

Торайғыров университетінің
ҒЫЛЫМИ ЖУРНАЛЫ

НАУЧНЫЙ ЖУРНАЛ
Торайғыров университета

ТОРАЙҒЫРОВ УНИВЕРСИТЕТІНІҢ ХАБАРШЫСЫ

Педагогикалық сериясы
1997 жылдан бастап шығады



ВЕСТНИК ТОРАЙҒЫРОВ УНИВЕРСИТЕТА

Педагогическая серия
Издается с 1997 года

ISSN 2710-2661

№ 4 (2025)

Павлодар

НАУЧНЫЙ ЖУРНАЛ
Торайгыров университета

Педагогическая серия
выходит 4 раза в год

СВИДЕТЕЛЬСТВО

о постановке на переучет периодического печатного издания,
информационного агентства и сетевого издания

№ KZ03VPY00029269

выдано

Министерством информации и коммуникаций
Республики Казахстан

Тематическая направленность

публикация материалов в области педагогики,
психологии и методики преподавания

Подписной индекс – 76137

<https://doi.org10.48081/UWEH8749>

Бас редакторы – главный редактор

Тулекова Г. М.

доктор PhD, профессор

Заместитель главного редактора

Жуматаева Е., *д.п.н., профессор*

Ответственный секретарь

Попандопуло А. С., *доктор PhD, профессор*

Редакция алкасы – Редакционная коллегия

Магауова А. С.,

д.п.н., профессор

Бекмагамбетова Р. К.,

д.п.н., профессор

Самекин А. С.,

доктор PhD, ассоц. профессор

Син Куэн Фунг Кеннет,

д.п.н., профессор (Китай)

Желвис Римантас,

д.п.н., к.псих.н., профессор (Литва)

Авагян А. В.,

д.п.н., ассоц. профессор (Армения)

Томас Чех,

д.п.н., доцент п.н. (Чешская Республика)

Шокубаева З. Ж.,

технический редактор

За достоверность материалов и рекламы ответственность несут авторы и рекламодатели

Редакция оставляет за собой право на отклонение материалов

При использовании материалов журнала ссылка на «Вестник Торайгыров университета» обязательна

ОҚЫТУДЫҢ ТЕОРИЯСЫ МЕН ӘДІСТЕМЕСІ

МРНТИ 14.35.07

<https://doi.org/10.48081/XKJT5386>

**Б. С. Абдрашилов¹, *Ш. Б. Алтыбаева², Л. Е. Шинетова³,
Г. Қ. Аширбек⁴, Д. Орлов⁵**

^{1,2,3,4,5}Национальный центр тестирования,

Республика Казахстан, г. Астана

¹ORCID: <https://orcid.org/0009-0002-1371-6211>

²ORCID: <https://orcid.org/0000-0003-0306-861X>

³ORCID: <https://orcid.org/0000-0003-4280-7999>

⁴ORCID: <https://orcid.org/0009-0006-6326-3088>

⁵ORCID: <https://orcid.org/0000-0002-7699-5345>

*e-mail: shugla@mail.ru

ВАЛИДАЦИЯ ТЕСТОВЫХ ЗАДАНИЙ ЕНТ: СОВРЕМЕННЫЕ ПСИХОМЕТРИЧЕСКИЕ ПОДХОДЫ

Настоящее исследование представляет собой психометрический анализ тестовых заданий по английскому языку, используемых в Едином национальном тестировании Республики Казахстан. База тестовых заданий была подготовлена экспертами, прошедшими обучение в рамках проекта Всемирного банка по модернизации образовательного оценивания. Для оценки надежности, валидности и дифференцирующей способности заданий в статье применена модель Раша. На основе данных, полученных в результате апробации тестов, был проведен анализ трудности заданий, их соответствия модели Раша и дифференциального функционирования (DIF). В апробации приняли участие 5000 выпускников школ со всего региона Казахстана, при этом для интерпретации результатов использовалась выборка, основанная на данных 399 участников. Основные результаты исследования показали высокую надежность инструментария (Item Reliability = 0,99), соответствие 92 % тестовых заданий установленным критериям валидности (Infit/Outfit MNSQ в пределах 0,5–1,5), а также выявление дифференциального функционирования заданий для учащихся с казахским и русским языками обучения.

Особое внимание уделено анализу дифференцирующей способности заданий различных типов, включая задания с одним правильным ответом, контекстные, с множественным выбором и на установление соответствий, что подтвердило способность теста эффективно оценивать уровень подготовки учащихся с соблюдением объективности.

Ключевые слова: ЕНТ, психометрический анализ, тест, модель Раиа, надежность, валидность, дифференцирующая способность, DIF.

Введение

Современная система образования Республики Казахстан играет ключевую роль в социально-экономическом развитии страны, формируя человеческий капитал как основной драйвер экономического роста. Как отметил Президент Касым-Жомарт Токаев, «в эпоху глобальной конкуренции именно качество образования и научного потенциала определяет место страны в мировом сообществе» [1]. В этом контексте особую актуальность приобретает совершенствование системы оценки образовательных достижений, обеспечивающей объективность и достоверность результатов. Единое национальное тестирование (далее – ЕНТ), являясь ключевым элементом национальной системы оценки качества образования, не только выполняет функцию своеобразного «фильтра» при переходе между образовательными уровнями, но и оказывает существенное влияние на формирование образовательных стратегий учащихся, определяя тем самым качество человеческого капитала страны в долгосрочной перспективе.

Масштабный проект Всемирного банка (далее – ВБ) «Модернизация среднего образования» (KZEMP/QCBS-11), реализованный в 2024 году по инициативе Министерства науки и высшего образования Республики Казахстан, представляет собой комплексную реформу оценочных процедур [2]. Данный проект направлен на разработку научно обоснованных методических руководств с учетом передового международного опыта в области педагогических измерений. В контексте этого, стандартизация ЕНТ, предусмотренная в Концепции развития высшего образования и науки Республики Казахстан на 2023–2029 годы [3], способствует его совершенствованию, обеспечивая более объективную и прозрачную систему оценивания, учитывающую как академические требования, так и общественные ожидания.

Настоящее исследование направлено на всесторонний анализ методических материалов, разработанных в рамках проекта ВБ, с акцентом на их теоретическое обоснование, структурные особенности и потенциал влияния на качество образовательного оценивания в Казахстане.

В рамках деятельности Национального центра тестирования (далее – Центр) были опубликованы научные статьи, освещающие результаты валидации и психометрического анализа разработанных тестовых заданий в ходе реализации проекта. В частности, работа Абрасилова Б. С. и др. [4] включает методический раздел, детализирующий ключевые аспекты исследования, проведенного в рамках проекта ВБ. В качестве основных инструментов исследования проведенного в рамках проекта ВБ использовались: анализ нормативно-правовой базы, изучение международного опыта, разработка и апробация тестовых заданий, а также применение статистико-математических методов обработки данных.

На первом этапе исследования осуществлялась подготовка разработчиков тестовых материалов, что являлось критически важным фактором обеспечения их качества. В рамках данного этапа был проведен специализированный обучающий курс общей продолжительностью 180 часов, в котором приняли участие 350 экспертов в области тестирования. Программа курса включала теоретическую подготовку по основам тестологии, изучение международных стандартов оценивания, а также практическую работу по созданию тестовых заданий с учетом предметной специфики и образовательных стандартов Республики Казахстан. Дополнительно был организован процесс экспертного анализа и редактирования тестов после их апробации. Обучение проводилось при координации Национального фонда образовательных исследований (NFER, Великобритания) с участием международных и национальных специалистов.

На основе проведенного исследования были разработаны методические рекомендации для специалистов, занимающихся созданием тестовых заданий.

Основным результатом проекта стало формулирование методологически обоснованной концепции модернизации системы педагогических измерений в контексте ЕНТ. Данная концепция реализовывалась по трем взаимосвязанным направлениям:

Проведение системного анализа международного опыта в области стандартизированного оценивания образовательных достижений.

Разработка и валидация базы тестовых заданий с применением современных психометрических моделей, включая теорию отклика на задание (Item Response Theory) и классическую теорию тестирования (Classical Test Theory).

Разработка институционального механизма подготовки высококвалифицированных специалистов в области тестологии, предполагающего сотрудничество с международными организациями, в том числе с консорциумом, в который входили Ginger SOFRECO (координация проекта), NFER (экспертиза в области образовательного оценивания), Capstan (лингвистическая адаптация) и TOO «BE Supply» (локальная координация).

На основе рекомендаций консорциума и с участием экспертов, прошедших обучение в рамках ВБ по модернизации образовательного оценивания, Центром была разработана база тестовых заданий для использования в ЕНТ.

Целью настоящего исследования является комплексный анализ качества разработанных тестовых заданий с учетом рекомендаций проекта ВБ и при участии специалистов, прошедших соответствующее обучение, с применением современных психометрических подходов, включая теорию отклика на задание (IRT). Результаты исследования направлены на обеспечение эмпирически обоснованных решений по дальнейшему совершенствованию системы оценки качества образования, основанных на принципах научной обоснованности, прозрачности и соответствия международным стандартам педагогических измерений.

Материалы и методы

Современные исследования в области педагогических измерений [5] подчеркивают важность научно обоснованного подхода к разработке тестовых материалов. Работы ведущих тестологов [6, 7] систематизируют ключевые требования к качеству тестов, особое внимание уделяя вопросам надежности и валидности.

Анализ международного опыта [8] показывает, что успешные системы стандартизированного тестирования (PISA, TIMSS) [9] базируются на спецификациях тестов, многоэтапной экспертизе заданий, комплексном психометрическом анализе, постоянном мониторинге качества.

Исследования [10; 11] подтверждают, что стандартизированные тесты, подобные ЕНТ, эффективны для прогнозирования академического успеха, особенно при комбинации с оценками школьной успеваемости, что снижает уровень отчислений на первом курсе. Как отмечают [12], оптимальные системы оценивания должны сочетать оценку знаний (достижения) и когнитивных навыков (способности), что повышает объективность отбора. Опыт SAT [13] и BMAT [14] показывает, что прозрачность и адаптация тестов под разные группы абитуриентов (включая учащихся из уязвимых слоев) критически важны для обеспечения равных возможностей, что актуально для модернизации ЕНТ. По данным Национальной ассоциации по консультированию при поступлении в колледж [15], стандартизированные тесты обеспечивают сопоставимость результатов и упрощают массовую обработку данных, но требуют строгой валидации заданий и регулярного обновления контента.

В исследовании применялся количественный описательный метод с использованием психометрического подхода на основе модели Раша (Rasch

model). Данный подход позволил оценить как уровень подготовленности испытуемых, так и характеристики тестовых заданий, включая их трудность и соответствие модели [16]. Для анализа данных использовалось программное обеспечение Winsteps, специально разработанное для реализации модели Раша. Оценивались такие параметры, как валидность, надежность, трудность заданий (в логитах), коэффициенты корреляции элементов с общим баллом, а также статистики соответствия модели (Infit и Outfit MNSQ и ZSTD).

В рамках исследования были использованы результаты апробации тестовых заданий. В общей апробации приняли участие 5000 выпускников школ из различных регионов Казахстана. В нашей работе рассматривается один вариант апробации, в котором приняли участие 399 выпускников общеобразовательных школ.

В качестве инструмента сбора данных использовался тест по английскому языку, состоящий из 40 тестовых заданий, из них 25 тестовых заданий с выбором одного правильного ответа, 5 тестовых заданий с выбором одного правильного ответа на основе одного контекста, 5 тестовых заданий с выбором одного или нескольких правильных ответов, 5 тестовых заданий на установление соответствия. Содержание заданий было направлено на проверку знаний и компетентности абитуриентов по английскому языку. Максимально возможный балл за выполнение всего теста составлял 50.

Формат ответов в тесте предполагал дихотомические задания («правильно» или «неправильно»), что соответствует базовой модели Раша. Однако задания с множественным выбором и на установление соответствия оценивались с использованием политомической шкалы, что могло потребовать применения расширенных версий модели Раша. При составлении матрицы для проведения анализа с использованием Winsteps задания были маркированы с учетом их типа: дихотомические и полиметрические.

Результаты и обсуждение

В ходе исследования проводилась оценка того, насколько эффективно каждая составляющая тестового задания выполняет свою функцию. Представленные в таблице 1 результаты психометрического анализа теста по английскому языку, включающего 40 заданий и апробированного на выборке из 399 выпускников средних школ с применением модели Раша, свидетельствуют о приемлемом уровне качества и надежности разработанного инструмента измерения.

Таблица 1 демонстрирует сводные психометрические характеристики теста. Средняя стандартная ошибка измерения трудности заданий, составляющая 0,12 логит, свидетельствует о высокой точности оценки трудности заданий на уровне всей выборки.

Таблица 1 – Сводные психометрические характеристики теста. Сводка по 40 измеренным (неэкстремальным) элементам

	Сум- марный балл	Коли- чество измере- ний	Изме- ре- ние	Стандарт- ная ошибка	Показатель соответствия		Показатель внешнего соответствия	
					MNSQ	ZSTD	MNSQ	ZSTD
Среднее значение	250,1	399,0	0,00	0,12	0,99	-0,16	1,04	0,03
Стандартная ошибка среднего	13,5	0,0	0,21	0,00	0,02	0,39	0,06	0,45
Стандартное отклонение популяции	84,1	0,0	1,29	0,03	0,14	2,43	0,35	2,80
Стандартное отклонение выборки	85,1	0,0	1,31	0,03	0,14	2,46	0,35	2,83
Максимальное значение	453,0	399,0	2,29	0,20	1,29	4,89	2,28	5,92
Минимальное значение	63,0	399,0	-2,69	0,07	0,71	-5,95	0,63	-5,56
Реля- тивная среднеквадратическая ошибка		0,12	Истинное стандартное отклонение	1,29	Коэффициент разделения	10,36	Коэффициент надежности элемента	0,99
Модельная среднеквадратическая ошибка измерения		0,12	Истинное стандартное отклонение	1,29	Коэффициент разделения	10,62	Коэффициент надежности элемента	0,99
Стандартная ошибка среднего значения элемента				0,21				
Корреляция между сырым баллом по элементу и его мерой = -0,78 (приблизительно из-за множественных группировок элементов)								

Средние значения статистик соответствия модели – Infit Mean-Square (MNSQ = 0,99) и Outfit Mean-Square (MNSQ = 1,04) – находятся вблизи оптимального значения 1,0. Согласно общепринятым критериям [26], значения MNSQ в диапазоне 0,5–1,5 указывают на приемлемое соответствие заданий модели Раша.

Надежность оценки трудности заданий (коэффициент альфа Кронбаха = 0,99) демонстрирует чрезвычайно высокую стабильность и воспроизводимость полученных оценок трудности для данной выборки респондентов.

Диапазон трудности заданий, варьирующийся от 2,69 до 2,29 логит, охватывает интервал в 4,98 логит. Это свидетельствует о способности измерительного инструмента эффективно дифференцировать респондентов с различным уровнем подготовленности.

Рассмотрим статистические характеристики заданий теста, чтобы оценить согласие тестового задания с моделью измерения.

В таблице 2 представлены результаты анализа отдельных тестовых заданий, полученные с использованием модели Раша. Психометрические характеристики каждого из 40 тестовых заданий включают уровень трудности, стандартную ошибку, показатели соответствия модели Раша и коэффициент корреляции.

Таблица 2 - Оценка параметров заданий

Балл	Кол-во	Изменение	Станд. ошибка	INFIT		OUTFIT		Точность бисер. корреляция		% правильных ответов		Задание	
				MNSQ	ZSTD	MNSQ	ZSTD	Набл.	Ожид.	Набл.	Ожид.		
1	2	3	4	5	6	7	8	9	10	11	12	13	14
	325	399	-1,46	0,14	1,01	0,16	1,26	1,79	0,32	0,37	84,2	82,5	A1
	196	399	0,37	0,11	1,22	4,89	1,31	5,01	0,22	0,42	60,8	68,2	A2
	116	399	1,40	0,12	1,25	4,34	1,58	5,26	0,13	0,37	67,3	73,4	A3
	99	399	1,65	0,12	1,15	2,32	1,42	3,37	0,18	0,35	74,9	76,4	A4
	286	399	-0,81	0,12	0,85	-2,63	0,78	-2,52	0,53	0,40	80,4	75,2	A5
	342	399	-1,83	0,15	0,94	-0,57	0,79	-1,31	0,41	0,35	86,2	86,3	A6
	285	399	-0,79	0,12	1,07	1,21	1,03	0,38	0,35	0,40	73,1	75,1	A7
	209	399	0,21	0,11	0,96	-0,95	0,91	-1,72	0,46	0,42	68,3	68,4	A8
	243	399	-0,22	0,11	0,93	-1,57	0,91	-1,50	0,48	0,42	75,6	70,3	A9
	297	399	-0,98	0,13	0,88	-1,90	0,78	-2,36	0,51	0,40	78,4	77,0	A10
	291	399	-0,88	0,12	1,03	0,54	1,09	0,94	0,36	0,40	76,6	76,0	A11
	134	399	1,15	0,12	1,14	2,83	1,27	3,14	0,24	0,38	66,8	71,0	A12
	260	399	-0,44	0,12	0,80	-4,18	0,76	-3,56	0,59	0,42	80,2	71,9	A13
	265	399	-0,51	0,12	0,98	-0,43	0,92	-1,06	0,44	0,41	72,9	72,4	A14
	198	399	0,34	0,11	1,03	0,65	1,01	0,11	0,40	0,42	64,3	68,2	A15
	351	399	-2,06	0,16	0,96	-0,36	0,79	-1,11	0,39	0,33	87,9	88,4	A16
	129	399	1,21	0,12	1,18	3,56	1,34	3,71	0,20	0,38	65,1	71,6	A17
	263	399	-0,48	0,12	0,86	-2,84	0,78	-3,20	0,54	0,41	74,9	72,2	A18
	159	399	0,82	0,11	1,11	2,37	1,29	4,04	0,28	0,40	68,8	69,0	A19

193	399	0,40	0,11	1,02	0,54	1,05	0,82	0,39	0,42	67,1	68,2	A20
272	399	1,12	0,08	1,01	0,14	1,01	0,16	0,50	0,50	56,5	56,6	B1
312	399	0,98	0,09	0,87	-2,15	0,88	-2,00	0,58	0,49	64,1	60,5	B2
453	399	-0,01	0,08	0,71	-5,30	0,70	-5,26	0,71	0,54	59,3	52,7	B3
254	399	1,22	0,08	0,87	-2,19	0,90	-1,21	0,56	0,50	61,1	57,5	B4
273	399	1,13	0,08	0,95	-0,86	0,94	-0,76	0,53	0,50	58,3	56,7	B5
168	399	1,63	0,08	1,27	3,30	1,56	3,46	0,32	0,46	65,8	64,9	B6
271	399	1,21	0,09	1,05	0,77	1,04	0,62	0,46	0,49	59,0	59,1	B7
171	399	1,86	0,09	1,04	0,52	1,17	1,62	0,39	0,44	66,1	66,2	B8
255	399	1,02	0,07	0,84	-2,69	0,83	-1,20	0,61	0,54	60,3	55,3	B9
118	399	1,98	0,09	1,24	2,40	1,94	4,04	0,24	0,40	71,6	74,2	B10
319	399	-1,35	0,14	1,13	1,65	1,36	2,56	0,24	0,38	81,2	81,2	C1
370	399	-2,69	0,20	0,88	-0,73	0,63	-1,47	0,42	0,29	92,7	93,0	C2
229	399	-0,04	0,11	0,97	-0,74	0,97	-0,45	0,45	0,42	70,6	69,3	C3
363	399	-2,42	0,19	0,94	-0,44	0,79	-0,89	0,37	0,31	91,2	91,3	C4
63	399	2,29	0,14	1,29	3,05	2,28	5,92	-0,07	0,29	83,2	84,2	C5
339	399	-1,76	0,15	0,89	-1,19	0,70	-1,98	0,46	0,35	86,9	85,6	C6
309	399	-1,17	0,13	0,93	-0,98	0,80	-1,84	0,46	0,39	79,6	79,2	C7
211	399	0,18	0,11	0,77	-5,95	0,72	-5,56	0,62	0,42	77,9	68,4	C8
316	399	-1,30	0,13	0,93	-1,01	0,75	-2,18	0,47	0,38	80,4	80,5	C9
297	399	-0,98	0,13	0,87	-2,08	0,77	-2,47	0,52	0,40	78,9	77,0	C10
С р . зн.	250,1 399,0	0,00	0,12	0,99	-0,16	1,04	0,03			73,0	72,4	
С т . откл	84,1 0,0	1,29	0,03	0,14	2,43	0,35	2,80			9,6	9,8	

В правой части таблицы записи расположены по порядку возрастания номера задания. Самым трудным заданием теста (обозначено красной линией) является задание C5 с уровнем трудности 2,29 логит. Из общего числа испытуемых (399 человек) с этим заданием справились 63 человека (что составляет 15,8 %). Самым легким заданием (обозначено зеленой линией) является задание C2 с уровнем трудности – 2,69 логит. С ним справились 370 человек (92,7 %).

Также по данным таблицы 2 видно, что показатели согласия (Infit/Outfit) находятся преимущественно в допустимом диапазоне. Значения Infit MNSQ всех заданий находятся в оптимальных пределах (0,5–1,5), что свидетельствует о согласованности большинства заданий с моделью.

В то же время Outfit MNSQ, более чувствительный к выбросам, указывает на проблемные задания: C5 (2,28) – крайне высокое значение, а также B10 (1,94) – оба требуют пересмотра или корректировки. Следует отметить, что задания с меткой «С» представляют собой контекстные задания, а задания с меткой «В» включают выбор нескольких правильных ответов или установление соответствий.

Корреляционный анализ показал, что задания с высокими значениями корреляции – B3 (0,71), A13 (0,59) и B2 (0,58) – хорошо соответствуют модели, тогда как задание C5 (–0,07) требует пересмотра из-за низкой корреляции [16].

При проведении психометрического анализа также был осуществлен анализ функционирования категорий заданий (Category Functioning), так как тест включает различные формы заданий, оценивание которых осуществляется по шкале 0, 1 и 2 балла. Этот анализ необходим для выявления эффективности каждого уровня сложности и корректности переходов между баллами, что позволяет более точно оценить дифференцирующие способности теста и его соответствие заданным критериям.

При проведении психометрического анализа был также рассмотрен анализ функционирования категорий заданий (Category Functioning). Тест включает задания, оцениваемые по шкале 0, 1 и 2 балла, что требует проверки эффективности каждого уровня сложности и корректности переходов между баллами для более точной оценки дифференцирующих способностей теста.

Пример анализа – задание B6 с несколькими правильными ответами, оцениваемое по шкале 0, 1 и 2 балла (таблица 3). Сложность задания составляет 1,63 logits, что указывает на средний уровень сложности, а для анализа использовались 399 ответов.

Анализ показал, что порог от 0 к 1 баллу (1,63 logits) соответствует уровню способности выше среднего, а порог от 1 к 2 баллам (3,09 logits)

требует высокого уровня подготовки. Fit-статистики категории 0 в пределах нормы, что подтверждает стабильность. Категория 1 имеет немного заниженные значения Infit и Outfit (0,93 и 0,82), что может свидетельствовать о ее предсказуемости. Категория 2 также имеет значения fit-статистик в пределах нормы.

Таблица 3 – Функционирование категорий заданий

Категория	Балл	Наблю- даемое ко л и чество	%	Наблю- даемое среднее	Ожида- е м о е среднее	INFIT MNSQ	O U T - FIT MNSQ	Порог Андри- ха	Мера
0	0	278	70	0,07	0,02	1,21	1,17	NONE	(0,17)
1	1	74	19	0,93	0,79	0,82	0,73	0,11	1,63
2	2	47	12	0,77*	1,31	1,45	2,38	-0,11	(3,09)

*НАБЛЮ-ДАЕМОЕ СРЕДНЕЕ - это среднее значение мер в категории. Это не оценка параметра
«NONE» означает, что порог не определен для этой категории (обычно для крайних категорий)

Теперь рассматриваем график функционирования категорий, который используется для анализа заданий с политомическими ответами (т.е. заданий с более чем двумя вариантами ответа), поскольку в формате ЕНТ рассматриваются различные формы заданий.

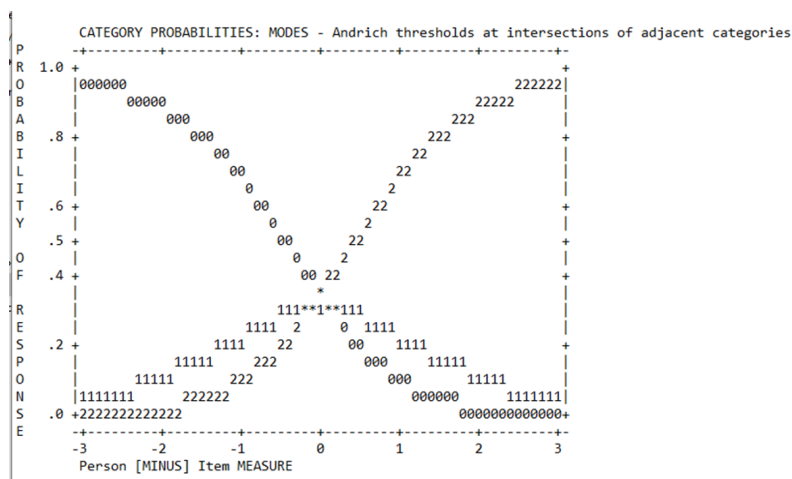


Рисунок 1 – Функционирование категорий заданий

По рисунку 1, анализируя функционирование категорий заданий, можно увидеть, что испытуемые с низкой подготовленностью преимущественно выбирают категорию «0» и с увеличением уровня подготовленности начинают выбирать категорию «1». Испытуемые со средней подготовленностью наиболее часто выбирают категорию «1», а с некоторой вероятностью – категорию «2». Испытуемые с высокой подготовленностью предпочитают категорию «2», при этом вероятность выбора категорий «0» и «1» значительно снижается.

Характеристическая кривая задания (Item Characteristic Curve, ICC) отображает вероятность правильного ответа в зависимости от уровня подготовленности испытуемого.

Кривая позволяет оценить, насколько задание эффективно различает испытуемых с разным уровнем знаний и насколько оно соответствует теоретической модели.

В данном исследовании представлены характеристические кривые для двух типов заданий: задания «А7», демонстрирующего хорошее соответствие модели, и задания «С5», которое, ввиду низкой корреляции (-0.07), требует переработки.

Характеристическая кривая задания А7 (Рисунок 2) демонстрирует типичную S-образную форму, что свидетельствует о корректной работе задания. Вероятность правильного ответа возрастает по мере повышения уровня подготовленности испытуемого, что указывает на его умеренную сложность.

Крутой наклон кривой в интервале от -2 до $+2$ на оси X отражает высокую различающую способность задания в этой зоне. Эмпирические данные (черные метки) в целом хорошо совпадают с модельной кривой (красная линия), подтверждая адекватность модели. Таким образом, задание А7 можно охарактеризовать как качественное, обладающее достаточной способностью дифференцировать испытуемых.

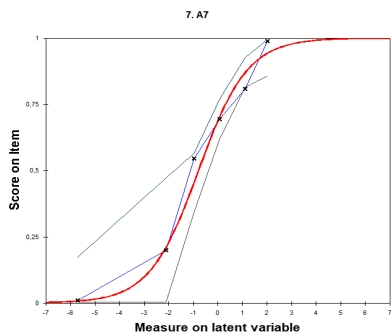


Рисунок 2 – Характеристические кривые заданий A7

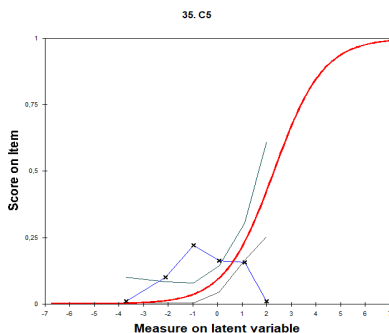


Рисунок 3 – Характеристические кривые заданий C5

На рисунке 3 видна меньшая крутизна кривой в диапазоне от -2 до +2, которая свидетельствует о сниженной различающей способности: задание менее дифференцирует испытуемых со средним уровнем знаний. Кроме того, эмпирические данные заметно отклоняются от предсказаний модели, особенно в диапазоне от -3 до +2, что указывает на несоответствие модели и необходимость доработки задания. В целом, задание C5 характеризуется повышенной сложностью, что требует его пересмотра.

Взаимосвязь между уровнем подготовленности испытуемых и трудностью тестовых заданий представлена на карте переменных «Испытуемые – Задание» (Рисунок 4). Информация, полученная из карты, показывает распределение уровней подготовленности респондентов (слева) и трудности тестовых заданий (справа) в диапазоне приблизительно от -3 до +3 логит.

Задания с меткой «А» (с одним правильным ответом) демонстрируют равномерное распределение по шкале сложности от -2 до +2 логит, что обеспечивает охват всех уровней сложности.

Политомические задания с метками «В» (с несколькими правильными ответами и задания на установление соответствия) сконцентрированы в диапазоне от +1 до +2 логит, что свидетельствует об их высокой сложности для большинства испытуемых.

Тест в основном содержит задания средней сложности, с включением нескольких очень сложных (C5) и нескольких легких заданий (A16, C4 и C2). Такое распределение позволяет более точно определить общий уровень сложности теста для обучающихся.

Таким образом, карта переменных демонстрирует наличие заданий разной степени трудности, что обеспечивает возможность адекватной оценки респондентов с различным уровнем подготовленности. Визуальный анализ подтверждает соответствие теста спецификации (testcenter.kz для подготовки к ЕНТ), предусматривающей три уровня сложности: базовый уровень – 50 %, средний уровень – 30 %, высокий уровень – 20 %.

В таблице 4 представлен анализ размерности, который показывает, что все задания теста измеряют одну и ту же способность.

Таблица 4 – Анализ размерности

	Собственное значение	Наблюдаемая дисперсия, %	Ожидаемая дисперсия, %
Общая необработанная дисперсия в наблюдениях	68,47	100,0	100,0
Необработанная дисперсия, объясненная мерами	28,47	41,6	40,9
Необработанная дисперсия, объясненная лицами	12,26	17,9	17,6
Необработанная дисперсия, объясненная элементами	16,21	23,7	23,3
Необработанная необъясненная дисперсия (итого)	40,00	58,4	59,1
Необъясненная дисперсия в 1-м контрасте	2,26	3,3	5,6
Необъясненная дисперсия во 2-м контрасте	2,11	3,1	5,3

Необъясненная дисперсия в 3-м контрасте	1,80	2,6	4,5
Необъясненная дисперсия в 4-м контрасте	1,57	2,3	3,9
Необъясненная дисперсия в 5-м контрасте	1,51	2,2	3,8

Анализ остаточной дисперсии показал, что в тесте нет ярко выраженных дополнительных измерений. Наибольшее значение контраста составляет 2,26, что ниже порогового уровня 3,0, при котором можно говорить о многомерности. Кроме того, доля необъясненной дисперсии по всем контрастам небольшая и ниже ожидаемой. Это означает, что задания в тесте в целом измеряют одну и ту же способность, то есть соответствуют одномерной модели.

Для анализа различий между группами использовался метод DIF (Differential Item Functioning), позволяющий выявить задания, функционирующие по-разному для испытуемых с одинаковым уровнем знаний, но из разных групп. В данном анализе сравнивались выпускники школ с казахским (синяя линия) и русским (красная линия) языками обучения; всего в исследовании приняли участие 399 человек, проходивших вариант теста, состоящий из 40 заданий по английскому языку.

Ось X представляет номера заданий, ось Y – величину DIF. Значение, близкое к нулю, указывает на отсутствие значимого дифференциала. Существенные отклонения в положительную или отрицательную сторону сигнализируют о потенциальной несправедливости задания для одной из групп. Наличие таких заданий требует дополнительной экспертизы на предмет лингвистической или культурной предвзятости [17].

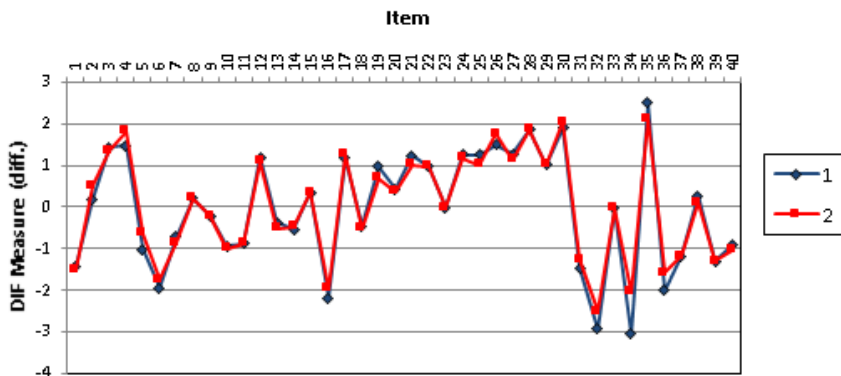


Рисунок 5 – DIF анализ

Анализ дифференциального функционирования заданий (DIF), проведенный между учащимися с казахским и русским языками обучения, показал, что некоторые задания теста по английскому языку демонстрируют незначительное различие в функционировании между группами. В частности, задание №35 (C5), относящееся к категории высокой сложности (уровень трудности – 2,29 логита), оказалось менее сложным для учащихся с казахским языком обучения по сравнению с учащимися с русским языком обучения. Этот результат требует дополнительного анализа для выявления возможных причин, таких как языковая или культурно-содержательная предвзятость задания.

Согласно графическому анализу, задания средней сложности - №3 (A3) с уровнем трудности 1,40 логита и №25 (B5) с уровнем 1,13 логита – оказались относительно легче для учащихся с русским языком обучения. Этот факт также требует дополнительного анализа. В то же время большинство заданий демонстрируют сопоставимый уровень сложности для обеих групп, что свидетельствует об общем равном уровне подготовленности участников тестирования. Это подтверждает, что выявленные случаи DIF имеют локальный характер и не оказывают существенного влияния на общую валидность теста. Рекомендуется провести дополнительную проверку и возможную корректировку задания №35.

В ходе исследования были проанализированы тестовые задания по английскому языку для ЕНТ. Результаты анализа подтверждают высокие показатели надежности и валидности, что свидетельствует о стабильности и точности используемых тестовых инструментов. В частности, параметры надежности, такие как Item Reliability = 0,99 и средняя стандартная ошибка измерения (0,12 логит), подтверждают высокую точность тестирования.

Значения Infit и Outfit MNSQ в пределах от 0,5 до 1,5 также свидетельствуют о соответствии модели Раша тестовых заданий, что подтверждает их валидность.

Анализ выявил, что задание №35 задание C5 демонстрирует отрицательную корреляцию с общим баллом теста ($r = -0,07$), что может свидетельствовать о его несоответствии модели и необходимости последующей корректировки. Характеристическая кривая данного задания также указывает на его повышенную сложность. Дополнительно, в результате DIF анализа было установлено, что задание C5 оказалось значительно легче для учащихся с казахским языком обучения, тогда как задания A3 и B5 – для учащихся с русским языком обучения. Эти данные подчеркивают важность учета культурных и языковых факторов при разработке тестовых заданий с целью обеспечения их объективности и справедливости для всех групп испытуемых.

На основании результатов анализа качества тестовых заданий рекомендовано пересмотреть задание № 35 с целью устранения выявленных проблем, связанных с низкими корреляционными показателями, повышенной сложностью и признаками дифференциального функционирования. В соответствии с данной рекомендацией задание было откорректировано: из формулировки устранена неоднозначность, допускавшая интерпретацию наличия двух правильных вариантов ответа.

Анализ распределения сложности тестовых заданий выявил положительную тенденцию: различные типы заданий, маркированные соответствующими метками, охватывают широкий диапазон уровней сложности. Особенно важно отметить, что задания с одним правильным ответом представлены во всем спектре сложности – от базового до высокого уровня, что позволяет эффективно дифференцировать уровень подготовки участников тестирования. Задания с меткой «B», включающие несколько правильных ответов и направленные на установление соответствий, органично дополняют структуру теста, обеспечивая разнообразие форматов и повышая доступность оценивания.

Таким образом, результаты проведенного исследования подтверждают высокий уровень качества тестовых заданий по английскому языку для ЕНТ, в том числе в аспекте распределения их сложности. В то же время выявлена необходимость корректировки отдельных заданий с целью повышения точности и справедливости оценивания. Важным направлением дальнейшей работы является учет культурно-языковых особенностей обучающихся, что будет способствовать повышению объективности, справедливости и валидности тестовой оценки.

Выводы

В рамках исследования была проведена комплексная психометрическая валидация тестовых заданий по английскому языку для ЕНТ в Республике Казахстан, подготовленных экспертами, прошедшими обучение в рамках проекта Всемирного банка. Применение модели Раша позволило оценить ключевые параметры качества тестовых заданий, включая их надежность, валидность и дифференцирующую способность.

Анализ данных, полученных в результате апробации тестовых материалов, выявил высокие показатели надежности инструментария (Item Reliability = 0,99) и соответствие 92 % заданий установленным критериям валидности (Infit/Outfit MNSQ в пределах 0,5-1,5). Эти результаты свидетельствуют о высоком качестве разработанных тестовых материалов и их пригодности для использования в системе ЕНТ.

Вместе с тем исследование показало наличие DIF ряда заданий для учащихся с казахским и русским языками обучения. Этот факт подчеркивает необходимость учета лингвистических и культурных факторов при разработке тестовых материалов, что обеспечит равные возможности для всех участников тестирования.

Полученные результаты имеют важное значение для специалистов в области педагогических измерений, разработчиков тестовых материалов и образовательных политиков, занимающихся модернизацией систем оценки образовательных достижений.

Реализация рекомендаций, основанных на результатах данного исследования, будет способствовать дальнейшему повышению качества и объективности системы ЕНТ, что, в свою очередь, окажет положительное влияние на качество образования в Республике Казахстан.

Список использованных источников

1 akorda.kz. Глава государства встретился с молодыми учеными в здании Национальной академии наук. [Электронный ресурс]. – URL: <https://www.akorda.kz/ru/glava-gosudarstva-vstretilsya-s-molodymi-uchenymi-v-zdani-nacionalnoy-akademii-nauk-314414>.

2 Финальный отчет за период 13 апреля – 9 июня 2024 г. в рамках контракта на оказание услуг № KZEMP/QCBS-11 «Улучшение системы оценивания ЕНТ». Консорциум «Ginger–SOFRECO» (Франция), National Foundation for Educational Research (Великобритания), ТОО «Be Supply» (Казахстан) и сApStAn (Бельгия). – 2024.

3 [adilet.zan.kz](https://adilet.zan.kz/rus/docs/P2400000471?utm_source). Концепция развития высшего образования и науки в Республике Казахстан на 2023-2029 годы. [Электронный ресурс]. – URL: https://adilet.zan.kz/rus/docs/P2400000471?utm_source.

4 Абдралилов, Б. С., Алтыбаева, Ш. Б., Шинетова, Л. Е. Результаты исследования системы ЕНТ, проведенного в рамках проекта Всемирного банка. Педагогические измерения. – №1 (2005). – С. 6–20. [Электронный ресурс]. URL: <https://testcenter-zhurnal.kz/index.php/sap/issue/view/1>.

5 American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. AERA. [Электронный ресурс]. – URL: <https://www.aera.net/publications/books/standards-for-educational-psychological-testing-2014-edition>.

6 Kline, T. Psychological testing: A practical approach to design and evaluation (2nd ed.). SAGE Publications. – 2023. <https://doi.org/1037/h0087036>.

7 Furr, R. M., & Bacharach, V. R. Psychometrics: An introduction (4th ed.). SAGE Publications. – 2021. [Электронный ресурс]. URL: https://www.researchgate.net/publication/230667917_Psychometrics_An_Introduction.

8 OECD. PISA 2022 technical report. OECD Publishing. <https://doi.org/10.1787/19963777>. – 2022.

9 Технический отчет TIMSS 2019, приложения 9.7 и 9.8. – 2019. [Электронный ресурс]. – URL: <https://timssandpirls.bc.edu/timss2019/methods/chapter-9.html>.

10 Makransky, G., Havmose, P., Vang, M. L., Andersen, T. E., & Nielsen, T. The predictive validity of using admissions testing and multiple mini-interviews in undergraduate university admissions. Higher Education Research & Development. – 2017. – 36(5). – P. 1003–1016. <https://doi.org/10.1080/07294360.2016.1263832>.

11 Yang, G. Are all admission sub-tests created equal? – Evidence from a National Key University in China. China Economic Review, 30. – 2014. – P. 600–617. <https://doi.org/10.1016/j.chieco.2013.12.002>.

12 Atkinson, R. C., & Geiser, S. Reflections on a century of College Admissions Tests. Educational Researcher. – 2009. – 38(9). – P. 665–676. – <https://doi.org/10.3102/0013189X09351981>.

13 Stringer, N. Aptitude tests versus school exams as selection tools for higher education and the case for assessing educational achievement in context. Research Papers in Education. – 2008. – 23(1). – P. 53–68. – <https://doi.org/10.1080/02671520701651771>.

14 McManus, I. C., Ferguson, E., Wakeford, R., Powis, D., & James, D. Predictive validity of the Biomedical Admissions Test: An evaluation

and case study. *Medical Teacher*. – 2011. – 33(1). – P. 53–57. <https://doi.org/10.3109/0142159X.2010.525267>.

15 National Association for College Admission Counseling. Report of the Commission on the Use of Standardized Tests in Undergraduate Admission. National Association for College Admission Counseling. – 2008.

16 **Bond, T. G., & Fox, C. M.** Applying the Rasch model fundamental measurement in the human sciences (3rd ed.). – 2015. DOI: 10.1111/j.1745-3984.2003.tb01103.

17 sciencedirect.com. Альфа-коэффициент Кронбаха. [Электронный ресурс]. – URL: <https://www.sciencedirect.com/topics/nursing-and-health-professions/cronbach-alpha-coefficient>.

References

1 akorda.kz. Glava gosudarstva vstretilsya s molodymi uchenyymi v zdanii Nacional'noy akademii nauk [The Head of State met with young scientists in the building of the National Academy of Sciences]. [Electronic resource]. – URL: <https://www.akorda.kz/ru/glava-gosudarstva-vstretilsya-s-molodymi-uchenyymi-v-zdanii-nacionalnoy-akademii-nauk-314414>.

2 Final'nyy otchet za period 13 aprelya – 9 iyunya 2024 g. v ramkah kontrakta na okazanie uslug № KZEMP/QCBS-11 «Uluchshenie sistemy ocenivaniya ENT». Konsorcium «Ginger–SOFRECO» (Franciya), National Foundation for Educational Research (Velikobritaniya), TOO «Be Supply» (Kazakhstan) i cApStAn (Bel'giya). – 2024.

3 adilet.zan.kz. koncepciya razvitiya vysshego obrazovaniya i nauki v Respublike Kazakhstan na 2023–2029 gody [The concept of higher education and science development in the Republic of Kazakhstan for 2023–2029.] [Electronic resource]. URL: https://adilet.zan.kz/rus/docs/P2400000471?utm_source.

4 **Abdrasilov, B. S., Altybaeva, Sh. B., Shinetova, L. E.** Rezul'taty issledovaniya sistemy ENT, provedennogo v ramkah proekta Vsemirnogo banka [The results of a study of the UNT system conducted as part of a World Bank project] Pedagogicheskie izmereniya. – 2005. – 1. – P. 6–20. [Electronic resource]. URL: <https://testcenter-zhurnal.kz/index.php/sap/issue/view/1>.

5 American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. Standards for educational and psychological testing. AERA. – 2014. [Electronic resource]. – URL: <https://www.acera.net/publications/books/standards-for-educational-psychological-testing-2014-edition>.

6 **Kline, T.** Psychological testing: A practical approach to design and evaluation (2nd ed.). // SAGE Publications. – 2023. – <https://doi.org/10.1037/h0087036>.

7 **Furr, R. M., & Bacharach, V. R.** Psychometrics: An introduction (4th ed.). SAGE Publications. – 2021. [Electronic resource]. – : https://www.researchgate.net/publication/230667917_Psychometrics_An_Introduction.

8 OECD. PISA 2022 technical report. OECD Publishing. – 2022. – <https://doi.org/10.1787/19963777>.

9 Tehnicheskij otchet TIMSS 2019, prilozheniya 9.7 i 9.8. [Electronic resource]. – URL: <https://timssandpirls.bc.edu/timss2019/methods/chapter-9.html>.

10 **Makransky, G., Havmose, P., Vang, M. L., Andersen, T. E., & Nielsen, T.** The predictive validity of using admissions testing and multiple mini-interviews in undergraduate university admissions. Higher Education Research & Development. – 2017. – 36(5). – P. 1003–1016. DOI: <https://doi.org/10.1080/07294360.2016.1263832>.

11 **Yang, G.** Are all admission sub-tests created equal? – Evidence from a National Key University in China. China Economic Review. – 2014. – 30. – P. 600–617. – <https://doi.org/10.1016/j.chieco.2013.12.002>.

12 **Atkinson, R. C., & Geiser, S.** Reflections on a century of College Admissions Tests. Educational Researcher. – 2009. – 38(9). – P. 665–676. – <https://doi.org/10.3102/0013189X09351981>.

13 **Stringer, N.** Aptitude tests versus school exams as selection tools for higher education and the case for assessing educational achievement in context. Research Papers in Education. – 2008. – 23(1). – P. 53–68. – <https://doi.org/10.1080/02671520701651771>.

14 **McManus, I. C., Ferguson, E., Wakeford, R., Powis, D., & James, D.** Predictive validity of the Biomedical Admissions Test: An evaluation and case study. Medical Teacher. – 2011. – 33(1). – P. 53–57. – <https://doi.org/10.3109/0142159X.2010.525267>.

15 National Association for College Admission Counseling. Report of the Commission on the Use of Standardized Tests in Undergraduate Admission. National Association for College Admission Counseling. – 2008.

16 **Bond, T. G., & Fox, C. M.** Applying the Rasch model fundamental measurement in the human sciences (3rd ed.). – 2015. <https://doi.org/10.1111/j.1745-3984.2003.tb01103>.

17 sciencedirect.com. Al'fa-koefficient Kronba [Alpha is the Cronbach's coefficient]. [Electronic resource]. – URL : <https://www.sciencedirect.com/topics/nursing-and-health-professions/cronbach-alpha-coefficient>.

Поступило в редакцию 26.05.25.

Поступило с исправлениями 30.10.25.

Принято в печать 25.11.25.

Б. С. Абдрасилов¹, *Ш. Б. Алтыбаева², Л. Е. Шинетова³,

Г. Қ. Аширбек⁴, Д. Орлов⁵

^{1,2,3,4,5}Ұлттық тестілеу орталығы,

Қазақстан Республикасы, Астана қ.

26.05.25. ж. баспаға түсті.

30.10.25 ж. түзетулерімен түсті.

25.11.25 ж. басып шығаруға қабылданды.

ҰБТ ТЕСТ ТАПСЫРМАЛАРЫН ВАЛИДАЦИЯЛАУ: ЗАМАНАУИ ПСИХОМЕТРИЯЛЫҚ ТӘСІЛДЕР

Зерттеу Қазақстан Республикасының Бірыңғай ұлттық тестілеуінде қолданылатын ағылшын тілі пәні бойынша тест тапсырмаларына психометриялық талдауын ұсынады. Тест тапсырмалары Дүниежүзілік банктің «Білім беруді бағалауды жаңғырту» жобасы аясында оқыған эксперттер құрастырған. Тапсырмалардың сенімділігін, жарамдылығын және дифференциялау қабілетін бағалау үшін Раш моделі қолданылды. Тестілеу нәтижелері бойынша алынған деректер негізінде тапсырмалардың қиындық деңгейі, Раш моделіне сәйкестігі және дифференциалдық жұмыс істеуі (DIF) талданды. Тестілеуге Қазақстанның әртүрлі аймақтарынан 5000 мектеп түлегі қатысты, ал нәтижелерді түсіндіру үшін 399 қатысушының деректері негізіндегі іріктеме пайдаланылды. Зерттеудің негізгі нәтижелері құралдың жоғары сенімділігін ($Item\ Reliability = 0,99$), тест тапсырмаларының 92 %-ның жарамдылық критерийлеріне сәйкестігін ($Inf\it/Out\it\it MNSQ\ 0,5-1,5$ шегінде) және қазақ және оқу тілдері бойынша оқушылар арасында тапсырмалардың дифференциалдық жұмыс істеуін анықтады. Бір дұрыс жауабы бар тапсырмалар, контекстік тапсырмалар, көптік таңдау және сәйкестендіру тапсырмалары сияқты әртүрлі типтегі тапсырмалардың дифференциялау қабілетіне ерекше назар аударылды, бұл тесттің оқушылардың дайындық деңгейін объективті бағалау қабілетін растады.

Кілтті сөздер: ҰБТ, психометриялық талдау, тест, Раш моделі, сенімділік, жарамдылық, дифференциялау қабілеті, DIF.

*B. S. Abdrasilov¹, *Sh. B. Altybayeva², L. E. Shinetova³,
G. K. Ashirbek⁴, D. Orlov⁵*

^{1,2,3,4,5}National Testing Center,

Republic of Kazakhstan, Astana

Received 26.05.25.

Received in revised form 30.10.25.

Accepted for publication 25.11.25.

VALIDATION OF UNT TEST TASKS: MODERN PSYCHOMETRIC APPROACHES

The present study is a psychometric analysis of English language test tasks used in the Unified National Testing of the Republic of Kazakhstan. The database of test tasks was prepared by experts who were trained as part of the World Bank's Educational assessment modernization project. The Rasch model is used in the article to assess the reliability, validity and differentiating ability of tasks. Based on the data obtained as a result of the testing, an analysis of the difficulty of the tasks, their compliance with the Rasch model and differential functioning (DIF) was carried out. 5,000 school graduates from all over the region of Kazakhstan took part in the testing, while a sample based on data from 399 participants was used to interpret the results. The main results of the study showed high reliability of the toolkit (Item Reliability = 0,99), compliance of 92 % of test tasks with established validity criteria (Infit/Outfit MNSQ in the range of 0,5–1,5), as well as identification of differential functioning of tasks for students with Kazakh and Russian languages of instruction. Special attention is paid to the analysis of the differentiating ability of tasks of various types, including tasks with one correct answer, contextual, multiple choice and matching tasks, which confirmed the ability of the test to effectively assess the level of students' education while maintaining objectivity.

Keywords: UNT, psychometric analysis, test, Rasch model, reliability, validity, differentiating ability, DIF.

Теруге 25.11.2025 ж. жіберілді. Басуға 30.12.2025 ж. қол қойылды.

Электронды баспа

9,42 Kb RAM

Шартты баспа табағы 31,59.

Таралымы 300 дана. Бағасы келісім бойынша.

Компьютерде беттеген З. Ж. Шокубаева

Корректорлар: А. Р. Омарова, Д. А. Қожас

Тапсырыс № 4485

Сдано в набор 21.11.2025 г. Подписано в печать 30.12.2025 г.

Электронное издание

9,42 Kb RAM

Усл.п.л. 32,63. Тираж 300 экз. Цена договорная.

Компьютерная верстка З. Ж. Шокубаева

Корректоры: А. Р. Омарова, Д. А. Қожас

Заказ № 4485

«Toraighyrov University» баспасынан басылып шығарылған

Торайғыров университеті

140008, Павлодар қ., Ломов к., 64, 137 каб.

«Toraighyrov University» баспасы

Торайғыров университеті

140008, Павлодар қ., Ломов к., 64, 137 каб.

8 (7182) 67-36-69

e-mail: kereku@tou.edu.kz

www.pedagogic-vestnik.tou.edu.kz