МРНТИ 14.01.11

Б.С. Абдрасилов[1], Ш.Б. Алтыбаева[2*], Л.Е. Шинетова[3]
[1,2,3]РГП на ПХВ «Национальный центр тестирования» МНВО РК, г. Астана, Республика Казахстан
*e-mail: shugla@mail.ru

[1]ORCID 0000-0002-2100-3564, [2]ORCID 0000-0003-0306-861X, [3]ORCID 0000-0003-4280-7999

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ СИСТЕМЫ ЕНТ, ПРОВЕДЕННОГО В РАМКАХ ПРОЕКТА ВСЕМИРНОГО БАНКА

Настоящее исследование представляет собой комплексный психометрический анализ системы Единого национального тестирования (ЕНТ) в Республике Казахстан, проведенный в рамках проекта Всемирного банка «Модернизация среднего образования» (далее - ВБ). Исследование основано на многоуровневом подходе, сочетающем три ключевых компонента: профессиональную подготовку 350 разработчиков тестов в соответствии с международными стандартами тестологии; создание и психометрическую валидацию базы из 8 848 тестовых заданий по 15 дисциплинам; применение комплементарных аналитических методов (CTT и модели Раша). Ключевые результаты исследования демонстрируют высокий уровень психометрической состоятельности разработанного инструментария: 90,7% заданий соответствуют критериям валидности и надежности. Выявлены статистически значимые различия в показателях надежности между предметными областями и языковыми версиями тестов, что требует дальнейшего изучения с учетом когнитивно-лингвистических факторов. Особое внимание уделено анализу дифференцирующей способности заданий и эффективности дистракторного аппарата. Результаты подтверждают корректность методологических подходов к разработке тестовых материалов и их способность достоверно дифференцировать испытуемых по уровню предметной подготовки. Практическая значимость исследования закладывают основы для проведения дальнейших исследований и выработки рекомендации в данной сфере. Исследование представляет интерес для специалистов в области психометрии, разработчиков тестовых материалов и образовательных политиков, занимающихся вопросами совершенствования систем оценивания образовательных достижений.

Ключевые слова: ЕНТ, психометрический анализ, валидность, надежность, модель Раша, дистракторный анализ, образовательное оценивание.

Введение

Единое национальное тестирование (далее - ЕНТ) на протяжении многих лет остается ключевым инструментом оценки академической подготовки выпускников средних образовательных учреждений Казахстана и играет решающую роль в процессе поступления в высшие учебные заведения. Данный экзамен является центральным компонентом образовательной системы страны, а его результаты выступают в качестве основного критерия для определения возможности поступления абитуриентов в университеты и получения ими образовательных грантов.

ЕНТ представляет собой комплексную систему оценивания, в рамках которой кандидаты сдают экзамены по трем обязательным предметам (грамотность чтения, математическая грамотность, история Казахстана) и двум предметам по выбору, в зависимости от специальности. Тестирование проводится на одном из трех языков и может быть сдано до пяти раз в год [1].

Формат и структура ЕНТ обусловлены необходимостью соответствия основным целям стандартизированного оценивания, включая объективное определение уровня подготовки поступающих, анализ их академических способностей и обеспечение справедливости в процессе поступления в высшие учебные заведения. Массовый характер тестирования требует оперативного подведения итогов и публикации результатов сразу после завершения экзаменационного процесса, что ставит перед системой оценивания задачу обеспечения надежности и валидности полученных данных.

Поскольку данные испытания являются экзаменом высокой ставки, они оказывают влияние на процесс обучения учащихся, их оценивание и усвоение учебного материала.

В этом контексте важным этапом стало завершение в 2021–2022 учебном году перехода среднего образования Казахстана на обновленное содержания, включая пересмотр стандартов, учебных программ, методов обучения и внедрение критериального оценивания. Как предусмотрено в Концепции развития дошкольного, среднего, технического и профессионального образования Республики Казахстан на 2023–2029 годы, дальнейшее совершенствование среднего образования направлено на разработку концептуальных и методологических основ, а также на периодический пересмотр учебных программ и учебных материалов с интервалом в 5–7 лет на основе анализа их эффективности [2].

Регулярный пересмотр образовательного контента рассматривается как необходимый инструмент для его соответствия современным требованиям и актуальным научным достижениям, что способствует повышению качества обучения. Это в свою очередь делает необходимым пересмотра методов и подходов к оцениванию поступающих для обеспечения справедливого доступа к высшему образованию.

Все эти изменения нацеливают нас на необходимость проведения исследований, ориентированных на оценку эффективности и объективности ЕНТ с позиций валидности используемых тестовых заданий. Одним из ключевых инструментов для достижения этих целей является психометрика – наука, занимающаяся измерением образовательных достижений и анализом качества тестовых заданий.

Психометрика определяет принципы разработки инструментов измерений в образовательных исследованиях, а также принципы работы с данными измерений. Она позволяет получить надежные и валидные данные, которые можно использовать для принятия управленческих решений или проверки исследовательских гипотез. Современная теория тестирования, включая модели Item Response Theory (IRT) и семейство моделей Раша, значительно расширила возможности анализа тестовых данных, обеспечивая объективность и надежность измерений [3].

В рамках проекта Всемирного банка «Модернизация среднего образования», с целью реализации задач по улучшению системы оценивания Единого национального тестирования в Республике Казахстан, были проведены комплексные мероприятия по совершенствованию ЕНТ, направленная на повышение качества тестовых заданий, надежности и объективности оценочных процедур [4].

Настоящая статья анализирует ключевые результаты проекта, включая разработку тестовых заданий, их апробацию и психометрический анализ, а также предлагает рекомендации по дальнейшему совершенствованию тестирования. Особое внимание уделяется анализу динамики изменения психометрических характеристик тестовых заданий ЕНТ. Актуальность исследования обусловлена возрастающим вниманием к проблемам стандартизированного тестирования, в частности, к качеству тестовых заданий, их справедливости, доступности и способности объективно оценивать академические способности поступающих.

Исследование основано на анализе материалов, подготовленных консорциумом Ginger-SOFRECO, Национальным фондом исследований в области образования (NFER), а также казахстанскими экспертами в рамках проекта ВБ.

Методы исследования

Ввиду отсутствия опубликованных материалов, всесторонне рассматривающих процесс разработки, апробации и психометрического анализа тестовых заданий в рамках ЕНТ, настоящее исследование опирается на комплексный анализ теоретических и эмпирических данных, а также на сотрудничество с международными и национальными экспертами в области образовательных измерений.

Исследовательская стратегия включала систематическое изучение научной литературы, нормативных документов, внутренних отчетов, а также первичный и вторичный анализ данных, предоставленных консорциумом Ginger-SOFRECO, NFER и специалистами

Национального центра тестирования (далее - НЦТ). Доступ к эмпирическим данным обеспечил возможность всесторонней оценки методологии формирования тестовых заданий, их апробации и последующего психометрического анализа.

В рамках исследования проведена всесторонняя эмпирическая работа, включавшая анализ нормативных документов и отчетности по разработке и применению тестовых заданий, а также количественный анализ данных апробации и статистических сводок их психометрических характеристик.

Дополнительно были проведены исследования экспертных заключений, направленных на определение параметров валидности и надежности тестового инструментария. Статистический анализ данных апробации был основан на исследовании частотных распределений, параметрических статистических показателей, а также применении методов корреляционного анализа [4].

Психометрический анализ тестовых заданий, реализованный посредством двух взаимодополняющих подходов:

- Классическая теория тестирования (CTT) – оценка сложности, дискриминационной способности и надежности заданий на основе традиционных статистических методов;
- Модель Раша – математическое моделирование параметров теста с целью обеспечения инвариантности измерений.

Особое внимание в рамках проекта уделялось профессиональной подготовке разработчиков тестовых заданий, поскольку квалификация специалистов непосредственно коррелирует с качеством инструментов оценивания. В ходе проекта была организована образовательная программа, в которой приняли участие 350 разработчиков, прошедших обучение по международным стандартам тестологии и психометрии. Учебная программа включала как теоретические курсы, так и практическую компоненту, связанную с конструированием и эмпирической валидацией тестовых заданий.

На данном этапе были разработаны и протестированы задания по 15 предметным областям. Их апробация проходила в два этапа: в первом этапе участвовали 94 883 учащихся, а во втором – 22 418 учащихся.

Таким образом, использование интегрированной методологической стратегии, сочетающей нормативно-правовой анализ, эмпирическую апробацию, статистическую обработку данных и психометрическую валидацию, обеспечило объективную оценку эффективности текущих механизмов тестирования и выработку рекомендаций для их дальнейшего совершенствования.

Результаты и обсуждение

В рамках механизма мониторинга качества, регламентированного методологическими принципами реализации Проекта, обучение специалистов по разработке тестовых заданий осуществлялось под эгидой NFER – ведущего независимого института в области образовательных исследований, аналитики и оценочных технологий Великобритании. Реализация данной образовательной программы, интегрирующей передовые международные и национальные методики, обеспечила профессиональную подготовку 350 специалистов, успешно завершивших четырехэтапное обучение, включавшее разработку, редакционную адаптацию и эмпирическую апробацию тестовых заданий. По итогам проекта было разработано 8 848 тестовых заданий, из них на казахском языке – 5 683, на русском языке – 2 462, на английском языке – 703.

В результате участники программы прошли комплексную 180-часовую подготовку по проектированию, валидации и пилотированию тестовых материалов, по завершении которой получили международные сертификаты, удостоверяющие их компетенции в области тестологических исследований и разработки стандартизированных оценочных инструментов.

Анализ апробации тестовых заданий ЕНТ. В рамках первой апробации тестовых заданий ЕНТ, реализованной в период февраль-март 2024 года, в соответствии с

методологией проекта было разработано 8 848 тестовых задания, из которых 4 897 подверглись процедуре пилотирования (более 65% из них на казахском языке). В апробации приняли участие 94 883 учащихся, которые проходили 209 тестовых вариантов. Однако вследствие неполного завершения тестирования отдельными респондентами и недостаточной численности выборки по ряду тестовых вариантов (менее 100 испытуемых) в итоговый анализ были включены данные 88 907 учащихся, охватывающие 156 тестовых вариантов.

Доля заданий с неудовлетворительными психометрическими характеристиками оказалась минимальной. В частности, в дисциплине «Основы права» (казахский и русский языки) проблемные задания не выявлены, тогда как наибольший уровень некорректных заданий зафиксирован по физике (казахский язык обучения) и составил 12,5%. Задания, не соответствующие установленным критериям психометрического качества, были исключены из дальнейшего тестирования, а разработчики осуществили корректировочные меры, основанные на статистическом анализе параметров сложности и эффективности дистракторов.

Анализ проведения второй апробации не менее 5000 тестовых заданий ЕНТ. В рамках настоящего исследования, проведенного в контексте проекта, выполнен анализ результатов второго пилотного тестирования ЕНТ, основанный на наборе эмпирических данных, предоставленных НЦТ в апреле 2024 года. В ходе анализа изучены ключевые психометрические характеристики тестовых заданий. Первичный набор данных включал 22 418 индивидуальных ответов респондентов, собранных в процессе апробации, а также 93 уникальных варианта теста. Важно отметить, что не все испытуемые завершили выполнение всех тестовых заданий, что обусловило вариативность выборки в зависимости от предметных областей и языковых версий теста.

В целях обеспечения всесторонней оценки качества тестовых заданий и выявления потенциальных дисбалансов в уровне сложности заданий был проведен комплексный психометрический анализ собранных данных. Особое внимание уделено оценке надежности теста, дифференцирующей способности заданий, а также соответствию эмпирических данных теоретическим моделям измерения. Комплексный характер анализа позволил установить объективные показатели качества тестового инструментария и выявить направления для его дальнейшего усовершенствования.

Методологическая основа психометрического анализа опиралась на современные подходы к оценке измерительных характеристик тестовых заданий. В частности, анализ проводился с использованием программной среды R [5], что обеспечило высокую точность расчетов и реплицируемость полученных результатов. В качестве первичного метода использовалась классическая теория тестирования (Classical Test Theory, CTT), реализация которой была осуществлена с помощью пакета R CTT [6]. В дополнение к этому для более глубокой интерпретации характеристик тестовых заданий применен анализ на основе модели Раша (Rasch Model), реализованный с использованием пакета R TAM [6]. Данный подход позволил провести детальную оценку сложности заданий, уровня их соответствия способности испытуемых и общей структуры теста.

Исходный набор данных включал все ответы учащихся, при этом неполные ответы были обработаны с учетом следующих критериев: задания со всеми пропущенными ответами исключались из анализа, тогда как частично пропущенные ответы оценивались как неверные.

Алгоритм обработки и оценивания данных тестирования. На этапе предварительной обработки эмпирических данных были сформированы матрицы, содержащие неоцененные и оцененные ответы испытуемых. Первичная матрица, представляющая необработанные данные, включала пять столбцов метаданных и 70 столбцов с ответами респондентов, зафиксированными в алфавитном представлении. В противоположность этому, матрица, подвергшаяся процессу оценивания, помимо аналогичных пяти столбцов метаданных,

содержала 75 столбцов, в которых ответы были конвертированы в бинарный формат (категории «верно»/«неверно»).

Процедура формирования оцененных и неоцененных матриц требует значительных вычислительных ресурсов, поскольку применение исключительно ручных методов обработки большого объема данных сопряжено с высокой вероятностью систематических и случайных ошибок. В связи с этим была разработана комплексная процедура автоматизированной обработки данных, реализованная посредством языка статистического программирования R. Конструирование алгоритма включало программирование итеративных функций, обеспечивающих верификацию и трансформацию ответов испытуемых для дальнейшего статистического анализа.

Разработанный алгоритм обладает модульной структурой и адаптируется к обработке тестовых массивов, содержащих 10, 20 и 40 заданий. Он реализует процедуру автоматического удаления данных на основании заданных критериев, осуществляет детекцию и разбиение комплексных заданий на подзадания, а также генерирует итоговые матрицы в соответствии с заданными параметрами. Особенность архитектуры алгоритма заключается в учете дифференцированного принципа разделения подзаданий для заданий, относящихся к диапазонам 31-35 и 36-40.

После завершения процедуры оценивания и сопоставления с соответствующими эталонными ключами, результирующие данные были подготовлены для проведения статистического анализа с использованием подходов классической тестовой теории (CTT) и модели Раша (Rasch Model).

Анализ по классической теории тестирования

В рамках анализа, основанного на классической теории тестирования (CTT), были получены ключевые статистические показатели, отражающие характеристики тестовых заданий и общие результаты тестируемых.

Одним из базовых параметров является размер выборки, который указывает общее число участников, чьи ответы были проанализированы. Для каждого задания вычисляется пропорция испытуемых, правильно ответивших на него (значение p), что позволяет оценить его сложность. *Значение p варьируется от 0 до 1*: чем выше показатель, тем легче задание, и наоборот [7]. Оптимальная структура теста предполагает наличие заданий разного уровня сложности, чтобы обеспечивать надежное измерение способностей всех тестируемых.

Дополнительно оцениваются показатели точности измерений. *Стандартная ошибка среднего общего балла (seM)* демонстрирует степень возможных отклонений выборочного среднего от истинного среднего значения, что важно для интерпретации надежности результатов. Также фиксируются минимальный и максимальный набранные баллы, а степень вариативности индивидуальных результатов отражается стандартным отклонением.

Одним из значимых показателей качества теста является *бисериальная корреляция (pbis),* которая показывает связь между результатами по отдельному заданию и общим итоговым баллом теста. Если данный коэффициент принимает отрицательное значение, задание может быть проблематичным и подлежит пересмотру или исключению. Низкие положительные значения также могут свидетельствовать о недостаточной дифференцирующей способности задания.

Общая надежность тестового варианта оценивается с использованием коэффициента внутренней согласованности (альфа Кронбаха). Этот показатель демонстрирует степень согласованности между заданиями теста и отражает его способность измерять единый конструкт. Для надежного теста коэффициент должен превышать 0,70, хотя в некоторых случаях допустимым считается значение выше 0,50 [8].

Дополнительно анализируется стандартная ошибка измерения (SEM), которая показывает уровень точности, с которым тест оценивает истинные способности тестируемых. Чем ниже SEM, тем выше вероятность того, что полученный результат соответствует реальному уровню знаний испытуемого [9].

Также в анализ включены частотные характеристики распределения ответов на задания, позволяющие оценить, насколько эффективно тестируемые распознают правильный вариант ответа и какие из альтернативных вариантов (дистракторов) оказываются наиболее привлекательными.

Данные показатели позволяют объективно оценивать качество тестовых материалов, их дифференцирующую способность и надежность, что способствует совершенствованию процесса тестирования и повышению его измерительной валидности.

Анализ Раша. Оценка сложности заданий осуществлялась преимущественно посредством анализа p-значений, дополнительно дополняемого применением модели Раша. Для измерения параметров тестовых заданий использовалась дихотомическая модель Раша, основанная на вероятностном подходе к оценке успешного выполнения задания в зависимости от уровня компетентности испытуемого и сложности самого задания. В рамках данного анализа вероятность корректного ответа при совпадении уровня способностей испытуемого и сложности задания принималась равной 0,50.

Статистическая обработка осуществлялась с использованием пакета R TAM [6], применяя методы предельного максимального правдоподобия (MME) и оценки ожидаемого апостериорного значения (EAP) для определения латентных параметров испытуемых. В ходе анализа задания, чья сложность превышала максимальный уровень компетентности испытуемых более чем на установленное пороговое значение (0,5 логитов), маркировались специальным обозначением, указывающим на их чрезмерную сложность. Аналогичным образом, задания, уровень сложности которых существенно уступал минимальному значению компетентности испытуемых, отмечались как слишком простые.

Для комплексной оценки качества вариантов теста, сформированных с учетом специфики предмета и языка тестирования, производился анализ распределения сложности заданий по тестовым формам. Например, в рамках второго этапа пилотного тестирования по дисциплине "Биология" на казахском языке оценивались 5 тестовых вариантов, включавших по 45 отдельных заданий, что в совокупности составляло 225 единиц тестового материала. Из общего числа заданий 95,1% демонстрировали положительную точечную бисериальную корреляцию, что свидетельствовало о соответствии минимальным критериям качества [4].

Дополнительно для каждого сочетания предмета и языка рассчитывалось среднее арифметическое значение альфа-коэффициентов по всем вариантам теста, выступавшее в качестве интегрального индикатора качества заданий. Итоговый анализ позволял оценить общее количество тестовых позиций, удовлетворяющих установленным критериям, что являлось ключевым параметром в определении эффективности проведенного пилотного исследования.

Оценка качества тестовых заданий и вариантов тестов. В рамках проведенного анализа были рассмотрены тестовые варианты, различающиеся по предметной области и языку тестирования. Оценка качества осуществлялась с применением количественных психометрических показателей, включающих: (1) коэффициент внутренней согласованности α (Кронбаха), отражающий степень однородности тестовых вариантов, (2) долю тестовых заданий, продемонстрировавших низкие статистические характеристики, а также (3) совокупное количество заданий, удовлетворяющих минимальным психометрическим требованиям.

Таблица 1 демонстрирует итоговое количество тестовых заданий и ответов испытуемых, включенных в аналитическую выборку, с дифференциацией по предметным областям и языкам тестирования. Качество тестовых вариантов для каждой предметно-языковой комбинации оценивалось посредством расчета среднего значения коэффициента α (Кронбаха), доли тестовых заданий с низкими статистическими характеристиками и совокупного количества заданий, обладающих приемлемыми психометрическими параметрами.

Следует отметить, что в аналитическую выборку включены исключительно те тестовые варианты, по которым было зарегистрировано участие не менее 100 испытуемых. Два

предмета – казахский и русский языки – не удовлетворяли данному критерию, вследствие чего оценка качества тестовых заданий по ним не проводилась.

Таблица 1 - Психометрические характеристики тестовых вариантов по предметам и языкам

Предмет	Язык	α (Надеж-ность)	% слабых заданий	Всего заданий	Достаточные задания	Учащиеся	Варианты теста
Биология	Казахский	0,83	4,9%	214	225	956	5
Биология	Русский	0,87	5,6%	85	90	201	1 (*1)
Химия	Казахский	0,77	11,1%	200	225	638	5
Химия	Русский	0,74	18,2%	108	135	103	0 (3**)
Английский язык	Казахский	0,83	6,7%	208	225	379	0 (*5)
Английский язык	Русский	0,77	18,1%	181	225	135	0 (**5)
География	Казахский	0,8	10,4%	120	135	733	3
География	Русский	0,81	10,0%	81	90	224	2
История Казахстана	Казахский	0,74	1,0%	99	100	3 680	5
История Казахстана	Русский	0,67	6,0%	94	100	868	5
Всемирная история	Казахский	0,87	2,7%	219	225	697	5
Всемирная история	Русский	0,75	11,9%	119	135	179	0 (*3)
Информатика	Казахский	0,84	8,1%	124	135	374	3
Информатика	Русский	0,72	20,3%	106	135	98	0 (**3)
Казахский язык	Казахский	0,91	2,2%	88	90	105	0 (*2)
Казахская литература	Казахский	0,89	5,6%	85	90	103	0 (*2)
Основы права	Казахский	0,88	0,0%	42	45	262	1
Основы права	Русский	0,87	2,2%	45	45	67	0 (*1)
Математическая грамотность	Казахский	0,53	2,0%	49	50	3 217	5
Математическая грамотность	Русский	0,57	2,5%	39	40	784	4
Математика	Казахский	0,82	4,5%	128	135	1 324	3
Математика	Русский	0,86	3,4%	86	90	281	2
Физика	Казахский	0,72	13,8%	193	225	715	5
Физика	Русский	0,82	11,3%	118	135	116	0 (**3)
Читательская грамотность	Казахский	0,6	4,0%	48	50	3 717	5
Читательская грамотность	Русский	0,39	10,0%	36	40	885	4
Русский язык	Русский	0,75	11,1%	40	45	38	0 (**1)
Русская литература	Русский	0,82	6,7%	42	45	42	0 (**1)

В рамках данного исследования проанализировано 3 305 тестовых заданий, из которых 2 997 продемонстрировали удовлетворительные психометрические характеристики, а именно

положительные значения коэффициента pbis. Полученные данные подверглись детализированному анализу, что позволило выявить, что 90,7% заданий соответствуют установленным критериям качества, что подтверждает их высокую надежность теста.

В рамках настоящего пилотного исследования была проанализирована выборка, включающая 22 418 испытуемых и 93 варианта тестов. Полученные результаты свидетельствуют о варьировании показателей качества тестовых вариантов по различным предметным областям и языкам тестирования в диапазоне от неудовлетворительного до высокого уровня. Например, коэффициент надёжности теста на читательскую грамотность (русский язык) составил 0,39, тогда как аналогичный показатель для теста по казахскому языку (для казахоязычных испытуемых) достиг 0,91, что соответствует высокому уровню внутренней согласованности.

Следует учитывать, что коэффициент α Кронбаха имеет ограничения, связанные с его зависимостью от числа тестовых заданий [10], в связи с чем низкие значения данного коэффициента не всегда однозначно свидетельствуют о недостаточном качестве теста.

Доля тестовых заданий с неудовлетворительными психометрическими характеристиками в большинстве вариантов тестов оставалась на низком уровне, что указывает на их соответствие установленным критериям качества.

Дистракторный анализ представляет собой исследование качества неверных, но правдоподобных ответов в тестовых заданиях закрытой формы. Дистракторы выполняют функцию когнитивных ловушек, ориентированных на дифференциацию уровня подготовленности испытуемых: они должны быть привлекательными для респондентов с низким уровнем знаний, но не вводить в заблуждение компетентных участников тестирования. Конструирование дистракторов направлено на повышение валидности тестовых заданий и обеспечение их дискриминативной способности. Эффективность дистракторов оценивается посредством анализа распределения выборов среди испытуемых, позволяя выявить их дидактическую значимость и диагностическую ценность в рамках тестового инструментария [11].

Дистракторный анализ представляет собой теоретико-эмпирическое исследование, направленное на оценку корректности и эффективности подбора вариантов ответов в тестовых заданиях. Данный анализ является неотъемлемым элементом разработки тестов и педагогических измерений. Его теоретическая составляющая охватывает вопросы содержательной и формальной структуры тестовых заданий, в то время как эмпирическая часть основана на проведении пилотного тестирования для выявления характеристик как самих заданий, так и предложенных вариантов ответов.

Рассмотрим пример, иллюстрирующий принцип функционирования дистракторов. Если тестовое задание с четырьмя вариантами ответов, один из которых является верным, выполняется группой из 300 испытуемых, и 150 из них выбирают правильный ответ, то оставшиеся 150 участников должны распределиться между тремя дистракторами приблизительно равномерно, то есть по 50 человек на каждый дистрактор. Такое распределение позволяет оценить эффективность дистракторов и их способность дифференцировать уровень подготовленности испытуемых.

Таблица 2 - Частотный анализ дистракторов

Вариант ответа	Число выборов	Процент выборов	Корреляция
A	21	7%	-0,178
B	89	30%	-0,086
C	59	20%	-0,325
D*	128	43%	0,43

Дистракторы, которые выбирают менее 5% неверно выполнивших задание тестируемых, считаются неработающими и должны быть изменены или удалены из теста [12].

Глубокий анализ правдоподобности дистракторов включает расчет коэффициентов корреляции для каждого дистрактора в тестовых заданиях, что позволяет оценить их дифференцирующую способность и выявить потенциальные недостатки в их формулировке.

Рассмотрим пример дистракторного анализа на основе конкретного тестового задания с одним правильным ответом из четырех предложенных вариантов, где верный ответ соответствует опции «D». Оценивание осуществлялось по дихотомической шкале: испытуемый получал 1 балл при выборе исключительно правильного ответа и 0 баллов в противном случае. В данном случае 128 участников тестирования ответили верно, получив 1 балл, тогда как 169 испытуемых выбрали неверные варианты и не получили баллов. В таблице 3 приведены данные, характеризующие распределение выборов по вариантам ответа.

Таблица 3 - Данные по вариантам задания с выбором 1 варианта из 4

Варианты ответов	**A**	**B**	**C**	**D***
Эмпирическая частота выбора дистракторов (%)	21 (7%)	89 (30%)	59 (20%)	128 (43%)

Следует отметить, что правильный ответ демонстрирует наибольшую привлекательность, а дистракторы, несмотря на их неравномерное распределение, выполняют свою функцию. Для углубленного анализа эффективности дистракторов целесообразно рассчитать коэффициенты корреляции между выбором правильного ответа, дистракторами и суммарным тестовым баллом.

Оптимальное функционирование дистрактора предполагает, что испытуемые с высоким уровнем подготовки избегают его выбора в качестве ответа, что отражается в отрицательном значении коэффициента корреляции (предпочтительно ниже -0,1). В свою очередь, для корректного ответа коэффициент корреляции должен быть положительным и превышать пороговое значение 0,3, что свидетельствует о его адекватной дифференцирующей способности. В таблице 4 представлены расчетные значения корреляционных коэффициентов для анализируемого тестового задания.

Таблица 4 - Значения корреляции

Варианты ответов	**A**	**B**	**C**	**D***
Коэффициенты корреляции	-0,178	-0,086	-0,325	0,43

Анализ показал, что данное тестовое задание обладает высокой дифференцирующей способностью: участники с высоким уровнем подготовки преимущественно выбирают правильный ответ, тогда как менее подготовленные испытуемые распределяют свои выборы среди дистракторов. Это свидетельствует о корректном функционировании задания и его способности эффективно дифференцировать уровень знаний тестируемых. Ниже представлено задание, характеристики которого были рассмотрены.

Два автомобиля движутся навстречу друг другу. Скорость первого автомобиля относительно Земли равна 30 км/ч, скорость второго - 40 км/ч. Определите модуль скорости первого автомобиля в системе отсчёта второго автомобиля.

А) 5 км/ч

B) 10 км/ч
C) 35 км/ч
D) 70 км/ч

Рассматриваемое тестовое задание относится к уровню А и тематическому разделу «Механика». Оно разработано в соответствии с учебной целью типовой учебной программы для 10 класса - «10.2.1.4: применение классического закона сложения скоростей и перемещений при решении задач» [13]. Дистракторы сформированы с учетом наиболее распространенных ошибок, характерных для испытуемых, что способствует повышению валидности задания и его диагностической ценности.

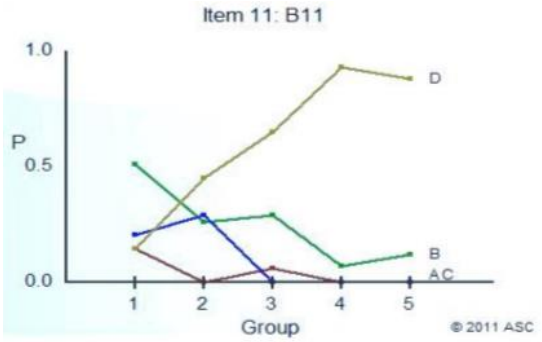

Рисунок 1 - Графическое представление результатов дистракторного анализа

Анализ графического представления данных, полученного с использованием программного обеспечения Iteman, выявляет четкие закономерности в распределении ответов испытуемых. Цветовая дифференциация линий на графике отражает характер выбора вариантов ответов среди пяти групп тестируемых, стратифицированных по уровню их суммарных результатов.

Наблюдается устойчивая корреляция между успешностью выполнения теста и частотой выбора верных ответов, что подтверждает эффективную дифференцирующую способность тестовых материалов. Правильные варианты демонстрируют выраженную тенденцию к возрастанию частоты выбора по мере увеличения общего балла испытуемых, что свидетельствует о высокой дискриминационной валидности заданий. В то же время дистракторы показывают обратную динамику - их выбор закономерно уменьшается у более подготовленных участников тестирования, что соответствует критериям качественно разработанного тестового инструментария.

Полученные результаты визуализации служат весомым подтверждением соответствия тестовых материалов современным психометрическим требованиям. Выявленные закономерности подкрепляют обоснованность методологических подходов к разработке и валидации заданий, а также убедительно свидетельствуют об их способности достоверно дифференцировать испытуемых по уровню предметной подготовки. Данные графического анализа представляют собой значимый аргумент в пользу качества созданного тестового инструментария и эффективности примененных процедур его психометрической проверки.

Заключение

В рамках реализации проекта Всемирного банка «Модернизация среднего образования» было проведено комплексное исследование, направленное на совершенствование системы Единого национального тестирования в Республике Казахстан. Основные достижения проекта заключаются в реализации двух взаимосвязанных стратегических задач: профессиональной подготовки специалистов в области педагогических измерений и разработки научно обоснованного инструментария для оценки образовательных достижений.

Ключевым результатом проекта стало формирование корпуса из 350 квалифицированных разработчиков тестовых материалов, прошедших специализированную подготовку по современным методикам тестологии. Разработанная 180-часовая программа профессионального развития позволила создать сообщество экспертов, обладающих компетенциями в области конструирования и валидации тестовых заданий.

В методологическом аспекте значимым достижением стало создание обширной базы тестовых материалов, включающей 8 848 валидированных заданий по 15 учебным предметам. Многоэтапная процедура апробации с участием репрезентативной выборки испытуемых (N=22 418) и применение комплементарных психометрических подходов (классическая теория тестирования и модель Раша) обеспечили высокую степень надежности разработанного инструментария. Психометрический анализ подтвердил соответствие 90,7% заданий установленным критериям качества, при этом были выявлены значимые различия в показателях надежности между различными предметными областями и языковыми версиями тестов.

Перспективные направления развития системы оценивания включают внедрение адаптивных алгоритмов тестирования, совершенствование методологии разработки дистракторов и создание многоуровневой системы мониторинга качества тестовых материалов. Реализация проекта заложила методологические основы для дальнейшей модернизации национальной системы оценивания в соответствии с международными стандартами, сохраняя при этом учет специфики отечественной образовательной системы.

Полученные результаты имеют существенное значение для образовательной политики и могут быть использованы при совершенствовании нормативно-методической базы, разработке стратегических документов и планировании дальнейших исследований. В качестве приоритетных направлений научного поиска следует рассматривать изучение когнитивных и лингвистических факторов тестирования, разработку интеллектуальных систем генерации тестовых заданий и анализ долгосрочного влияния модернизированной системы оценивания на образовательные результаты.

Реализация проекта создала прочную основу для дальнейшего развития системы ЕНТ, что будет способствовать повышению качества образования в Республике Казахстан. В этом контексте ключевым направлением становится разработка научно обоснованных решений, основанных на представленных рекомендациях, которые акцентируют важность исследований валидации тестовых заданий ЕНТ.

Для достижения данной цели необходимо систематическое научно-методическое обоснование разработки тестовых заданий, а также применение методов, продемонстрировавших свою эффективность в совершенствовании оценочных процедур. В рамках проекта были выявлены оптимальные подходы к проведению научных исследований в НЦТ, что, в свою очередь, подчеркивает необходимость их институционализации и активного внедрения научной деятельности в данную сферу.

Благодарность, конфликт интересов

Авторы выражают благодарность Министерству науки и высшего образования Республики Казахстан и консорциуму Ginger-SOFRECO за поддержку проекта.

Литература

1. adilet.zan.kz. Об утверждении Правил проведения единого национального тестирования. Приказ Министра образования и науки Республики Казахстан от 2 мая 2017 года № 204. https://adilet.zan.kz/kaz/docs/V1700015173.

2. adilet.zan.kz Концепции развития дошкольного, среднего, технического и профессионального образования Республики Казахстан на 2023 – 2029 годы. https://adilet.zan.kz/rus/docs/P2300000249.

3. Карданова Е.Ю., Иванова А.Е. Психометрические исследования: современные методы и новые возможности для образования. Вопросы образования / Educational Studies Moscow, № 3. – 2023. - сс. 8-19. https://doi.org/10.17323/vo-2023-17951.

4. 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 cApStAn (Бельгия).

5. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. – 2023. https://www.R-project.org.

6. Willse, J.T. (2018). CTT: Classical Test Theory Functions. R package version 2.3.3. https://CRAN.R-project.org/package=CTT.

7. Robitzsch, A., Kiefer, T., & Wu, M. (2022). TAM: Test Analysis Modules. R package version 4.1-4. https://CRAN.Rproject.org/package=TAM.

8. sciencedirect.com Альфа-коэффициент Кронбаха. https/www.sciencedirect.com/topics/nursing-and-health-professions/cronbach-alpha coefficient.

9. Маслак А.А., Поздняков С.А. Анализ качества тестовых заданий с выбором одного правильного ответа //Методические рекомендации. - Славянск-на-Кубани: Издательский центр СГПИ, 2009 – 50 с.

10. Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 27(2), 53-55. doi:10.5116/ijme.4dfb.8dfd.

11. cyberleninka.ru. Аванесов В.С. Дистракторный анализ. https://cyberleninka.ru/article/n/distraktornyy-analiz-1/viewer.

12. Аванесов В.С. Применение тестовых форм в e-learning с проведением дистракторного анализа. Образовательные технологии. № 3/2013. https://iedtech.ru/files/journal/2013/3/avanesov.pdf

13. adilet.zan.kz. Об утверждении государственных общеобязательных стандартов образования всех уровней образования. Приказ Министра образования и науки Республики Казахстан от 31 октября 2018 года № 604. http://adilet.zan.kz/rus/docs/V1800017669.

Б.С. Абдрасилов, Ш.Б. Алтыбаева, Л.Е. Шинетова

ДҮНИЕЖҮЗІЛІК БАНК ЖОБАСЫ АЯСЫНДА ЖҮРГІЗІЛГЕН ҰБТ ЖҮЙЕСІН ЗЕРТТЕУ НӘТИЖЕЛЕРІ

Бұл зерттеуде Дүниежүзілік Банктің «Орта білім беруді жаңғырту» (бұдан әрі - ДБ) жобасы аясында жүргізілген Қазақстан Республикасындағы Ұлттық бірыңғай тестілеу (бұдан әрі - ҰБТ) жүйесіне кешенді психометриялық талдауы ұсынылады. Зерттеу келесі үш негізгі компонентті біріктіретін көп деңгейлі тәсілге негізделген: халықаралық тестология стандарттарына сәйкес 350 тест әзірлеушілерді кәсіби даярлау; 15 пән бойынша 8 848 тест тапсырмаларының базасын құру және психометриялық валидация; қосымша аналитикалық әдістерді қолдану (CTT және Rash модельдері). Зерттеудің негізгі нәтижелері әзірленген құралдардың психометриялық сәйкестігінің жоғары деңгейін көрсетеді: тапсырмалардың 90,7% влидтілік пен сенімділік критерийлеріне сәйкес келеді. Пәндік салалар мен тесттердің тілдік нұсқалары арасындағы сенімділік көрсеткіштерінен статистикалық маңызды айырмашылықтар анықталды, бұл когнитивті-лингвистикалық факторларды ескере отырып, одан әрі зерттеуді қажет етеді. Тапсырмалардың саралау қабілетін және дистракциялық аппараттың тиімділігін талдауға ерекше назар аударылады. Нәтижелер тест материалдарын әзірлеудің әдіснамалық тәсілдерінің дұрыстығын және олардың пәндік дайындық деңгейі бойынша субъектілерді сенімді түрде саралау қабілетін растайды. Зерттеудің практикалық маңыздылығы одан әрі зерттеулер жүргізу және осы салада ұсыныстар әзірлеу үшін негіз қалайды. Зерттеу психометрия саласындағы мамандарды, тест материалдарын әзірлеушілерді және білім беру жетістіктерін бағалау жүйесін жетілдірумен айналысатын білім беру саласындағы саясаткерлерді қызықтырады.

Түйін сөздер: ҰБТ, психометриялық талдау, валидттілік, сенімділік, Rash моделі, дистракторлық талдау, білім беруді бағалау.

B.S. Abdrasilov, Sh. Altybayeva, L.E. Shinetova

THE RESULTS OF THE UNT SYSTEM STUDY CONDUCTED BY WITHIN THE FRAMEWORK OF THE WORLD BANK PROJECT

The present study is a comprehensive psychometric analysis of the Unified National Testing (UNT) system in the Republic of Kazakhstan, conducted within the framework of the World Bank's project «Modernization of Secondary Education» (hereinafter - WB). The study is based on a multi-level approach combining three key components: professional training of 350 test developers in accordance with international standards of testology; creation and psychometric validation of a database of 8,848 test tasks in 15 disciplines; application of complementary analytical methods (CTT and Rasch models). The key results of the study demonstrate a high level of psychometric consistency of the developed toolkit: 90.7% of tasks meet the criteria of validity and reliability. Statistically significant differences in reliability indicators between the subject areas and the language versions of the tests have been revealed, which requires further study taking into account cognitive and linguistic factors. Special attention is paid to the analysis of the differentiating ability of tasks and the effectiveness of the distractor apparatus. The results confirm the correctness of methodological approaches to the development of test materials and their ability to reliably differentiate subjects by the level of subject training. The practical significance of the research lays the foundation for further research and recommendations in this area. The study is of interest to specialists in the field of psychometry, developers of test materials and educational policy makers involved in improving educational achievement assessment systems.

Keywords: UNT, psychometric analysis, validity, reliability, Rasch model, distractor analysis, educational assessment.

References

1. adilet.zan.kz. On approval of the 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. https://adilet.zan.kz/kaz/docs/V1700015173.

2. adilet.zan.kz Concepts for the development of preschool, secondary, technical and vocational education in the Republic of Kazakhstan for 2023-2029. https://adilet.zan.kz/rus/docs/P2300000249.

3. Kardanova E.Y., Ivanova A.E. (2023) Psychometric research: modern methods and new educational opportunities. Questions of Education / Educational Studies Moscow, No. 3, pp. 8-19. https://doi.org/10.17323/vo-2023-17951

4. Final report for the period April 13 – June 9, 2024 under the service contract no. KZEMP/QCBS-11 «Improvement of the UNT assessment system». The Ginger–SOFRECO Consortium (France), the National Foundation for Educational Research (Great Britain), Be Supply LLP (Kazakhstan) and cApStAn (Belgium).

5. R Core Team. (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org.

6. Willse, J. T. (2018). CTT: Classical Test Theory Functions. R package version 2.3.3. https://CRAN.R-project.org/package=CTT.

7. Robitzsch, A., Kiefer, T., & Wu, M. (2022). TAM: Test Analysis Modules. R package version 4.1-4. https://CRAN.Rproject.org/package=TAM.

8. sciencedirect.com Alpha is the Cronbach's coefficient. https/www.sciencedirect.com/topics/nursing-and-health-professions/cronbach-alpha coefficient.

9. Maslak A.A., Pozdnyakov S.A. (2009) Quality analysis of test tasks with the choice of one correct answer //Methodological recommendations. Slavyansk-on-Kuban: SSPI Publishing Center, 50 p.

10. Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 27(2), 53-55. doi:10.5116/ijme.4dfb.8dfd.

11. cyberleninka.ru. Avanesov V. S. Distractor analysis. https://cyberleninka.ru/article/n/distraktornyy-analiz-1/viewer.

12. Avanesov V.S. Application of test forms in e-learning with distractor analysis. Educational technologies. № 3/2013. https://iedtech.ru/files/journal/2013/3/avanesov.pdf

13. adilet.zan.kz. On the approval of state mandatory standards of education at all levels of education. Order of the Minister of Education and Science of the Republic of Kazakhstan dated October 31, 2018 No. 604. http://adilet.zan.kz/rus/docs/V1800017669.

Авторлар туралы мәлімет:

Абдрасилов Болатбек Серикбаевич - ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК директоры, Родниковая көшесі, 1/1, 010008, Астана, Қазақстан, e-mail: uto@testcenter.kz.

Алтыбаева Шугыла Болатовна - ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК Орта және жоғары білім берудегі тест тапсырмаларын қалыптастыру басқармасының басшысы, Родниковая көшесі, 1/1, 010008, Астана, Қазақстан, e-mail: shugla@mail.ru

Шинетова Ляззат Ермековна - ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК Жоғары, Ғылыми зерттеулер мен психометрика зертханасының меңгерушісі, Родниковая көшесі, 1/1, 010008, Астана, Қазақстан, e-mail: shinetovalyazzat24@gmail.com.

Сведения об авторах:

Абдрасилов Болатбек Серикбаевич – директор РГП на ПХВ «Национальный центр тестирования» МНВО РК, ул. Родниковая, 1/1, 010008, Астана, Казахстан, e-mail: uto@testcenter.kz.

Алтыбаева Шугыла Болатовна - руководитель Управления по формированию тестовых заданий для высшего и среднего образования РГП на ПХВ «Национальный центр тестирования» МНВО РК, ул. Родниковая, 1/1, 010008, Астана, Казахстан, e-mail: shugla@mail.ru

Шинетова Ляззат Ермековна - заведующий лабораторией научных исследований и психометрики РГП на ПХВ «Национальный центр тестирования» МНВО РК, ул. Родниковая, 1/1, 010008, Астана, Казахстан, e-mail: shinetovalyazzat24@gmail.com.

Information about authors:

Abdrasilov Bolatbek Serikbayevich – Director of the National Testing Center of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Rodnikovaya STR., 1/1, 010008, Astana, Kazakhstan, e-mail: uto@testcenter.kz.

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, Rodnikovaya STR., 1/1, 010008, Astana, Kazakhstan, e-mail: shugla@mail.ru.

Shinetova Lyazzat Ermekovna – 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, Rodnikovaya STR., 1/1, 010008, Astana, Kazakhstan, e-mail: shinetovalyazzat24@gmail.com.