury Park : Sage, 1991. – 174 p.
8. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests / G. Rasch. – Chicago : University of Chicago Press, 1980. – 199 p.
9. Тест тапсырмаларының базасын қалыптастыру қағидалары. – Астана : Ұлттық тестілеу орталығы, 2026. – 12–20 б.
10. DeVellis R. F. Scale Development: Theory and Applications / R. F. DeVellis. – 4th ed. – Los Angeles : Sage, 2017. – 280 p.
11. Linacre J. M. A User's Guide to Winsteps Rasch-Model Computer Programs / J. M. Linacre. – Chicago : Winsteps.com, 2023. – 250 p.
12. Field A. Discovering Statistics Using IBM SPSS Statistics / A. Field. – 5th ed. – London : Sage, 2018. – 1104 p.
13. Boone W. J. Rasch Analysis in the Human Sciences / W. J. Boone, J. R. Staver, M. S. Yale. – Dordrecht : Springer, 2014. – 482 p.
14. Wright B. D. Rating Scale Analysis/B.D. Wright, G. N. Masters. - Chicago : MESA Press, 1982. - 206p.
15. Brookhart S. M. How to Create and Use Rubrics for Formative Assessment and Grading / S. M. Brookhart. – Alexandria : ASCD, 2013. – 160 p.

Ауезханова А.Ж., Ерсұлтанова Т.С., Куандық А.К., Ален С.А.

ПСИХОМЕТРИЧЕСКИЙ АНАЛИЗ АПРОБАЦИИ ТЕСТИРОВАНИЯ ДЛЯ ОЦЕНКИ ЗНАНИЙ ПЕДАГОГОВ ПО ФИЗИКЕ В 2025 ГОДУ

Данное исследование посвящено психометрическому анализу тестовых заданий по физике, применяемых для оценки уровня знаний педагогов в Республике Казахстан. Исследование базируется на результатах апробации педагогического тестирования (ПББ) 2025 года. В анализе использовалась модель Раша для оценки надежности, валидности и дифференциационной способности теста. Несмотря на то,

что в апробации приняли участие более 6000 педагогов, для анализа была использована репрезентативная выборка из 118 участников. Результаты исследования показали высокую надежность теста (Item Reliability = 0,99), 92% заданий соответствовали критериям валидности (Infit/Outfit MNSQ в диапазоне 0,5–1,5). Кроме того, оценивался уровень сложности заданий в логитной шкале и проводился дистракторный анализ. Некоторые задания выявили несоответствия и требуют пересмотра. Сопоставление уровня знаний участников и сложности заданий показало, что тест эффективно оценивает уровень знаний педагогов в широком диапазоне. С 2022 года введенный тематический анализ позволяет педагогам глубже понять свои результаты. Полученные данные подтверждают, что используемый тест является надежным и валидным инструментом для объективной оценки предметных знаний педагогов и формируют рекомендации по дальнейшему совершенствованию заданий.

Ключевые слова: оценка знаний педагогов, психометрический анализ, тест, модель Раша, надежность, валидность.

A. Auezhanova, T. Ersultanova, A. Kuandyk, S. Alen

PSYCHOMETRIC ANALYSIS OF THE 2025 PILOT TESTING FOR ASSESSING THE KNOWLEDGE OF PHYSICS TEACHERS

This study focuses on the psychometric analysis of physics test items used to assess teachers' knowledge in the Republic of Kazakhstan. The research is based on the results of the 2025 pilot of the Teacher Knowledge Assessment (TKA). The Rasch measurement model was applied to evaluate the reliability, validity, and discriminatory power of the test. Although over 6,000 teachers participated in the pilot, a representative sample of 118 participants was analyzed. The results demonstrated high test reliability (Item Reliability = 0.99), with 92% of items meeting validity criteria (Infit/Outfit MNSQ within 0.5–1.5). Additionally, item difficulty was measured on the logit scale, and a distractor analysis was conducted. Some items revealed inconsistencies and require revision. The alignment between participants' ability levels and item difficulty indicated that the test effectively differentiates teachers across a broad range of knowledge levels. Since 2022, the introduction of thematic analysis has allowed teachers to gain deeper insights into their results. Overall, the findings confirm that the test is a reliable and valid instrument for the objective assessment of teachers' subject knowledge and provide recommendations for further item refinement.

Keywords: teacher knowledge assessment, psychometric analysis, test, Rasch model, reliability, validity.

References

1. Ministry of Education and Science of the Republic of Kazakhstan. (2010). On approval of the rules for certification of pedagogical workers (Order No. 16). Astana.
2. Ministry of Education of the Republic of Kazakhstan. (2016). On approval of rules and conditions for certification of teachers (Order No. 83). Astana.
3. OECD. (2019). *TALIS 2018 results (Volume I): Teachers and school leaders as lifelong learners*. OECD Publishing. <https://doi.org/10.1787/1d0bc92a-en>
4. Educational Testing Service. (2022). *Praxis test technical manual*. <https://doi.org/10.1080/02619768.2017.1315399>
5. Darling-Hammond, L. (2017). Teacher education around the world: What can we learn? *European Journal of Teacher Education*, 40(3), 291–309. <https://doi.org/10.1080/02619768.2017.1315399>
6. Bond, T. G., Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge. <https://doi.org/10.4324/9781315814698>
7. Hambleton, R. K., Swaminathan, H., Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.
8. Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.
9. National Testing Center. (2026). *Rules for the formation of the database of test items* (pp. 12–20). Astana.
10. DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Sage.
11. Linacre, J. M. (2023). *A user's guide to Winsteps Rasch-model computer programs*. Winsteps.com.

12. Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage.
13. Boone, W. J., Staver, J. R., Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.
14. Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.
15. Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.

Автор туралы мәліметтер:

Ауезханова Анар Жумағалиевна (автор-корреспондент) – ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, Орта білімдегі педагогтердің білімін бағалау тест тапсырмаларын қалыптастыру ғылыми-әдістемелік зертханасының меңгерушісі, Астана қ., Қазақстан Республикасы, email: auezkhanova72@gmail.com

Ерсұлтанова Турсынкуль Серикбаевна - ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, Педагогтердің білімін бағалау тест тапсырмаларын қалыптастыру басқармасының басшысы, Астана қ., Қазақстан Республикасы, email: Tursynkul-07@mail.ru

Қуандық Аманжол Қасқырбайұлы - ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, Орта білімдегі педагогтердің білімін бағалау тест тапсырмаларын қалыптастыру ғылыми-әдістемелік зертханасының аға сарапшысы, Астана қ., Қазақстан Республикасы, email: Amanzhol.uandy@bk.ru

Әлен Сабира Әділқызы - ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, Орта білімдегі педагогтердің білімін бағалау тест тапсырмаларын қалыптастыру ғылыми-әдістемелік зертханасының жетекші сарапшысы, Астана қ., Қазақстан Республикасы, email: Sabira.alen@bk.ru

Сведения об авторах:

Ауезханова Анар Жумағалиевна (автор-корреспондент) - РГП на ПХВ «Национальный центр тестирования» МНВО РК, заведующий научно-методической лабораторией формирования тестовых заданий оценки знаний педагогов среднего образования, Астана, Республика Казахстан, email: auezkhanova72@gmail.com

Ерсұлтанова Турсынкуль Серикбаевна - РГП на ПХВ «Национальный центр тестирования» МНВО РК, руководитель управления формирования тестовых заданий оценки знаний педагогов, г. Астана, Республика Казахстан, email: Tursynkul-07@mail.ru

Қуандық Аманжол Қасқырбайұлы - РГП на ПХВ «Национальный центр тестирования» МНВО РК, старший эксперт научно-методической лаборатории формирования тестовых заданий оценки знаний педагогов среднего образования, г. Астана, Республика Казахстан, email: Amanzhol.uandy@bk.ru

Әлен Сабира Әділқызы - РГП на ПХВ «Национальный центр тестирования» МНВО РК, ведущий эксперт научно-методической лаборатории формирования тестовых заданий оценки знаний педагогов среднего образования, г. Астана, Республика Казахстан, email: Sabira.alen@bk.ru

Information about the authors:

Auezhanova Anar (corresponding author) – Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Head of the Scientific-Methodological Laboratory for Developing Test Items for Assessing the Knowledge of Secondary Education Teachers, Astana, Republic of Kazakhstan, email: auezkhanova72@gmail.com

Ersultanova Tursynkul - Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Head of the Department for Developing Test Items for Assessing Teacher Knowledge, Astana, Republic of Kazakhstan, email: Tursynkul-07@mail.ru

Kuandyk Amanzhol - Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Senior Expert of the Scientific-Methodological Laboratory for Developing Test Items for Assessing the Knowledge of Secondary Education Teachers, Astana, Republic of Kazakhstan, email: Amanzhol.uandy@bk.ru

Alen Sabira - Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Leading Expert of the Scientific-Methodological Laboratory for Developing Test Items for Assessing the Knowledge of Secondary Education Teachers, Astana, Republic of Kazakhstan, email: Sabira.alen@bk.ru

ЖАСАНДЫ ИНТЕЛЛЕКТ ЖӘНЕ БІЛІМ БЕРУДІ БАҒАЛАУДАҒЫ ИННОВАЦИЯ

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ИННОВАЦИИ В ОЦЕНИВАНИИ ОБРАЗОВАНИЯ

ARTIFICIAL INTELLIGENCE AND INNOVATIONS IN EDUCATIONAL ASSESSMENT

К.В. Тарасова

Центр психометрики и измерений в образовании, г. Москва, Российская Федерация

e-mail: ktarasova@hse.ru

ORCID: 0000-0002-3915-3165

ПСИХОМЕТРИКА И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: НОВАЯ ПАРАДИГМА ОБРАЗОВАТЕЛЬНОГО ОЦЕНИВАНИЯ

В статье анализируется, каким образом искусственный интеллект трансформирует современное образовательное оценивание и почему эти изменения требуют пересмотра привычных психометрических оснований измерения. Отправной точкой выступает противоречие между быстрым внедрением генеративного ИИ в практики разработки заданий, автоматической проверки и адаптивного тестирования и недостаточной методологической проработанностью этих процессов. Цель статьи состоит в том, чтобы концептуализировать возникающую психометрическую повестку и показать, что ИИ следует рассматривать не как внешний технический модуль, а как фактор, меняющий архитектуру оценивания на всех этапах - от проектирования конструкта до интерпретации результатов. Статья имеет концептуально-аналитический характер и структурирована вокруг трех ключевых направлений: автоматической генерации тестовых заданий и стимульного материала с использованием больших языковых моделей; автоскоринга текстовых и графических открытых ответов; оценивания ИИ-грамотности и иных сложных компетенций в цифровой среде. Показано, что потенциал ИИ особенно высок в масштабируемой генерации контента, обработке мультимодальных данных и построении интерактивных сценариев оценивания, однако технологическая эффективность сама по себе не гарантирует валидности, справедливости и интерпретируемости результатов. В качестве центрального вывода обосновывается необходимость гибридной модели, в которой ИИ расширяет операционные возможности оценивания, но не отменяет психометрическую экспертизу и контроль человека в контуре принятия решений. Особое значение в статье придается подходу Evidence-Centered Design как рамке для перехода от декларативного измерения к поведенческому оцениванию действий обучающегося в среде взаимодействия с ИИ. Тем самым статья вносит вклад в развитие методологии оценивания в условиях цифровой трансформации образования и задает ориентиры для научно обоснованной интеграции ИИ в измерительные системы.

Ключевые слова: психометрика, искусственный интеллект, образовательное оценивание, автоматическая генерация заданий.

Введение

Искусственный интеллект (ИИ) сегодня все чаще рассматривается не как отдельный технологический инструмент, а как инфраструктурная технология, меняющая саму логику образовательных процессов. За несколько десятилетий развитие ИИ прошло путь от относительно простых алгоритмических моделей, таких как линейная и логистическая регрессия, к сложным архитектурам глубокого обучения и большим языковым моделям (large language models, LLM), основанным на трансформерах. Именно LLM (GPT-линейка, GigaChat, YandexGPT и другие) стали «лицом» современного ИИ для широкой аудитории, поскольку демонстрируют способность порождать связный текст, поддерживать контекст на больших отрезках дискурса и учитывать тонкие смысловые нюансы.

Расширение сфер применения ИИ, от промышленности и медицины до экономики, искусства и повседневных сервисов, неизбежно отражается и на образовании. В образовательной среде ИИ уже используется в интеллектуальных обучающих системах, адаптивном обучении, рекомендательных системах, прогнозировании образовательных и

карьерных траекторий, автоматизированном оценивании и предоставлении персонализированной обратной связи. При этом оценивание и измерение становятся одной из ключевых областей, где ИИ не только позволяет оптимизировать отдельные процедуры (например, автоматическую проверку заданий с открытым ответом), но и радикально пересматривает саму архитектуру инструментов тестирования, от проектирования конструкта до анализа результатов.

Психометрика в условиях широкого внедрения ИИ приобретает ключевую роль: технологии расширяют возможности оценивания - от генерации заданий и мультимодальных форматов до адаптивных процедур и автоматического скоринга, - но одновременно требуют пересмотра того, как обеспечиваются валидность, устойчивость параметров, контроль предвзятости и корректная интеграция новых типов данных в существующие модели измерения.

Материалы и методы исследования

Автоматическая генерация заданий и стимульного материала. Одной из наиболее динамично развивающихся областей является автоматическая генерация стимульного материала и тестовых заданий. Потребность в такого рода решениях обусловлена, во-первых, масштабом современных мониторинговых исследований и экзаменов с высокими ставками, требующих больших и регулярно обновляемых банков заданий, и, во-вторых, переходом к более сложным форматам оценивания, включая интерактивные и мультимодальные сценарии. Ручная разработка большого количества заданий и стимульного материала в этих условиях становится крайне затратной и часто ограничивает разнообразие форматов.

Эволюция технологий генерации контента для оценивания хорошо иллюстрирует общую логику развития ИИ. Ранние подходы к автоматической генерации заданий опирались на относительно простые модели: либо на подбор числовых параметров в рамках заранее заданных шаблонов, либо на формирование так называемых моделей заданий с радикальными и сопроводительными элементами (radicals and incidentals). Радикальные элементы задают психометрически существенные характеристики задания (например, тип когнитивной операции), тогда как сопроводительные (детали контекста, конкретные числовые значения, варианты ответов) могут варьироваться без изменения латентной структуры. Такой подход позволял получать «клоны» заданий с предсказуемыми психометрическими свойствами, но оставался жестко ограниченным рамками вручную прописанных моделей: он плохо масштабируется, слабо поддерживает контентное разнообразие и практически не выходит за пределы однотипных, относительно простых когнитивных требований.

Появление трансформеров и последующее развитие больших языковых моделей радикально изменило ситуацию. Модели, такие как BERT, GPT-х и их аналоги, обучаются на гигантских корпусах текстов и содержат триллионы параметров [1-3]. Масштаб и архитектура этих моделей позволяют им генерировать длинные, контекстно согласованные и стилистически разнообразные тексты, которые по своим поверхностным характеристикам часто неотличимы от текстов, созданных человеком. Это открывает принципиально новые возможности для генерации стимульного материала и заданий для образовательного оценивания - от информационных и художественных текстов до описаний сложных ситуаций и сценариев.

Ряд эмпирических исследований демонстрирует потенциал LLM для решения задач, связанных с разработкой тестового контента. Так, в работах команды Duolingo English Test GPT-3 использовалась для генерации текстов и заданий для взрослых обучающихся, при этом весь процесс строился в формате «человек в цикле ИИ» (human in the loop): эксперты по оцениванию отбирали и редактировали варианты, проверяя их на наличие ошибок, предвзятости и соответствие спецификациям теста [4, 5]. Аналогичный подход реализован и в исследованиях,

посвященных генерации текстов для оценки читательской грамотности на уровне начальной школы (например, в рамках PIRLS): LLM генерирует множество вариантов текстов, после чего проводится многоступенчатый отбор на основе показателей сложности и экспертной оценки содержания [6].

Важно, что в этих работах генеративные возможности ИИ интегрируются в существующие психометрические рамки, а не подменяют их: автоматически созданные тексты не рассматриваются как «готовые» задания, а являются исходным материалом, который проходит процедуру отбора, калибровки и валидации, сопоставимую по строгости с традиционной разработкой тестов.

Следующий этап развития связан с мультимодальными и иммерсивными форматами. Генеративные модели позволяют создавать не только текст, но и изображения, аудио- и видеоконтент, что существенно расширяет спектр доступных стимулов. В оценивании это выражается в переходе от статичных текстовых задач к мультимодальным сценариям, где текст комбинируется с визуальной информацией, звуком, динамическими объектами и интерактивными элементами [7, 8]. Использование AR/VR-сред, дополненных генеративными возможностями ИИ, делает возможным моделирование сложных ситуаций реальной деятельности и оценку поведения человека в контекстах, приближенных к реальным [9-11]. При этом когнитивные и поведенческие индикаторы могут фиксироваться автоматически, а параметры среды - адаптироваться в режиме реального времени в зависимости от действий участника.

Однако переход к таким форматам обостряет и ряд фундаментальных вопросов психометрики. Несмотря на впечатляющие примеры использования LLM и других моделей ИИ, большинство существующих исследований по автоматической генерации заданий до сих пор слабым образом опираются на теорию измерений [12, 13]. Во многих случаях целевые конструкторы, образовательные цели и когнитивные уровни (например, в терминах таксономии Блума) не артикулируются явно; генерация фокусируется на заданиях, проверяющих запоминание и базовое понимание, тогда как более высокие уровни (применение, анализ, оценка, создание) остаются в тени. Немногочисленные работы, в которых LLM целенаправленно используется для генерации заданий более высоких когнитивных уровней (например, в физике для старшей школы [14]), показывают, что это технически возможно, но требует гораздо более тщательного проектирования промптов и интеграции теоретических моделей обучения в сам процесс генерации.

В этой связи особое значение приобретает промптинг - разработка запросов к языковой модели, которые явно кодируют цели обучения, ожидаемый когнитивный уровень, формат задания, критерии качества и ограничения по содержанию [1, 15, 16]. Техники zero-shot, one-shot и few-shot prompting, instructional prompting и другие подходы позволяют не только управлять формальными характеристиками генерируемого материала, но и «встраивать» в него образовательную логику: таксономии, спецификации теста, требуемые типы когнитивных операций. Фактически, промпт становится новой единицей проектирования в психометрике, связывающей между собой модель конструктора, модель задания и модель оценивания.

Наконец, существующие результаты ясно демонстрируют необходимость сохранения человека «в цикле» при использовании ИИ для задач образовательного оценивания. Лингвистическая связность и внешняя правдоподобность автоматически сгенерированных текстов не гарантируют их содержательной валидности, отсутствия предвзятости или соответствия нормативным требованиям. Поэтому наиболее перспективными выглядят гибридные модели, в которых ИИ используется как инструмент масштабируемой генерации и предварительного отбора материалов, тогда как психометрическая экспертиза, калибровка и интерпретация остаются в фокусе профессионального сообщества.

Результаты и обсуждение

Применение ИИ для автоскоринга графических и текстовых открытых ответов. Автоматическая проверка открытых ответов (текстовых, графических, аудио- и видеоформатов) становится одним из центральных направлений применения ИИ в оценивании. Традиционно задания с конструируемым ответом опирались на экспертов, что связано не только с высокими временными и финансовыми затратами, но и с неизбежной субъективностью: оценка подвержена эффектам гало, склонности к усреднению или крайним баллам, влиянию усталости и контекста проверки [18]. Технологии автоскоринга изначально задумывались как способ уменьшения вариативности, связанной с человеческим фактором, при сохранении экспертного контроля.

Автоматическая оценка открытых текстовых ответов опирается на методы обработки естественного языка (NLP), которые преобразуют текст в числовые представления - от частотных матриц TF-IDF до эмбедингов [18]. Эти представления используются для обучения моделей автоскоринга. Различают краткие открытые ответы и развернутые тексты, такие как эссе. В первом случае акцент делается на содержательной корректности, тогда как в эссе важны также структура, связность и стиль [19].

Исследования автоскоринга коротких ответов демонстрируют высокую вариативность точности — от 65% до 99% [20, 21]. Автоматическая оценка эссе остаётся более сложной задачей. Традиционные модели, использующие статистические, стилевые и контекстные признаки, достигают умеренной точности [22, 23]. Современные LLM лучше работают с «сырыми» текстами, но пока уступают специализированным моделям, особенно по критериям аргументации [24]. Таким образом, ИИ способен существенно поддерживать проверку открытых ответов, однако полностью автономный автоскоринг без участия экспертов остаётся недостижимым [25].

Параллельно развивается автоматическая оценка графических решений. Здесь ключевой инструмент - сверточные нейронные сети, способные извлекать структурные признаки из изображений. В исследованиях на материале заданий TIMSS-2019 показано, что CNN могут с очень высокой точностью классифицировать графические ответы (точность предсказания баллов в диапазоне 97–99 % и выше) [27, 28]. В клинической практике аналогичные подходы применяются для анализа рисунков (например, в тесте рисования часов) с целью диагностики когнитивных нарушений [29–31]. В исследованиях креативности CNN используются для автоматической оценки рисуночных тестов творческого мышления, демонстрируя приемлемую точность как при классификации по уровням, так и при предсказании непрерывных баллов [32–34].

Таким образом, ИИ уже сегодня позволяет существенно снизить нагрузку на проверяющих и повысить операционную эффективность оценивания графических решений и заданий открытого типа. При этом психометрическая повестка смещается от вопроса «можно ли автоматизировать проверку?» к вопросу «насколько валидны и интерпретируемы получаемые оценки и какова роль эксперта в таких гибридных системах?».

Оценивание ИИ-грамотности и сложных компетенций. Быстрое распространение генеративного ИИ делает очевидным, что традиционные модели образовательного оценивания оказываются недостаточными. Если ещё недавно ИИ воспринимался как вспомогательный инструмент для отдельных технических задач, то сегодня он формирует новую инфраструктуру образовательной среды, изменяя способы взаимодействия учащихся с информацией, а следовательно, и способы измерения их компетентностей. Наиболее остро это проявляется в области ИИ-грамотности - комплексного конструкта, включающего знания, навыки взаимодействие с ИИ, способность к критическому суждению и этическому использованию

алгоритмов.

Международные мониторинги, исследовательские группы и университетские инициативы уже предпринимают попытки осмыслить, что именно следует относить к ИИ-грамотности и как ее оценивать. Однако подавляющее большинство существующих инструментов опирается на самоотчёты и тесты знаний [35], которые фиксируют лишь декларативный уровень: «что человек знает об ИИ» или «как он воспринимает свои навыки» - но почти не говорят о том, как человек реально действует в ситуации взаимодействия с ИИ. Это приводит к признанному сегодня эпистемологическому тупику: восприятие собственной компетентности и реальное умение работать с ИИ не совпадают. Эмпирические данные подтверждают разрыв между интересом и действительным умением использовать ИИ, особенно у университетских студентов [36].

Именно в ответ на этот вызов нами была разработана первая в России и одна из первых в мире операциональная модель измерения AI Literacy, построенная на принципах Evidence-Centered Design (ECD) [37]. В отличие от статичных самоотчетных шкал, модель задаёт чёткую психометрически интерпретируемую структуру, которая объединяет когнитивные, процедурные, этические и метакогнитивные компоненты. В центре внимания - не знание принципов ИИ само по себе, а способность осознанно и ответственно использовать ИИ, понимать ограничения алгоритмов, замечать ошибки моделей, регулировать степень автономии ИИ в процессе решения задач.

Ключевым технологическим элементом новой парадигмы является переход от декларативного к поведенческому измерению, что радикально меняет саму архитектуру оценивания. В симуляционной тестовой среде участник вступает в реальное или моделируемое взаимодействие с LLM: формулирует задания, уточняет промпты, проверяет ответы ИИ, интерпретирует их в контексте предложенной ситуации. Такие задания строятся как комплексные сценарии, а каждое действие респондента становится наблюдаемым индикатором конструкции. Таким образом, измерение ИИ-грамотности становится операционализированным в терминах доказательств, что полностью соответствует логике ECD: конструкт - модель задания - модель доказательств - интерпретация.

Таким образом, оценивание ИИ-грамотности выходит за рамки традиционных инструментов и становится частью более широкой задачи - измерения сложных, многокомпонентных компетенций в цифровой среде. В этой логике ИИ используется не только как объект измерения, но и как компонент самой среды оценивания, создающий новые возможности для адаптивных сценариев, мультимодальных стимулов, фиксации процессуальных данных и анализа стратегий поведения.

Такое сочетание теоретической ясности (ECD) и технологических механизмов (LLM, симуляционные среды, мультимодальные данные) формирует новую психометрическую парадигму, где ключевой задачей становится не просто оценить знания, а реконструировать структуру компетентного действия в условиях ИИ-опосредованной деятельности.

Заключение

Интеграция искусственного интеллекта в образовательное оценивание приводит к формированию новой методологической рамки, в которой измерение становится частью единой цифровой экосистемы. Генерация заданий с использованием больших языковых моделей, мультимодальные форматы, автоматическая проверка открытых ответов и анализ процессуальных данных открывают возможности для более глубокого и точного оценивания сложных компетенций. Однако реализация этих возможностей требует устойчивой теоретической опоры: психометрика остается ключевой дисциплиной, обеспечивающей валидность, справедливость и интерпретируемость результатов, полученных с помощью

инструментов оценивания, созданных с использованием ИИ.

Важным условием развития этой области становится международное сотрудничество. Центр психометрики НИУ ВШЭ ведет активное взаимодействие с Шанхайским институтом искусственного интеллекта в образовании, Нанкинским университетом, Сианьским университетом Цзяотун–Ливерпуль (XJTLU) и Национальным центром тестирования Республики Казахстан. Эти партнерства позволяют сопоставлять модели в разных контекстах, проводить совместные исследования и укреплять научные основания для внедрения ИИ в оценивание.

Таким образом, дальнейшее развитие ИИ в образовательных измерениях возможно только при сочетании технологических инноваций с психометрической строгостью и международной научной кооперацией, обеспечивающей качество, надежность и этическую устойчивость создаваемых инструментов.

Список литературы

1. Brown T. B. et al. Language models are few-shot learners // arXiv preprint. - 2020. - arXiv:2005.14165. - <https://doi.org/10.48550/arXiv.2005.14165>
2. Devlin J. BERT: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint. - 2018. - arXiv:1810.04805.
3. Radford A. et al. Language models are unsupervised multitask learners // OpenAI blog. - 2019. - Vol. 1, no. 8. - P. 9.
4. Darwish S. M., Mohamed S. K. Automated essay evaluation based on fusion of fuzzy ontology and latent semantic analysis // The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019). - Springer International Publishing, 2020. - P. 566-575. - https://doi.org/10.1007/978-3-030-14118-9_57
5. Attali Y. et al. The interactive reading task: Transformer-based automatic item generation // Frontiers in Artificial Intelligence. - 2022. - Vol. 5. - P. 934124. - <https://doi.org/10.3389/frai.2022.903077>
6. Bezirhan U., von Davier M. Automated reading passage generation with OpenAI's large language model // Computers and Education: Artificial Intelligence. - 2023. - Vol. 5. - P. 100161. - <https://doi.org/10.1016/j.caeai.2023.100161>
7. Bulut O. et al. The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges // arXiv preprint. - 2024. - arXiv:2406.18900. <https://doi.org/10.48550/arXiv.2406.18900>
8. Sharma K., Giannakos M. Multimodal data capabilities for learning: What can multimodal data tell us about learning? // British Journal of Educational Technology. - 2020. - Vol. 51, no. 5. - P. 1450-1484. - <https://doi.org/10.1111/bjet.12993>
9. Banville F. et al. Validation of a sorting task implemented in the virtual multitasking task-2 and effect of aging // Human Interface and the Management of Information. Information in Applications and Services: 20th International Conference, HIMI 2018. Proceedings, Part II. - Springer International Publishing, 2018. - P. 41-54. - https://doi.org/10.1007/978-3-319-92046-7_4
10. Al-Ansi A. M. et al. Analyzing augmented reality (AR) and virtual reality (VR) recent development in education // Social Sciences & Humanities Open. - 2023. - Vol. 8, no. 1. - P. 100532. - <https://doi.org/10.1016/j.ssaho.2023.100532>
11. Al Balushi J. S. G. et al. Incorporating Artificial Intelligence Powered Immersive Realities to Improve Learning using Virtual Reality (VR) and Augmented Reality (AR) Technology // 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC). - IEEE, 2024. - P. 760-765. - <https://doi.org/10.1109/ICAAIC60222.2024.10575046>
12. Von Davier M. Automated item generation with recurrent neural networks // Psychometrika. - 2018. - Vol. 83, no. 4. - P. 847-857. - <https://doi.org/10.1007/s11336-018-9608-y>
13. Tan B. et al. A Review of Automatic Item Generation Techniques Leveraging Large Language Models // OSF Preprints. - 2024. - <https://doi.org/10.35542/osf.io/6d8tj>
14. Omopekunola M. O., Kardanova E. Y. Automatic generation of physics items with Large Language

Models (LLMs) // REID (Research and Evaluation in Education). - 2024. - Vol. 10, no. 2. - P. 168-185. - <https://doi.org/10.21831/reid.v10i2.76864>

15. Bozkurt A., Sharma R. C. Generative AI and prompt engineering: The art of whispering to let the genie out of the algorithmic world // Asian Journal of Distance Education. - 2023. - Vol. 18, no. 2. - P. i–vii. - <https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/749>

16. Liu P. et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing // ACM Computing Surveys. - 2023. - Vol. 55, no. 9. - P. 1-35. - <https://doi.org/10.1145/3560815>

17. Haley D. T. et al. Measuring improvement in latent semantic analysis-based marking systems: using a computer to mark questions about HTML // Proceedings of the 9th Australasian Conference on Computing Education (ACE 2007). - Ballarat: Australian Computer Society, 2007. - Vol. 66. - P. 35-42.

18. Jurafsky D., Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. - Online manuscript, 2024. - <https://web.stanford.edu/~jurafsky/slp3>

19. Higgins D. et al. Evaluating multiple aspects of coherence in student essays // Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. - 2004. - P. 185-192. - <https://aclanthology.org/N04-1024/>

20. Ariely M., Nazaretsky T., Alexandron G. Machine learning and Hebrew NLP for automated assessment of open-ended questions in biology // International Journal of Artificial Intelligence in Education. - 2023. - Vol. 33, no. 1. - P. 1-34. - <https://doi.org/10.1007/s40593-021-00283-x>

21. Raz T. et al. Automated Scoring of Open-Ended Question Complexity: A Large Language Model Approach // arXiv preprint. - 2024. - DOI:10.21203/rs.3.rs-3890828/v1

22. Burstein J. et al. A theoretical assessment ecosystem for a digital-first assessment - The Duolingo English Test // Duolingo Research Report DRR-21-04. - 2021. - <https://duolingo-papers.s3.amazonaws.com/reports/DRR-21-04.pdf>

23. Mathias S., Bhattacharyya P. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). - 2018.

24. Masikisiki B., Marivate V., Hlophe Y. Investigating the Efficacy of Large Language Models in Reflective Assessment Methods through Chain of Thought Prompting // Proceedings of the 4th African Human Computer Interaction Conference (Africhi '23). - 2023. - P. 44-49. - <https://doi.org/10.48550/arXiv.2310.00272>

25. Kostic M. et al. LLMs in Automated Essay Evaluation: A Case Study // Proceedings of the AAAI Symposium Series. - 2024. - Vol. 3, no. 1. - P. 143-147. - <https://doi.org/10.1609/aaais.v3i1.31193>

26. Schneider J. et al. Towards LLM-based Autograding for Short Textual Answers // arXiv preprint. - 2023. - <https://doi.org/10.48550/arXiv.2309.11508>

27. Von Davier M., Tyack L., Khorramdel L. Scoring graphical responses in TIMSS 2019 using artificial neural networks // Educational and Psychological Measurement. - 2023. - Vol. 83, no. 3. - P. 556-585. - <https://doi.org/10.1177/00131644221098021>

28. Tyack L., Khorramdel L., von Davier M. Using Convolutional Neural Networks to Automatically Score Eight TIMSS 2019 Graphical Response Items // Computers and Education: Artificial Intelligence. - 2024. - Vol. 6. - P. 100249. - <https://doi.org/10.1016/j.caeai.2024.100249>

29. Youn Y. C. et al. Use of the Clock Drawing Test and the Rey–Osterrieth Complex Figure Test-copy with convolutional neural networks to predict cognitive impairment // Alzheimer's Research & Therapy. - 2021. - Vol. 13. - P. 1-7. - <https://doi.org/10.1186/s13195-021-00821-8>

30. Ruengchaijatuporn N. et al. An explainable self-attention deep neural network for detecting mild cognitive impairment using multi-input digital drawing tasks // Alzheimer's Research & Therapy. - 2022. - Vol. 14, no. 1. - P. 111. - <https://doi.org/10.1186/s13195-022-01043-2>

31. Raksasat R. et al. Attentive pairwise interaction network for AI-assisted clock drawing test assessment of early visuospatial deficits // Scientific Reports. - 2023. - Vol. 13, no. 1. - P. 18113. - <https://doi.org/10.1038/s41598-023-44723-1>

32. Cropley D. H., Marrone R. L. Automated scoring of figural creativity using a convolutional neural network // Psychology of Aesthetics, Creativity, and the Arts. - 2022. - <https://doi.org/10.1037/aca0000510>

33. Panfilova A. S., Valueva E. A., Ilyin I. Y. The application of explainable artificial intelligence methods

to models for automatic creativity assessment // *Frontiers in Artificial Intelligence*. - 2024. - Vol. 7. - P. 1310518. - <https://doi.org/10.3389/frai.2024.1310518>

34. И. Л. Угланова, Е. С. Гельвер, С. В. Тарасов [и др.] Оценивание креативности на основе анализа изображений с помощью нейронных сетей // *Искусственный интеллект и принятие решений*. - 2021. - № 1. - С. 86-97. - <https://doi.org/10.14357/20718594210108>

35. Lintner T. A systematic review of AI literacy scales // *npj Science of Learning*. - 2024. - Vol. 9, no. 1. - P. 50. - <https://doi.org/10.1038/s41539-024-00264-4>

36. Иванова А. Е., Тарасова К. В., Талов Д. П. Между интересом и умением: как студенты воспринимают и применяют ИИ // *Высшее образование в России*. - 2025. - Т. 34, № 8-9. - С. 9-32. - <https://doi.org/10.31992/0869-3617-2025-34-8-9-32>

37. Mislevy R. J., Steinberg L. S., Almond R. G. On the structure of educational assessments // *Measurement: Interdisciplinary Research and Perspectives*. - 2003. - Vol. 1, no. 1. - P. 3-62. - https://doi.org/10.1207/S15366359MEA0101_02

К.В. Тарасова

ПСИХОМЕТРИКА ЖӘНЕ ЖАСАНДЫ ИНТЕЛЛЕКТ: БІЛІМ БЕРУДІ БАҒАЛАУДЫҢ ЖАҢА ПАРАДИГМАСЫ

Мақалада жасанды интеллектінің білім берудегі бағалау жүйелерін қалай өзгертетіні және осы өзгерістер психометриялық өлшеудің дәстүрлі негіздерін қайта қарауды неге талап ететіні талданады. Зерттеудің бастапқы мәселесі айқын: генеративті ИИ тест тапсырмаларын әзірлеуге, ашық жауаптарды автоматты бағалауға, бейімделмелі тестілеуге және мультимодальды бағалау орталарына жедел енгізіліп жатқанымен, бұл үдерістердің өлшеу логикасы мен психометриялық негіздемесі көбіне жеткілікті деңгейде айқындалмаған. Мақаланың мақсаты - қалыптасып келе жатқан осы ғылыми күн тәртібін тұжырымдамалық тұрғыдан сипаттау және ИИ-ді жай ғана қосымша техникалық құрал емес, конструктивті анықтаудан бастап нәтижелерді интерпретациялауға дейінгі бүкіл бағалау архитектурасын өзгертетін фактор ретінде көрсету. Мақала концептуалдық-аналитикалық сипатқа ие және үш негізгі бағыт төңірегінде құрылған: үлкен тілдік модельдер көмегімен тест тапсырмалары мен стимулдық материалдарды автоматты генерациялау; мәтіндік және графикалық ашық жауаптарды ИИ арқылы бағалау; сондай-ақ цифрлық ортадағы ИИ сауаттылығы мен басқа да күрделі құзыреттерді өлшеу. Талдау нәтижесінде ИИ мазмұнды ауқымды түрде генерациялау, мультимодальды тапсырмалар құрастыру, процессуалдық деректерді тіркеу және бағалау тиімділігін арттыру тұрғысынан зор мүмкіндік беретіні көрсетіледі. Алайда мұндай технологиялық артықшылықтардың өзі валидтілікке, әділдікке, интерпретациялануға және теориялық негізділікке автоматты түрде кепіл болмайды. Осыған байланысты мақалада ИИ бағалау үдерістерін кеңейтетін, бірақ психометриялық сараптама, валидация және адам бақылауы шешуші болып қалатын гибридік модель негізделеді. Сонымен бірге Evidence-Centered Design тәсілі өзіндік есепке сүйенетін декларативті бағалаудан ИИ-мен әрекеттесу жағдайындағы нақты мінез-құлықтық әрекеттерді бағалауға өтудің әдіснамалық негізі ретінде ұсынылады. Осылайша, мақала ИИ-делдалды бағалау әдіснамасының дамуына үлес қосып, технологиялық инновацияларды психометриялық қатаңдықпен ұштастырудың ғылыми негізделген бағытын ұсынады.

Түйін сөздер: психометрика, жасанды интеллект, білім беруді бағалау, тапсырмаларды автоматты түрде құрастыру.

K.V. Tarasova

PSYCHOMETRICS AND ARTIFICIAL INTELLIGENCE: A NEW PARADIGM OF EDUCATIONAL ASSESSMENT

This article examines how artificial intelligence is reshaping educational assessment and argues that these developments require a reconsideration of the psychometric foundations of measurement. The paper starts from a clear tension: while generative AI is rapidly being incorporated into item development, automated scoring,

adaptive testing, and multimodal assessment environments, the measurement logic underlying these practices often remains insufficiently articulated. The aim of the article is to conceptualize this emerging agenda and to demonstrate that AI should not be treated merely as an external technical tool, but as a force that reconfigures assessment architecture from construct definition to score interpretation. The paper is conceptual and analytical in nature and is organized around three core domains: automated generation of assessment stimuli and test items with large language models; AI-based scoring of open-ended textual and graphical responses; and the assessment of AI literacy and other complex competencies in digitally mediated environments. The analysis shows that AI substantially expands the possibilities of scalable content generation, multimodal task design, process-data capture, and operational efficiency. At the same time, these advantages do not in themselves ensure validity, fairness, interpretability, or theoretical coherence. The article therefore argues for a hybrid assessment model in which AI supports and extends assessment operations, while psychometric expertise, validation procedures, and human oversight remain indispensable. Particular attention is given to Evidence-Centered Design as a methodological framework for moving from declarative and self-report-based approaches toward behavior-based assessment of learners' actions in interaction with AI systems. In this way, the article contributes to the emerging methodology of AI-mediated assessment and offers a theoretically grounded perspective for integrating technological innovation with psychometric rigor in contemporary educational measurement.

Keywords: psychometrics, artificial intelligence, educational assessment, automatic item generation.

References

1. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2005.14165>
2. Devlin, J. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*. <https://arxiv.org/abs/1810.04805>
3. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.
4. Darwish, S. M., & Mohamed, S. K. (2020). Automated essay evaluation based on fusion of fuzzy ontology and latent semantic analysis. In *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019)* (pp. 566–575). Springer International Publishing. https://doi.org/10.1007/978-3-030-14118-9_57
5. Attali, Y., Davier, M. V., Moore, C., Steier, M., & Burstein, J. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5, Article 934124. <https://doi.org/10.3389/frai.2022.903077>
6. Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence*, 5, Article 100161. <https://doi.org/10.1016/j.caeai.2023.100161>
7. Bulut, O., Erhelfer, E. S., von Davier, M., & Kardanova, E. Y. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2406.18900>
8. Sharma, K., & Giannakos, M. (2020). Multimodal data capabilities for learning: What can multimodal data tell us about learning? *British Journal of Educational Technology*, 51(5), 1450–1484. <https://doi.org/10.1111/bjet.12993>
9. Banville, F., Nolin, P., & Labrosse, F. (2018). Validation of a sorting task implemented in the virtual multitasking task-2 and effect of aging. In *Human Interface and the Management of Information. Information in Applications and Services: 20th International Conference, HIMI 2018. Proceedings, Part II* (pp. 41–54). Springer International Publishing. https://doi.org/10.1007/978-3-319-92046-7_4
10. Al-Ansi, A. M., Jaboob, M., Akour, M., & Al-Ansi, A. (2023). Analyzing augmented reality (AR) and virtual reality (VR) recent development in education. *Social Sciences & Humanities Open*, 8(1), Article 100532. <https://doi.org/10.1016/j.ssaho.2023.100532>
11. Al Balushi, J. S. G., Al-Ansi, A. M., Akour, M., & Al-Ansi, A. (2024). Incorporating artificial intelligence powered immersive realities to improve learning using virtual reality (VR) and augmented reality

- (AR) technology. In *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)* (pp. 760–765). IEEE. <https://doi.org/10.1109/ICAAIC60222.2024.10575046>
12. Von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika*, 83(4), 847–857. <https://doi.org/10.1007/s11336-018-9608-y>
 13. Tan, B., Liu, L., & Liu, G. (2024). A review of automatic item generation techniques leveraging large language models. *OSF Preprints*. <https://doi.org/10.35542/osf.io/6d8tj>
 14. Omopekunola, M. O., & Kardanova, E. Y. (2024). Automatic generation of physics items with Large Language Models (LLMs). *REID (Research and Evaluation in Education)*, 10(2), 168–185. <https://doi.org/10.21831/reid.v10i2.76864>
 15. Bozkurt, A., & Sharma, R. C. (2023). Generative AI and prompt engineering: The art of whispering to let the genie out of the algorithmic world. *Asian Journal of Distance Education*, 18(2), i–vii. <https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/749>
 16. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1–35. <https://doi.org/10.1145/3560815>
 17. Haley, D. T., Thomas, P., de Roeck, A., & Petre, M. (2007). Measuring improvement in latent semantic analysis-based marking systems: using a computer to mark questions about HTML. In *Proceedings of the 9th Australasian Conference on Computing Education (ACE 2007)* (Vol. 66, pp. 35–42). Australian Computer Society.
 18. Jurafsky, D., & Martin, J. H. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Online manuscript. <https://web.stanford.edu/~jurafsky/slp3>
 19. Higgins, D., Burstein, J., Marcu, D., & Gentile, C. (2004). Evaluating multiple aspects of coherence in student essays. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004* (pp. 185–192). <https://aclanthology.org/N04-1024/>
 20. Ariely, M., Nazaretsky, T., & Alexandron, G. (2023). Machine learning and Hebrew NLP for automated assessment of open-ended questions in biology. *International Journal of Artificial Intelligence in Education*, 33(1), 1–34. <https://doi.org/10.1007/s40593-021-00283-x>
 21. Raz, T., Nazaretsky, T., & Alexandron, G. (2024). Automated scoring of open-ended question complexity: A large language model approach. *arXiv preprint*. <https://doi.org/10.21203/rs.3.rs-3890828/v1>
 22. Burstein, J., LaFlair, G. T., Kunnan, A. J., & von Davier, M. (2021). A theoretical assessment ecosystem for a digital-first assessment — The Duolingo English Test. *Duolingo Research Report DRR-21-04*. <https://duolingo-papers.s3.amazonaws.com/reports/DRR-21-04.pdf>
 23. Mathias, S., & Bhattacharyya, P. (2018). ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
 24. Masikisiki, B., Marivate, V., & Hlophe, Y. (2023). Investigating the efficacy of large language models in reflective assessment methods through chain of thought prompting. In *Proceedings of the 4th African Human Computer Interaction Conference (Africhi '23)* (pp. 44–49). <https://doi.org/10.48550/arXiv.2310.00272>
 25. Kostic, M., Stevanovic, A., & Jovanovic, J. (2024). LLMs in automated essay evaluation: A case study. *Proceedings of the AAAI Symposium Series*, 3(1), 143–147. <https://doi.org/10.1609/aaais.v3i1.31193>
 26. Schneider, J., Di Mitri, D., Limongelli, C., & Drachsler, H. (2023). Towards LLM-based autograding for short textual answers. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2309.11508>
 27. Von Davier, M., Tyack, L., & Khorramdel, L. (2023). Scoring graphical responses in TIMSS 2019 using artificial neural networks. *Educational and Psychological Measurement*, 83(3), 556–585. <https://doi.org/10.1177/00131644221098021>
 28. Tyack, L., Khorramdel, L., & von Davier, M. (2024). Using convolutional neural networks to automatically score eight TIMSS 2019 graphical response items. *Computers and Education: Artificial Intelligence*, 6, Article 100249. <https://doi.org/10.1016/j.caeai.2024.100249>
 29. Youn, Y. C., Kim, J., Ryu, J. H., Park, J. H., & Kim, H. J. (2021). Use of the clock drawing test and the Rey–Osterrieth complex figure test-copy with convolutional neural networks to predict cognitive impairment. *Alzheimer's Research & Therapy*, 13, 1–7. <https://doi.org/10.1186/s13195-021-00821-8>

30. Ruengchaijatuporn, N., Phutrakool, P., Hemrungronj, S., & Suchato, A. (2022). An explainable self-attention deep neural network for detecting mild cognitive impairment using multi-input digital drawing tasks. *Alzheimer's Research & Therapy*, 14(1), Article 111. <https://doi.org/10.1186/s13195-022-01043-2>
31. Raksasat, R., Phutrakool, P., Ruengchaijatuporn, N., Hemrungronj, S., & Suchato, A. (2023). Attentive pairwise interaction network for AI-assisted clock drawing test assessment of early visuospatial deficits. *Scientific Reports*, 13(1), Article 18113. <https://doi.org/10.1038/s41598-023-44723-1>
32. Copley, D. H., & Marrone, R. L. (2022). Automated scoring of figural creativity using a convolutional neural network. *Psychology of Aesthetics, Creativity, and the Arts*. <https://doi.org/10.1037/aca0000510>
33. Panfilova, A. S., Valueva, E. A., & Ilyin, I. Y. (2024). The application of explainable artificial intelligence methods to models for automatic creativity assessment. *Frontiers in Artificial Intelligence*, 7, Article 1310518. <https://doi.org/10.3389/frai.2024.1310518>
34. Uglanova, I. L., Gelfer, E. S., & Tarasov, S. V. (2021). Otsenivanie kreativnosti na osnove analiza izobrazheniy s pomoshchyu neyronnykh setey [Creativity assessment based on image analysis using neural networks]. *Iskusstvennyy intellekt i prinyatie resheniy* [Artificial Intelligence and Decision Making], (1), 86–97. <https://doi.org/10.14357/20718594210108>
35. Lintner, T. (2024). A systematic review of AI literacy scales. *npj Science of Learning*, 9(1), Article 50. <https://doi.org/10.1038/s41539-024-00264-4>
36. Ivanova, A. E., Tarasova, K. V., & Talov, D. P. (2025). Mezhdru interesom i umeniem: kak studenty vosprinimayut i primenyayut II [Between interest and skill: How students perceive and use AI]. *Vysyshee obrazovanie v Rossii* [Higher Education in Russia], 34(8-9), 9–32. <https://doi.org/10.31992/0869-3617-2025-34-8-9-9-32>
37. Mislavy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Авторлар туралы мәлімет:

Тарасова Ксения Вадимовна – педагогика ғылымдарының кандидаты, доцент, Білім институтының Психометрия және білімді бағалау орталығының директоры, Мәскеу, Ресей Федерациясы, ktarasova@hse.ru

Сведения об авторах:

Тарасова Ксения Вадимовна – кандидат педагогических наук, доцент, директор Центра психометрики и измерений в образовании Института образования, Москва, Российская Федерация, ktarasova@hse.ru

Information about authors:

Tarasova Kseniya Vadimovna – Candidate of Pedagogical Sciences, Director of the Center for Psychometrics and Educational Measurement, Institute of Education, Moscow, Russian Federation, ktarasova@hse.ru