Валидизация Единого национального тестирования в Казахстане: доказательная основа справедливого оценивания

Болатбек Абдрасилов, Ляззат Шинетова, Алина Иванова, Елена Карданова, Жанар Бейсенова, Гульдана Жабаева, Давид Орлов, Гулнур Аширбек

Статья поступила в редакцию в июле 2025 г.

Абдрасилов Болатбек Серикбаевич — кандидат физико-математических наук, доктор биологических наук, директор, Национальный центр тестирования Республики Казахстан. E-mail: bolatbek_s@mail.ru. ORCID: https://orcid.org/0009-0002-1371-6211

Шинетова Ляззат Ермековна — заведующая лабораторией, Национальный центр тестирования Республики Казахстан. E-mail: shinetovalyazzat24@gmail.com. ORCID: https://orcid.org/0000-0003-4280-7999

Иванова Алина Евгеньевна — кандидат наук об образовании, старший научный сотрудник Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики». E-mail: aeivanova@hse.ru. ORCID: https://orcid.org/0000-0003-3340-7651

Карданова Елена Юрьевна — кандидат физико-математических наук, научный руководитель Центра психометрики и измерений в образовании Института образования, Национальный исследовательский университет «Высшая школа экономики». Адрес: 101000 Москва, Потаповский пер., 16, стр. 10. E-mail: ekardanova@hse.ru. ORCID: https://orcid.org/0000-0003-2280-1258 (контактное лицо для переписки)

Бейсенова Жанар Жумагазиновна — ведущий научный сотрудник, Национальный центр тестирования Республики Казахстан. E-mail: zhanar-beisenova@mail.ru. ORCID: https://orcid.org/0000-0002-7347-1538

Жабаева Гульдана Куатовна — ведущий научный сотрудник, Национальный центр тестирования Республики Казахстан. E-mail: guldanazhabayeva@gmail.com. ORCID: https://orcid.org/0009-0006-1003-6316

Орлов Давид — доктор философии (PhD) в области евразийских исследований, главный научный сотрудник, Национальный центр тестирования Республики Казахстан. E-mail: dorlov@nu.edu.kz. ORCID: https://orcid.org/0000-0002-7699-5345

Аширбек Гулнур Құрманғалиқызы — заведующий лабораторией, Национальный центр тестирования Республики Казахстан. E-mail: gulnur_ashirbek@mail.ru. ORCID: https://orcid.org/0009-0006-6326-3088

Аннотация Во всем мире к вступительным экзаменам в вузы, как к экзаменам с высокими ставками, предъявляются серьезные требования: они должны быть максимально объективными, надежными и справедливыми. Это означает, что процесс их разработки, проведения и оценки должен быть тщательно контролируемым и стандартизированным, чтобы исключить любую возможность предвзятости. В Казахстане прием в вузы осуществляется по результатам государственного вступительного экзамена — Единого национального тестирования (ЕНТ).

Проведено исследование с целью оценки валидности результатов Единого национального тестирования в Казахстане на примере математики. Основное внимание уделено двум блокам свидетельств валидности: на основе внутренней структуры теста и на основе связи результатов тестирования с другими переменными, в частности с критериями последующей успешности студентов на этапе получения высшего образования. Теоретической рамкой исследования послужила теория валидности С. Мессика. Серия валидизационных исследований проводилась согласно методологическим требованиям объединенных Стандартов образовательного и психологического тестирования *American Educational Research Association, American Psychological Association* и *National Council on Measurement in Education* с использованием подходов классической теории тестирования, современной теории тестирования, а также статистического анализа, включая иерархическое линейное моделирование. Полученные результаты свидетельствуют о том, что экзамен ЕНТ по математике обладает высоким психометрическим качеством. При этом общий балл ЕНТ по пяти предметам и отдельно балл ЕНТ по математике позволяют надежно прогнозировать будущую успешность обучения студентов в вузе. Установлено, что качество использованных статистических моделей и объясняемая ими дисперсия оценок студентов за первый семестр во многом согласуются с результатами исследований, проведенных в других странах, включая Россию и США.

Ключевые слова экзамены с высокими ставками, валидизация, прогностическая валидность, ЕНТ, Казахстан

Для цитирования Абдрасилов Б.С., Шинетова Л.Е., Иванова А.Е., Карданова Е.Ю., Бейсенова Ж.Ж., Жабаева Г.К., Орлов Д., Аширбек Г.К. (2026) Валидизация Единого национального тестирования в Казахстане: доказательная основа справедливого оценивания. *Вопросы образования / Educational Studies Moscow.* https://doi.org/10.17323/vo-2026-27950

Validation of the Unified National Testing in Kazakhstan: Evidence-Based Framework for Fair Testing

Bolatbek Abdrasilov, Lyazzat Shinetova, Alina Ivanova, Elena Kardanova, Zhanar Beisenova, Guldana Zhabayeva, David Orlov, Gulnur Ashirbek

Bolatbek S. Abdrasilov — Candidate of Sciences in Physics and Mathematics, Doctor of Sciences in Biology, Director, National Testing Center (Kazakhstan). E-mail: bolatbek_s@mail.ru. ORCID: https://orcid.org/0009-0002-1371-6211

Lyazzat E. Shinetova — Head of the Psychometric Laboratory, National Testing Center (Kazakhstan). E-mail: shinetovalyazzat24@gmail.com. ORCID: https://orcid.org/0000-0003-4280-7999

Alina E. Ivanova — Candidate of Sciences in Education (PhD), Senior Researcher, Center for Psychometrics and Measurement in Education, Institute of Education, HSE University. E-mail: aeivanova@hse.ru. ORCID: https://orcid.org/0000-0003-3340-7651

Elena Yu. Kardanova — Candidate of Sciences (PhD) in Physics and Mathematics, Scientific Supervisor of the Center for Psychometrics and Measurement in Education, Institute of Education, HSE University. Address: 16/10 Potapovsky lane, 101000 Moscow, Russian Federation. E-mail: ekardanova@hse.ru. ORCID: https://orcid.org/0000-0003-2280-1258 (corresponding author)

Zhanar Zh.Beisenova — Senior Research Fellow, National Testing Center (Kazakhstan). E-mail: zhanar-beisenova@mail.ru. ORCID: https://orcid.org/0000-0002-7347-1538

Guldana K. Zhabayeva — Senior Research Fellow, National Testing Center (Kazakhstan). E-mail: guldanazhabayeva@gmail.com. ORCID: https://orcid.org/0009-0006-1003-6316

David Orlov — PhD in Eurasian Studies, Chief Research Fellow, National Testing Center (Kazakhstan). E-mail: dorlov@nu.edu.kz. ORCID: https://orcid.org/0000-0002-7699-5345

Gulnur K. Ashirbek — Head of Laboratory, National Testing Center (Kazakhstan). E-mail: gulnur_ashirbek@mail.ru. ORCID: https://orcid.org/0009-0006-6326-3088

Abstract

Worldwide, high-stakes entrance examinations are subject to rigorous requirements: they must be as objective, reliable, and fair as possible. This entails that the processes of their development, administration, and scoring are carefully controlled and standardized to eliminate any possibility of bias. In Kazakhstan, university admission is based on the results of the state entrance examination — the Unified National Testing (UNT). The aim of this study is to investigate the validity of the UNT results in Kazakhstan, focusing on the mathematics component. Particular attention is given to two sources of validity evidence: based on internal test structure and based on relationship with other variables, specifically criteria of students' subsequent academic success in higher education. The theoretical framework for this research is grounded in S. Messick's validity theory. A series of validation studies were conducted in accordance with the methodological standards set forth by American Educational Research Association, American Psychological Association and National Council on Measurement in Education, employing classical test theory, modern test theory approaches, and statistical analyses including hierarchical linear modeling. The findings indicate that the UNT mathematics exam demonstrates strong psychometric properties. Furthermore, both the overall UNT score across five subjects and the mathematics effectively predict students' future academic performance at university. The quality of the statistical models used and the proportion of variance explained in students' first-semester grades are largely consistent with results reported by researchers from other countries, including Russia and the United States.

Keywords

high-stakes exams, validation, predictive validity, UNT, Kazakhstan

For citing

Abdrasilov B.S., Shinetova L.E., Ivanova A.E, Kardanova E.Yu., Beisenova Zh.Zh., Zhabayeva G.K., Orlov D., Ashirbek G.K. (2026) Validation of the Unified National Testing in Kazakhstan: Evidence-Based Framework for Fair Testing. *Voprosy obrazovaniya / Educational Studies Moscow*. https://doi.org/10.17323/vo-2026-27950

Приемлемость централизованных экзаменов как способа отбора абитуриентов в высшие учебные заведения уже в течение долгого времени активно обсуждается как профессиональным сообществом специалистов в области образования, так и широкой общественностью. Самые важные вопросы, которые наиболее часто поднимаются в рамках таких дискуссий, — это эффективность централизованных экзаменов как средства выравнивания доступа к высшему образованию и их прогностическая валидность в отношении будущей академической успеваемости. О том, что централизованные экзамены признаны испытанием, подходящим для справедливого отбора студентов, которые могут успешно завершить университетскую программу, свидетельствует их продолжительное использование в разных странах мира [Garg, Li, Monachou, 2020; Zwick, 2019]. По данным Организации экономического сотрудничества и развития, 31 из 38 государств — членов организации проводят национальные вступительные экзамены в университеты [OECD, 2020].

По этому же пути пошли и почти все государства постсоветского пространства [Bethell, Zabulionis, 2012]. Основными целями введения централизованных вступительных экзаменов в этих странах были обеспечение равного доступа всех обучающихся к высшему образованию, поиск и продвижение наиболее талантливых студентов посредством выделения стипендий и грантов на образование, повышение прозрачности вступительных экзаменов в вузы, предотвращение нарушений и коррупции в этой сфере и в итоге улучшение качества подготовки выпускников вузов как необходимое условие экономического роста стран.

Модели вступительных экзаменов в вузы различаются в разных странах: экзамены могут быть обязательными или опциональными, фокусироваться на знании содержания дисциплины или на навыках, они различаются по частоте проведения и политике пересдач, являются единственным основанием оценивания или используются совместно с другими источниками информации. Однако, независимо от модели, к вступительным экзаменам, как к экзаменам с высокими ставками, предъявляются высокие требования: они должны обеспечивать максимально объективные, надежные и справедливые результаты[1].

[1] Под объективностью в данной статье понимается минимизация субъективного вклада экзаменаторов или экспертов в оценивание абитуриентов через использование заранее определенных критериев оценки. Надежность означает, что результаты тестирования устойчивы и воспроизводимы: при повторном администрировании или использовании альтернативной формы вариации ошибки остаются минимальными, а истинная дисперсия доминирует над ошибочной. Справедливость — это комплексное требование, которое должно обеспечиваться на каждом этапе тестирования (от формулировки заданий до интерпретации результатов), но в техническом смысле оно означает необходимость устранять нерелевантные конструкту источни-

В Казахстане прием в вузы осуществляется по результатам государственного вступительного экзамена — Единого национального тестирования (ЕНТ). Он введен в 2004 г., и в период с 2004 по 2024 г. в среднем 126,3 тыс. абитуриентов ежегодно сдавали этот экзамен. Основная цель ЕНТ — объективная оценка знаний абитуриентов и распределение государственных образовательных грантов на получение высшего образования.

Несмотря на более чем двадцатилетнюю историю существования ЕНТ, посвященных ему исследований в академической литературе немного. В ряде работ анализировалась связь результатов экзамена с демографическими характеристиками абитуриентов, такими как пол, место проживания, язык проведения экзамена [Mingisheva, 2023; Amangeldiyeva, 2024; Shabdenova, Satybayeva, 2024], а также с использованием услуг частных репетиторов и социально-экономическими показателями абитуриентов [Смагулова, Сатанов, Кадирова, 2025; Hajar, Abenova, 2021]. Нам удалось обнаружить только одну работу, в которой исследуется психометрическое качество тестов ЕНТ: в ней проведен общий анализ почти 9 тыс. тестовых заданий по 15 дисциплинам и подтверждено высокое психометрическое качество инструментария ЕНТ [Абдрасилов, Алтыбаева, Шинетова, 2025].

Целью настоящей работы является исследование валидности результатов Единого национального тестирования в Казахстане с акцентом на прогностическую валидность. В качестве теоретической рамки исследования выбрана теория валидности С. Мессика [Messik, 1994], отраженная в наиболее комплексных из современных стандартов измерений в образовании — Стандартах образовательного и психологического тестирования (*Standards for Educational and Psychological Testing)* [American Educational Research Association et al., 2014] и уточненная в недавних исследованиях, внесших весомый вклад в развитие теории валидности [Sireci, 2021; Sireci, Benítez, 2023].

1. Понятие валидности применительно к централизованному тестированию при отборе в вузы

1.1. Исследования валидности результатов вступительных экзаменов: мировой опыт

О валидности любого экзамена, включая вступительный, судят по тому, насколько вероятно с его помощью получить достоверную информацию и реализовать таким образом заявленную цель экзамена. В Стандартах образовательного и психологического тестирования [American Educational Research Association et al., 2014] закреплено требование обоснования валидности как основы для применения теста, особенно если речь идет об экзаменах с высокими ставками, по результатам которых принимаются жизнен-

ки вариации, контролировать различное функционирование заданий (DIF), и в более широком смысле — в перспективе отслеживать последствия применения теста.

но важные решения. Вступительные экзамены в вузы безусловно относятся к этой категории испытаний.

Вступительные экзамены в вузы — это не просто административный барьер. Это решающий инструмент отбора, от которого зависит справедливость доступа к высшему образованию, будущая академическая успеваемость студентов и их карьера, а также эффективность государственных инвестиций в образование [Stemler, 2012]. Поэтому к вступительным экзаменам, как к экзаменам с высокими ставками, предъявляются особые требования: весь процесс их разработки, процедуры проведения и оценивания тщательно регламентированы. Обоснование объективности, надежности и справедливости результатов экзамена обеспечивает валидность его результатов и принимаемых на его основе решений, повышает доверие к экзамену в обществе.

В условиях высокой значимости вступительных испытаний большое внимание уделяется психометрическому качеству инструментария экзаменов — как отдельных заданий, так и тестов в целом, а также параллельности вариантов теста, справедливости оценивания и отсутствию предвзятости по отношению к какой-либо группе учащихся. Организации, занимающиеся разработкой тестов для вступительных экзаменов, проводят масштабные исследования валидности и надежности, регулярно публикуют технические отчеты, содержащие обоснование психометрического качества инструментария. Например, в России разработку тестов для Единого государственного экзамена проводит Федеральный институт педагогических измерений[2], на официальном сайте которого можно найти всю информацию об экзамене, включая кодификаторы содержания, спецификации тестов, их демоверсии, а также технические отчеты.

К числу старейших стандартизированных вступительных испытаний относятся американские экзамены *Scholastic Assessment Test*, SAT [College Board, 2025] и *American College Testing*, ACT [ACT, 2024]), появившиеся в 1926 и 1959 гг. соответственно. Их разработке, валидизации и последующему анализу вторичных данных посвящен большой массив публикаций. Для обоснования качества заданий тестов SAT и ACT, выравнивания данных по разным вариантам заданий, проверки справедливости измерений по отношению к отдельным группам тестируемых, выделенных, например, по полу или этнической принадлежности, используется современная теория тестирования (*Item Response Theory*, IRT) [De Ayala, 2018]. Использование IRT обеспечивает справедливость результатов и сопоставимость данных между разными версиями и разными группами участников, сдающих тест по разным вариантам и в разное время. В отношении обоих тестов подтвержде-

[2] https://fipi.ru/ege

на высокая надежность измерений: коэффициент надежности альфа Кронбаха и для SAT, и для ACT выше 0,9 [Amrein-Beardsley et al., 2025].

Применительно к вступительным экзаменам особое внимание уделяется прогностической валидности — их способности эффективно предсказывать будущие академические успехи обучающихся в высшем образовании [Marini et al., 2020; Sanchez, 2025]. Традиционный подход к оценке прогностической валидности основывается на выявлении корреляционной связи между результатами вступительных испытаний и средним баллом за первый год обучения в университете (*First Year Grade Point Average*, FYGPA) [Bai, Chi, Qian, 2014; Migliaretti et al., 2017]. Исследования прогностической валидности SAT имеют богатую историю. Одно из первых — метаанализ Д. Фишмана и Э. Пазанелла, которые на основании 147 исследований, проведенных в период с 1950 по 1960 г., установили, что коэффициенты корреляции между баллами SAT и школьными оценками, с одной стороны, и средним баллом за первый год обучения (FYGPA) — с другой, варьируют от 0,34 до 0,82 [Fishman, Pasanella, 1960]. Р. Морган исследовал прогностическую валидность SAT и школьных оценок (*High School Grade Point Average*, HSGPA) в период с 1976 по 1985 г. и обнаружил, что SAT и HSGPA в совокупности показывали более высокую прогностическую валидность, чем по отдельности: корреляции SAT с FYGPA составляли в среднем около 0,4, корреляции HSGPA с FYGPA были сопоставимыми или немного выше, а их комбинация давала коэффициент до 0,60 и выше, тем самым была подтверждена взаимодополняемость этих показателей [Morgan, 1989].

Схожие результаты получены в метааналитическом исследовании с использованием тестов ACT [Westrick et al., 2015]. Оценивалась сила взаимосвязей комбинированных баллов ACT, школьных оценок, а также показателей социально-экономического статуса (СЭС) с академической успеваемостью и продолжением обучения на 2-м и 3-м курсах четырехгодичных колледжей и университетов. На выборке около 200 тыс. студентов из 50 учебных заведений авторы показали, что комбинированные баллы ACT и средний балл аттестата (*Grade Point Average*, GPA) имеют высокую корреляцию с успеваемостью на 1-м курсе, которая, в свою очередь, является наилучшим предиктором продолжения обучения на 2-м и 3-м курсах, а СЭС оказался хоть и значимым, но слабым предиктором.

Таким образом, в многочисленных исследованиях убедительно показано, что стандартизированные вступительные тесты, такие как SAT и ACT, могут служить значимыми предикторами академической успеваемости студентов в высших учебных заведениях. По данным этих исследований, коэффициент корреляции баллов по стандартизированным вступительным тестам с показате-

лем академической успешности студента находится в диапазоне от 0,3 до 0,6. Эти цифры в некоторой степени могут служить ориентиром для оценки прогностической валидности вступительных испытаний в других странах.

Одна из первых статей, посвященных прогностической валидности Единого государственного экзамена в России, вышла в 2014 г. — через год после того, как дипломы о высшем образовании получил первый поток студентов, в большинстве своем поступавших в вузы по результатам ЕГЭ [Хавенсон, Соловьева, 2014]. Авторы с помощью линейного регрессионного анализа на выборке из почти 19 тыс. студентов, обучавшихся в вузах на разных направлениях подготовки, оценили способность Единого государственного экзамена предсказывать дальнейшую успеваемость студентов в вузе. Использовались и суммарный балл ЕГЭ, и баллы ЕГЭ по отдельным предметам. Среднее значение коэффициента детерминации для моделей с суммарным баллом ЕГЭ составило порядка 0,20, т.е. в среднем по разным направлениям подготовки успеваемость студентов на 1-м курсе на 20% объяснялась только одним фактором — баллом вступительных экзаменов в форме ЕГЭ. На разных факультетах этот показатель варьировал от 15 до 35%. Предсказательная способность баллов ЕГЭ по отдельным предметам была примерно одинакова, но баллы ЕГЭ по математике и русскому языку оказались лучшими предикторами для подавляющего большинства направлений.

Таким образом, результаты стандартизированных выпускных или вступительных тестов в разных странах являются достаточно надежными предикторами последующей академической успешности обучающегося. Разброс оценок в исследованиях часто возникает из-за различий в контексте, обусловленных такими факторами, как уровень селективности учебных заведений, неоднородность учебных планов, различия в стандартах оценивания и набор используемых сопутствующих предикторов, среди которых могут быть демографические данные, показатели СЭС, школьные оценки, университетские оценки и рейтинги [Burton, Ramist, 2001].

1.2. Теория валидности С. Мессика

С. Мессик определил валидность как интегральную оценку степени обоснованности интерпретаций и действий на основе результатов тестирования [Messick, 1994]. В его унифицированной концепции валидность — это комплексное суждение, объединяющее эмпирические доказательства, теоретические обоснования, а также ценностные и социальные последствия. Именно его определение валидности стало основой требований к проведению процедур валидизации, сформулированных в Стандартах образовательного и психологического тестирования 2014 г. [American Educational Research Association et al., 2014].

С. Мессик считает валидность, надежность, сопоставимость и справедливость не только техническими условиями корректных измерений, хотя они, безусловно, таковыми являются, но и социальными ценностями, которые имеют отношение к любым оценочным суждениям и решениям [Messick, 1994]. По мнению С. Сиречи [Sireci, 2021], именно профессиональные ценности специалистов в измерениях во многом определяют, что и как будет измеряться, как именно будут интерпретироваться и использоваться результаты тестов, а также каким образом будет обосновываться валидность решений, основанных на результатах тестирования. С. Сиречи сформулировал эти определяющие ценностные суждения следующим образом:

1) каждый человек способен к обучению;
2) нет различий в способности к обучению между группами, определяемыми по расовой, этнической или половой принадлежности;
3) любые образовательные тесты в той или иной степени подвержены искажениям и имеют ошибку измерения;
4) образовательные тесты могут давать ценную информацию для улучшения обучения или подтверждения компетентности;
5) все случаи использования результатов тестов должны быть должным образом обоснованы доказательствами валидности.

В русле рассматриваемого подхода сами инструменты измерения не являются «валидными» или «невалидными». Валидными или невалидными могут быть выводы, сделанные на основе результатов тестирования, причем для конкретной цели, в конкретном контексте, для определенной группы и зачастую в конкретный исторический момент. При этом не существует единого доказательства, которое могло бы подтвердить валидность результатов тестирования, вместо этого происходит постоянный процесс сбора доказательств валидности. Другими словами, валидность не устанавливается раз и навсегда.

Согласно стандартам Американской ассоциации исследователей в области образования [American Educational Research Association et al., 2014], валидность подтверждается с помощью пяти групп свидетельств: на основе содержания теста, процесса порождения ответа на вопросы теста, внутренней структуры теста, связи результатов тестирования с другими переменными и последствий тестирования.

Свидетельства валидности теста по содержанию (содержательная валидность) получают, проверяя соответствие содержания теста измеряемому конструкту. Для сбора таких свидетельств

проводится систематический обзор исследований конструкта и смежных с ним областей и/или привлекаются эксперты в той области, к которой принадлежит измеряемый конструкт.

Подтверждение содержательной валидности теста является крайне важным этапом валидизации тестов с высокими ставками. Свидетельствами валидности теста по содержанию могут служить результаты анализа содержания теста на соответствие оцениваемым областям знания; оценки полноты, релевантности и достаточности этих областей для предполагаемого использования результатов; рекомендации и заключения экспертов по содержанию теста. К примеру, согласно результатам систематического обзора [Amrein-Beardsley et al., 2025], при анализе валидности SAT на основании всестороннего описания содержания оцениваемого конструкта было показано, что SAT не предназначен для измерения способностей — ранее результаты применения теста с этой целью были предметом его критики, — а оценивает когнитивные навыки, приобретенные как в школе, так и вне ее [Messick, Jungeblut, 1981], а также что SAT и ACT являются наилучшими испытаниями для определения готовности к колледжу [Linn, 2009]. В то же время другая группа исследователей установила, что по причине существенных различий в образовательных приоритетах между американскими штатами, а также между университетами ни SAT, ни ACT нельзя считать удовлетворительными инструментами ни для оценки школьной программы обучения, ни для оценки знаний и навыков, важных с точки зрения готовности к колледжу [Atkinson, Geiser, 2009].

Свидетельства на основе процесса порождения ответа отражают когнитивные и психологические условия, влияющие на выполнение теста. Такие свидетельства собирают, например, с помощью когнитивных интервью с тестируемыми, предназначенных для того, чтобы понять, как они воспринимают вопросы, инструкции и проч. В 18 статьях из числа рассмотренных в обзоре валидности SAT и ACT [Amrein-Beardsley et al., 2025] представлены доказательства валидности, связанные с ответами, которые авторы охарактеризовали как нейтральные — не подтверждающие и не опровергающие обоснованность использования этих тестов. С точки зрения этого типа валидности обсуждается также уровень когнитивной нагрузки SAT и ACT [Aguinis, Culpepper, Pierce, 2016]. Тщательно задокументированные процедуры стандартизации проведения тестирования, инструкций, регулирующих процесс ответа на тест, также могут быть отнесены в эту группу свидетельств.

Свидетельства на основе внутренней структуры теста получают, когда изучают взаимосвязи между заданиями теста, включая надежность, факторную структуру, характеристики заданий теста. Эти свидетельства собирают путем тщательного психометриче-

ского анализа результатов тестирования. В выборке упомянутого ранее систематического обзора [Amrein-Beardsley et al., 2025] как минимум восемь разных работ представили доказательства валидности тестов SAT и ACT, основанные на их внутренней структуре. Свидетельства на основе внутренней структуры тестов применительно к вступительным экзаменам достаточно хорошо изучены и рассмотрены в предыдущем разделе статьи.

Свидетельства на основе связи с другими переменными представляют собой результаты оценки связи результатов тестирования и теми или иными внешними данными, например, с показателями, полученными при применении аналогичных инструментов тестирования. Применительно к вступительным экзаменам в эту группу свидетельств можно отнести анализ прогностической силы инструмента, а также меру справедливости функционирования заданий теста в разрезе разных групп тестируемых. Оба этих направления оценки валидности рассмотрены в предыдущем разделе на примере тестов SAT и ACT.

Свидетельства на основе последствий тестирования — это результаты оценки влияния теста, как позитивного, так и негативного, на тестируемых. Эта группа свидетельств наименее исследованная, поскольку предполагает, по сути, большой временной разрыв между тестированием и оценкой его последствий. Примером такого рода работ является серия исследований С. Лэйн [Lane, 2014], которая изучала последствия внедрения Программы оценки успеваемости штата Мэриленд (*Maryland School Performance Assessment Program*, MSPAP) в США и показала положительное влияние этой программы на результаты деятельности школ. Большинство учебных мероприятий по математике проводилось в соответствии с требованиями MSPAP, и школы, уделявшие, чтобы соответствовать этим требованиям, больше внимания практикоориентированному обучению, включая развитие у учащихся навыков критического мышления и логического рассуждения, получили более высокие баллы MSPAP по сравнению со школами, не перестроившими обучение в соответствии с программой. Другой пример — изучение свидетельств валидности на основе последствий для национальных экзаменов на Кипре [Michaelides, 2014]. Автор выяснял, как студенты воспринимают цели экзаменов, понимают их технические характеристики и интерпретируют результаты своих испытаний. В интервью со студентами, незадолго до этого сдававшими экзамены, обнаружилось, что у них есть некоторые сомнения в справедливости и вообще уместности системы экзаменов, что автор назвал непредвиденным последствием тестирования.

Представленный подход к оценке валидности результатов тестирования помогает исследователям и практикам систематизировать доказательства, делая выводы более обоснованны-

ми и надежными. Однако нам не удалось обнаружить ни одной комплексной статьи, в которой применительно к конкретному экзамену с высокими ставками была бы реализована идея сбора всех свидетельств валидности на основе требований измерительных стандартов 2014 г. Публикуется только техническая отчетная документация, например [College Board, 2025]. Тем не менее свежий систематический обзор литературы [Amrein-Beardsley et al., 2025], который охватил более 70 научных публикаций с 1969 по 2019 г., посвященных валидизации SAT и ACT, дает наглядное представление как об усилиях, которые исследователи прилагают к проведению валидизации, так и о необходимости таких усилий. В валидизационном исследовании казахстанского ЕНТ, представленном в данной статье, мы осуществили максимально комплексную оценку в рамках имеющихся данных и фактически возможного объема одной публикации.

2. Обоснование валидности результатов ЕНТ в Казахстане

Настоящее исследование имеет целью обоснование валидности результатов Единого национального тестирования в Казахстане на основе теории валидности С. Мессика.

Для ответа на главный исследовательский вопрос — какие свидетельства подтверждают валидность интерпретации результатов ЕНТ как экзамена, успешно описывающего текущую и предсказывающего будущую академическую успешность обучающегося, — мы последовательно проводили валидизационные исследования согласно Стандартам образовательного и психологического тестирования [American Educational Research Association et al., 2014].

В Республике Казахстан ответственность за разработку и проведение ЕНТ возложена на Национальный центр тестирования. Центр активно сотрудничает с международными организациями и экспертами в области тестирования, такими как *The AQA Group* и *National Foundation for Educational Research* в Великобритании, *Educational Testing Service* в США, адаптируя международные стандарты и методики для повышения качества ЕНТ. Особое внимание уделяется валидности результатов тестирования по содержанию. Применительно к ЕНТ содержательная валидность отражает, насколько содержание тестов соответствует объему знаний, умений и навыков, предусмотренных Государственным общеобразовательным стандартом образования и учебными программами средней школы Республики Казахстан.

Для обеспечения и поддержания содержательной валидности ЕНТ для каждого предмета разрабатываются подробные спецификации тестов, которые детально описывают структуру теста, распределение заданий по разделам учебной программы, типы заданий, уровни сложности, а также критерии оценивания. Публи-

кация этих спецификаций на официальном сайте Национального центра тестирования[3] обеспечивает прозрачность тестирования для абитуриентов и педагогического сообщества. Тестовые задания для ЕНТ разрабатывают предметные комиссии, состоящие из высококвалифицированных учителей, преподавателей вузов и методистов. Все тесты подвергаются процедуре многоуровневой экспертной оценки на соответствие содержанию утвержденных школьных программ и учебников.

На этапе разработки тестов собираются и свидетельства валидности на основе организации ответа. Разработчики проводят когнитивные лаборатории с тестируемыми, чтобы понять, как они воспринимают вопросы, исключить двойные толкования и ошибки, убедиться, что задания не содержат культурной, гендерной и иной предвзятости. Данный процесс является итеративным и включает апробацию заданий на репрезентативных выборках учащихся.

В данной статье основное внимание уделено двум группам свидетельств валидности — на основе внутренней структуры теста и на основе связи результатов с другими переменными. Для обоснования валидности по этим группам мы провели тщательный психометрический анализ теста ЕНТ по математике. Выбор именно этого предмета обусловлен тем, что в поиске альтернативных путей экономического роста, стремясь снизить зависимость от природных ресурсов, сохранить и усилить конкурентоспособность на быстро развивающемся мировом технологическом и промышленном рынке, Казахстан сделал развитие STEM, т.е. естественных наук, технологий, инженерии и математики, национальным приоритетом [Karayev, Duisenova, 2024]. Далее была исследована прогностическая валидность результатов ЕНТ. Она критически важна для вступительного экзамена в Казахстане, так как без доказательств связи между результатами экзамена и дальнейшей успешностью обучения в вузе использование теста может привести к нежелательным социальным последствиям и неэффективному финансированию высшего образования.

Пятая группа свидетельств — на основе последствий тестирования — предполагает, как отмечалось выше, установление последствий тестирования с течением времени и находится вне фокуса данной статьи, так как требует отдельного, причем лонгитюдного, исследования.

2.1. Методология исследования

2.1.1. Инструмент

ЕНТ состоит из 120 заданий и включает три обязательных предмета: грамотность чтения (10 заданий), математическую грамотность (10 заданий) и историю Казахстана (20 заданий), а также два предмета по выбору в зависимости от специализации абитури-

[3] https://testcenter.kz

ента (по 40 заданий каждый). Например, те, кто желает получить высшее образование в области компьютерных или инженерных наук, должны сдавать математику и физику. Задания могут быть трех типов: с выбором одного или нескольких правильных ответов, на установление соответствия и контекстные задания. Все задания оцениваются дихотомически, за исключением последних 10 заданий в предметах по выбору, которые оцениваются политомически: максимальный балл — 2. Максимально возможный балл по всему тесту равен 140. Приказом Министерства науки и высшего образования Республики Казахстан установлены проходные баллы по тесту, которые варьируют от 50 до 75 в зависимости от вуза, предметов по выбору и формы обучения. При этом вузы могут устанавливать собственные проходные баллы по группам образовательных программ.

2.1.2. Процедура

ЕНТ можно сдавать пять раз в год. Основное тестирование проводится с мая по июль, и по его результатам определяются получатели государственных грантов для обучения в вузах. В течение этого периода абитуриенты могут сдавать экзамен два раза и самостоятельно записываются на определенную дату сдачи экзамена. Лучший результат засчитывается как окончательный. Остальные три сдачи ЕНТ проходят в январе, марте и августе, однако сдающие экзамен в эти периоды имеют право только на платное обучение или могут использовать попытку в качестве репетиции.

С 2021 г. ЕНТ осуществляется в компьютерном формате. Его можно проходить в 40 региональных центрах тестирования, распределенных по всей стране, на трех языках — казахском, русском и английском. Продолжительность теста составляет 240 минут, в течение которых абитуриенты могут свободно перемещаться между разделами онлайн-платформы и возвращаться к ранее пропущенным заданиям, ограничений времени на выполнение тестов по отдельным предметам не установлено. Все задания проверяются автоматически, и результаты тестирования становятся доступны сразу после завершения экзамена.

2.1.3. Данные

Проведенное исследование состоит из двух этапов. На первом этапе выполнен подробный психометрический анализ теста по математике, включая анализ параллельности вариантов. В совокупности проведенные процедуры позволили собрать свидетельства валидности на основе внутренней структуры тестов. На втором этапе свидетельства валидности устанавливались на основе связей результатов тестирования с другими переменными, в частности исследована прогностическая валидность результатов ЕНТ — общего балла и теста по математике.

Данные для исследования предоставлены Национальным центром тестирования Республики Казахстан. Для анализа психометрического качества теста по математике случайным образом отобраны восемь вариантов теста — четыре на казахском языке и четыре на русском. Данные для анализа — баллы, полученные участниками тестирования по всем заданиям теста.

Для обоснования валидности на основе анализа связей с другими переменными использовались следующие данные:

- результаты двух попыток ЕНТ основного тестирования 2024 г., включая общий балл за экзамен, а также баллы по пяти предметам, входящим в тест: по трем обязательным предметам (история Казахстана, грамотность чтения, математическая грамотность) и двум предметам по выбору — физике и математике. Все баллы представлены на шкале первичных баллов;
- результаты первой сессии у студентов, обучающихся в вузе, включая средний балл успеваемости и оценки по математике и физике;
- средний балл успеваемости (GPA) — средневзвешенный балл студента по всем дисциплинам на основе 4-балльной шкалы, который рассчитывается с учетом оценки по предмету и числа кредитов за предмет. Оценки по предметам представлены на 100-балльной шкале, принятой в вузах Казахстана;
- индивидуальная информация о студенте: пол, язык сдачи ЕНТ и некоторые характеристики школы (тип учебного заведения и его местоположение), а также информация о вузе, образовательной программе и форме обучения (платная или бюджетная).

Все данные анонимизированы, при этом каждый студент был представлен уникальным идентификатором, что позволило сопоставить его результаты ЕНТ с успеваемостью в вузе.

2.1.4. Выборка

Выборка первого этапа исследования состоит из учащихся, которые сдавали ЕНТ по отобранным восьми вариантам теста ЕНТ по математике. Выборка второго этапа исследования состоит из студентов 1-го курса вузов Казахстана, которые сдавали ЕНТ в 2024 г., отдав предпочтение в качестве предметов по выбору математике и физике, поступили в вузы и сдавали экзамен по математике в первую учебную сессию зимой 2025 г. Выборка исследования, таким образом, сформирована из нескольких частей. В первую часть вошли 29 855 студентов (74% юноши, 60% студентов из городской местности, 81% сдавали ЕНТ на казахском языке), участвовавших в основной сдаче ЕНТ в 2024 г. Вторая часть представляет всю совокупность студентов вузов Казахстана, сдававших математику в первую сессию в 2025 г., — 47 613 человек (89%

имеют экзаменационную оценку по математике на момент сбора данных). Объединенная выборка, которая используется для анализа свидетельств прогностической силы ЕНТ, состоит из пересечения этих двух выборок и включает 11 276 наблюдений. В объединенной выборке 66% юношей, 56% респондентов проживали в городской местности, 86% сдавали экзамен на казахском языке. Дополнительная описательная статистика этой выборки приведена в табл. 1.

Таблица 1. **Описательная статистика выборки исследования**

Переменная	*n*	Среднее	SD	Минимум	Максимум
Балл ЕНТ по пяти предметам	11 276	72,68	20,64	24	138,00
ЕНТ «История Казахстана»	11 276	11,04	3,14	2	20,00
ЕНТ «Читательская грамотность»	11 276	8,19	1,48	0	10,00
ЕНТ «Математическая грамотность»	11 276	6,85	2,07	0	10,00
ЕНТ «Физика»	11 276	20,54	8,28	3	50,00
ЕНТ «Математика»	11 276	26,05	10,29	1	50,00
GPA в вузе	11 276	2,59	0,73	0	3,94
Экзамен в вузе по математике	10 441	68,95	21,50	0	100,00

2.1.5. Анализ

2.1.5.1. Свидетельства на основе внутренней структуры

Качество тестовых заданий и тестов в целом играет решающую роль в обеспечении надежных и валидных измерений в экзаменах с высокими ставками. Именно поэтому при обосновании валидности измерений ЕНТ мы уделили особое внимание анализу психометрического качества теста.

Анализ проводился в рамках классической теории тестирования [Algina, Penfield, 2009] и в рамках современной теории тестирования (*Item Response Theory*, IRT) [De Ayala, 2018]. Применение классической теории тестирования позволило сделать начальные выводы о качестве теста и включало первичный анализ заданий всех вариантов теста: оценены их трудность и дискриминативность, а также надежность каждого варианта. Кроме того, оценивалась эффективность дистракторов — неверных вариантов ответа в заданиях с выбором одного или нескольких правильных ответов из предложенных. Классический анализ проводился с помощью программы *Jamovi*[4] и пакета CTT в R [Sheng, 2019].

Более глубокий анализ осуществлен в рамках Раш-моделирования современной теории тестирования. Для дихотомических заданий использовалась дихотомическая модель Раша [Wright, Stone, 1979], а для заданий, оцениваемых политомически, — ее расширение, модель частичного оценивания Partial Credit Model (PCM) [Wright, Masters, 1982]. Эти модели фиксируют взаимосвязь между измеряемой латентной характеристикой испытуемо-

[4] https://www.jamovi.org

го (в нашем случае уровнем подготовленности) и вероятностью того, что он правильно выполнит задание (или получит определенный балл по заданию). Применение моделей позволяет оценить трудности заданий, а также отдельных их шагов для политомических заданий, и уровни подготовленности испытуемых на одной метрической шкале с указанием ошибки измерения каждого параметра. Выбор моделей семейства Раша обусловлен практическими соображениями. Это простейшие модели IRT, широко применяемые в педагогическом тестировании, предлагающие хорошо разработанный аппарат для анализа тестовых заданий и ответных категорий, а также анализа теста в целом — его размерности и надежности. Для проведения анализа использовалось программное обеспечение *Winsteps* версии 5.2.3.0[5].

В рамках моделей Раша проведен анализ каждого варианта теста, и он включал следующие этапы:

- *анализ соответствия данных модели*. Для оценки степени соответствия данных модели Раша мы использовали общепринятую в Раш-моделировании статистику, основанную на стандартизированных остатках, которые представляют собой взвешенную разницу между наблюдаемым и ожидаемым в рамках модели ответом (в терминах выходных данных *Winsteps* INFIT MNSQ). Эта статистика принимает значения, близкие к 1, если задание находится в хорошем согласии с моделью Раша. Значения в интервале (0,8; 1,2) считаются приемлемыми для тестов с высокими ставками [Smith, 2000];
- *исследование размерности*. Тест является одномерным, если все его задания нацелены на измерение одного конструкта. Для исследования размерности теста по математике мы использовали анализ главных компонент (*Principal Component Analysis*, PCA) стандартизированных остатков. Теоретически, если тест одномерный, остатки должны быть случайным шумом и корреляции между остатками должны быть близки к нулю. В этом случае PCA стандартизированных остатков должен генерировать собственные значения меньше 2, а распределение дисперсии между компонентами должно быть равномерным [Smith, 2002];
- *анализ заданий теста*. На данном этапе мы проанализировали психометрические характеристики отдельных заданий теста. Для политомических заданий мы исследовали также качество функционирования ответных категорий;
- *анализ надежности*. Для анализа надежности мы использовали индекс надежности, который по значению и интерпретации близок к классической надежности. Дополнительно мы использовали альтернативную статистику — индекс разделя-

[5] https://www.winsteps.com

ющей способности теста (*Person Separation Index*), который сравнивает «истинное» стандартное отклонение мер испытуемых (урегулированное на ошибку измерения) с их ошибкой измерения. Данный индекс можно использовать для расчета количества страт — статистически различных уровней мер испытуемых (отделенных на 3 ошибки измерения) [Smith, 2001];

- *оценка трудности теста*. Чтобы выяснить, насколько данный тест оказался трудным для тестируемых, мы построили карту переменных [Wright, Stone, 1979], которая показывает относительное распределение заданий и испытуемых на общей шкале;
- *справедливость измерений.* Задания могут по-разному функционировать (*Differential Item Functioning*, DIF) на разных группах тестируемых, например у юношей и девушек. Такие задания могут дискриминировать одну из групп, создавая угрозу справедливости оценивания. DIF возникает, когда учащиеся с одинаковым уровнем измеряемой черты имеют разную вероятность правильного ответа на задание. В данном исследовании использовался метод *Mantel — Haenszel*, один из наиболее популярных способов обнаружения заданий с DIF [Dorans, 1989]. Суть метода состоит в вычислении обычной статистики χ^2 для трехмерной таблицы сопряженности, размерностями которой служат балл по тесту, принадлежность к группе и результат выполнения задания.

После Раш-моделирования проведен *анализ параллельности вариантов.* В экзаменах с высокими ставками, как правило, используется много вариантов заданий. Они разрабатываются на основе одной спецификации и предполагаются параллельными: результат тестируемого не должен зависеть от того, какой вариант ему достался. Параллельность вариантов оценивается статистическими методами. В данном исследовании использовался критерий χ^2 однородности выборок, состоящий в сравнении распределений первичных баллов разных вариантов. Дополнительно мы исследовали распределения баллов по каждому варианту теста, а также сопоставили средние трудности и дискриминативности по каждому варианту.

2.1.5.2. Свидетельства на основе связей с другими переменными: анализ прогностической валидности

На втором этапе анализа исследовалась прогностическая валидность результатов ЕНТ. О наличии свидетельств прогностической валидности ЕНТ можно говорить, если баллы ЕНТ — общий балл по сумме пяти предметов и отдельно балл по математике — будут предсказывать дальнейшую успеваемость студентов в вузе. Для анализа использованы GPA студентов за первый семестр в вузе, балл за экзамен по математике за первый семестр, а также предикторы: общий балл ЕНТ, балл по математике, пол студен-

та, язык сдачи ЕНТ, сельский или городской район школы. Метод анализа — иерархическое линейное моделирование (*Hierarchical Linear Modeling*, HLM), позволяющее учесть случайный эффект на уровне вуза. Анализ выполнен с помощью пакета *lme4*[6].

Исходя из структуры данных и объема пропущенных значений, для построения моделей применялось полное удаление строк с пропущенными значениями (*listwise deletion*). Такой подход допустим, поскольку выборка остается достаточно большой для анализа (более 11 тыс. наблюдений), а доля пропущенных данных среди ключевых предикторов — низкая (не более 5–10%). Все модели строились на подмножествах данных, очищенных от пропущенных значений в используемых переменных.

Для оценки значимости предикторов в объяснении академической успеваемости студентов построены четыре блока моделей, в которых в качестве итоговой (зависимой) переменной выступает GPA (блок моделей 1) и экзамен по математике (блок моделей 2), а основным предиктором является балл ЕНТ по математике, и аналогичные модели, где основным предиктором выступает общий балл ЕНТ по пяти предметам (блоки моделей 3 и 4 соответственно). Значимость предикторов оценивалась с помощью критерия Вальда (*Wald test*), применение которого позволило определить, существенно ли улучшают предсказательную силу модели различные предикторы, учитывая иерархическую структуру данных.

2.2. Результаты

2.2.1. Обоснование психометрического качества инструментария ЕНТ: свидетельства валидности на основе внутренней структуры

В табл. 2 приведены обобщенные результаты анализа восьми вариантов теста по математике в рамках классической теории тестирования. Размер выборки варьирует от 335 до 348 человек для казахского языка и от 111 до 134 человек для русского. Такое соотношение вполне соответствует реалиям совокупной популяции абитуриентов: большинство из них сдавали экзамен на казахском языке. Средний балл по тесту варьирует от 24,9 до 28,5 при максимально возможном 50, стандартное отклонение (SD) — от 10,8 до 12,3. Такие баллы означают, что в целом тест оказался достаточно легким и на казахском языке, и на русском, что подтверждается и показателем средней трудности теста, значения которого находятся в диапазоне от 0,62 до 0,71. Все варианты характеризуются достаточно высокой средней дискриминативностью — от 0,43 до 0,51, при этом число заданий в вариантах с низкой дискриминативностью (< 0,1) не более двух. Показатель надежности — коэффициент альфа Кронбаха — имеет высокие значения для всех вариантов и составляет от 0,91 до 0,93. Таким образом, все варианты теста по математике обладают высокой внутренней согласованностью и являются надежными инструментами измерения.

[6] https://cran.r-project.org/package=lme4

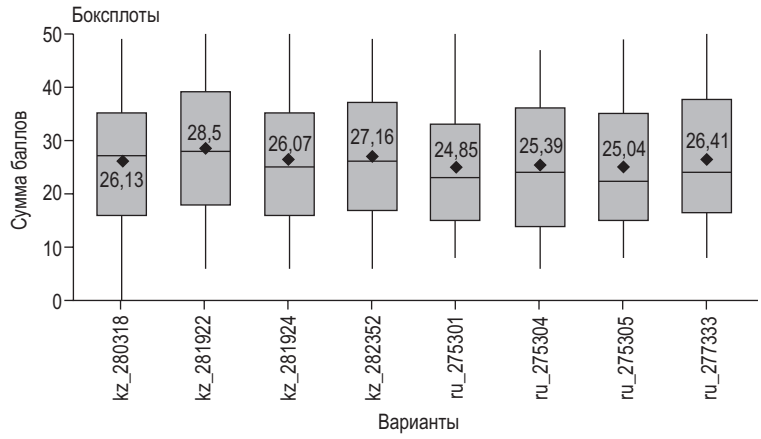
Таблица 2. **Обобщенные результаты анализа вариантов теста по математике в классической теории тестирования**

Вариант	Язык	Число участников	Средний балл (SD)	Надежность (альфа Кронбаха)	Средняя трудность	Средняя дискриминативность	Число заданий с дискриминативностью <0,1
281922	Казахский	335	28,5 (12)	0,93	0,71	0,51	0
281924	Казахский	348	26,1 (11,5)	0,93	0,65	0,48	1
282352	Казахский	341	27,2 (11,2)	0,92	0,68	0,45	1
280318	Казахский	337	26,1 (11,3)	0,92	0,65	0,44	2
297333	Русский	111	26,4 (11,9)	0,93	0,66	0,47	0
275301	Русский	115	24,9 (10,8)	0,91	0,62	0,43	1
275304	Русский	125	25,4 (12,3)	0,93	0,64	0,49	2
275305	Русский	134	25 (11,8)	0,93	0,63	0,48	1

Дополнительный анализ дистракторов показал, что у небольшого числа заданий (от 1 до 6 по разным вариантам теста) есть неэффективные дистракторы. Такой показатель не является критическим, но свидетельствует о необходимости дальнейшей работы над заданиями.

Из табл. 2 видно, что все показатели для всех вариантов достаточно близки. На рис. 1 представлены распределения первичных баллов для каждого варианта теста, а на рис. 2 — сравнение средних значений трудности и дискриминативности по вариантам. Близость всех показателей позволяет предположить, что варианты параллельны как внутри одной языковой группы, так и между языками, что и было проверено.

Рис. 1. **Сравнение распределения первичных баллов по вариантам**



Примечание: График представляет собой боксплоты («усиковая диаграмма»), где серый прямоугольник – это межквартильный размах значений, срединной линией отмечена медиана, «усы» показывают значения от минимального до максимального, и черным ромбом обозначены средние для каждого варианта.

Рис. 2. **Сравнение классических показателей вариантов**

Чтобы оценить параллельность тестовых вариантов, мы проверили однородность распределений первичных баллов, полученных участниками, по заранее заданным пяти интервалам: 0–10, 11–20, 21–30, 31–40 и 41–50 баллов. Все первичные баллы тестируемых были отнесены к соответствующим интервалам, после чего для каждого варианта теста построена частотная таблица распределения по этим интервалам. Сопоставление распределений осуществлялось с использованием критерия согласия χ^2 для таблицы сопряженности. Полученное значение критерия ($\chi^2 = 33$; $df = 28$; $p = 0{,}221$) не выявило статистически значимых различий между вариантами — следовательно, тестовые варианты можно считать параллельными по уровню трудности и структуре распределения баллов.

Таким образом, проведенный анализ дает основания считать все восемь вариантов теста по математике надежными инструментами измерения с точки зрения классической теории тестирования. Более того, варианты могут быть признаны параллельными.

Хорошее психометрическое качество теста подтверждается и результатами анализа в рамках современной теории тестирования, представленными в табл. 3.

Почти все задания находятся в согласии с моделью и имеют удовлетворительные психометрические характеристики, однако в каждом варианте есть несколько заданий, находящихся в неудовлетворительном согласии с моделью (столбец 4 в табл. 3). Эти задания были проанализированы. В частности, для них построены модельные и эмпирические характеристические кривые, показывающие ожидаемое и реальное распределение ответов

Таблица 3. **Обобщенные результаты анализа вариантов теста по математике в IRT**

Номер варианта	Язык	Число участников	Число заданий *misfit*	Надежность (*Person Reliability*)	Индекс разделяющей способности (*Person Separation*)	*Real RMSE*	*Mean Persons Measure* (SD)	Размерность: собственные значения (дисперсия)	Число заданий с DIF
1	2	3	4	5	6	7	8	9	10
280318	Казахский	337	3	0,91	3,14	0,38	0,02 (1,18)	3,2 (8,1%)	0
281922	Казахский	335	4	0,91	3,12	0,44	0,56 (1,36)	2,9 (7,2%)	0
281924	Казахский	348	3	0,91	3,27	0,4	0,27 (1,32)	2,9 (7,4%)	0
282352	Казахский	341	3	0,91	3,19	0,38	0,27 (1,21)	2,4 (5,9%)	0
275301	Русский	115	2	0,9	2,97	0,39	0,02 (1,16)	2,6 (6,4%)	0
275304	Русский	125	3	0,92	3,32	0,37	0,09 (1,22)	2,5 (6,3%)	0
275305	Русский	134	3	0,92	3,32	0,38	0,08 (1,27)	2,6 (6,5%)	1
277333	Русский	111	3	0,9	2,98	0,4	0,15 (1,2)	2,6 (6,6%)	0

испытуемых. Все эти задания оказались очень трудными (процент выполнения для большинства из них не более 10%), и их плохое согласие с моделью, вероятно, объясняется тем, что экзаменуемые давали ответы наугад. В целом абсолютное большинство заданий находятся в хорошем согласии с моделью.

Индекс надежности принимает значения в диапазоне от 0,9 до 0,92, а индекс разделяющей способности теста варьирует от 2,97 до 3,32. Вычисляя количество страт, т.е. статистически различных уровней мер испытуемых, получаем, что все варианты теста позволяют разделить участников на четыре группы с разным уровнем подготовленности. Полученный показатель является высоким.

Результаты исследования размерности теста с помощью PCA стандартизированных остатков представлены в столбце 9 табл. 3. Собственные значения для первого компонента матрицы корреляции остатков варьируют от 2,4 до 3,2, т.е. немного превышают рекомендуемый порог — 2. Кроме того, на первый компонент приходится несколько больший процент дисперсии — от 5,9 до 8,1%. Анализ заданий, имеющих наибольшие по модулю нагрузки по первому компоненту, показал, что задания выделяются в группы по тематическому принципу и формату заданий. Также мы проанализировали скорректированную на ошибку измерения корреляцию между кластерами заданий в остатках (*disattenuated correlation*) для каждого варианта теста и установили, что все показатели находятся в диапазоне 0,82–1,0 и в одном случае составляют 0,74, что в соответствии с рекомендациями *Winsteps Manual*[7] сви-

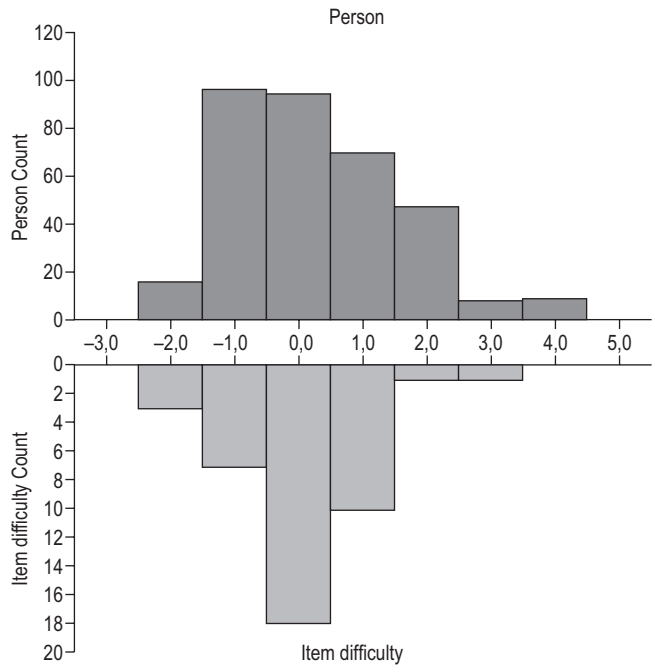
[7] https://www.winsteps.com/winman/table23_1.htm

детельствует об отсутствии необходимости выделения нескольких размерностей. Учитывая результаты анализа, можно сделать вывод, что для всех вариантов теста угрозой неодномерности можно пренебречь.

В последнем столбце табл. 3 показано число заданий, характеризующихся гендерным DIF, и только одно задание функционирует так, что дает преимущество девушкам.

На рис. 3 представлена карта переменных одного из вариантов теста, показывающая относительное распределение заданий и испытуемых на общей метрической шкале. На карте испытуемые представлены в верхней части, а задания — в нижней. Более трудные задания и учащиеся с высокими результатами расположены в правой части карты, а более легкие задания и учащиеся с низкими результатами — в левой. Учащиеся распределены широко — это значит, что тест способен дифференцировать учащихся с разным уровнем подготовки. Анализ распределения заданий относительно выборки испытуемых показывает, что тест хорошо центрирован по трудности, хотя наблюдается недостаток трудных заданий, подходящих для учащихся с высоким уровнем подготовки, что, впрочем, не является проблемой ввиду небольшого числа таких учащихся в выборке.

Рис. 3. **Карта переменных**



Обобщая результаты психометрического анализа теста ЕНТ по математике, можно сделать вывод, что тест является качествен-

ным измерительным инструментом: все проанализированные варианты теста могут быть признаны существенно одномерными, имеют высокую надежность, хорошо подходят целевой аудитории по трудности, абсолютное большинство заданий имеет хорошее психометрическое качество, варианты теста близки между собой по основным характеристикам и могут считаться параллельными.

2.2.2. Прогностическая способность экзамена: свидетельства валидности на основе связей с другими переменными

Чтобы оценить, насколько хорошо баллы ЕНТ предсказывают успеваемость студентов в вузе, мы работали с доступными обширными данными на выборке, описанной в табл. 1. Результаты анализа значимости всех включенных предикторов для каждого блока рассматриваемых моделей представлены в табл. 4.

Таблица 4. **Оценка вклада предикторов с помощью теста Вальда**

Переменные	GPA в вузе		Экзамен по математике в вузе	
Предикторы	*Модель 1.1* (χ^2, *p*)	*Модель 1.2* (χ^2, *p*)	*Модель 2.1* (χ^2, *p*)	*Модель 2.2* (χ^2, *p*)
(Интерцепт)	1013; $p < 0{,}001$	1190,2; $p < 0{,}001$	1263; $p < 0{,}001$	1236,9; $p < 0{,}001$
ЕНТ «Математика»	1515; $p < 0{,}001$	1339,9; $p < 0{,}001$	1183, $p < 0{,}001$	1080,7; $p < 0{,}001$
Пол (мужской)	–	433,1; $p < 0{,}001$	–	146,4; $p < 0{,}001$
Язык (казахский)	–	13,17; $p = 0{,}0003$	–	0,16; $p = 0{,}69$
Район	–	0,99; $p = 0{,}319$	–	2,60; $p = 0{,}11$
Предикторы	*Модель 3.1* (χ^2, *p*)	*Модель 3.2* (χ^2, *p*)	*Модель 4.1* (χ^2, *p*)	*Модель 4.2* (χ^2, *p*)
(Интерцепт)	604; $p < 0{,}001$	783,1; $p < 0{,}001$	601; $p < 0{,}001$	693,67; $p < 0{,}001$
ЕНТ 5 предметов	1850; $p < 0{,}001$	1735,1; $p < 0{,}001$	1279; $p < 0{,}001$	1196,30; $p < 0{,}001$
Пол (мужской)	–	453,5; $p < 0{,}001$	–	156,12; $p < 0{,}001$
Язык (казахский)	–	52,47; $p < 0{,}001$	–	12,89; $p < 0{,}001$
Район	–	1,54; $p = 0{,}21$	–	1,76; $p = 0{,}18$

Во всех моделях баллы ЕНТ — и только по математике (модели 1.1 — 2.2), и общий балл по сумме пяти предметов (модели 3.1 — 4.2) — оказались значимым предиктором как общей вузовской успеваемости, так и балла за первый экзамен по математике в вузе. В моделях 1.2, 2.2, 3.2, 4.2 включение пола, родного языка и типа населенного пункта значимо улучшило результаты.

Для интерпретации относительной силы предикторов ниже приведены стандартизованные коэффициенты для всех моделей. Таблица 5 позволяет оценить вклад каждого предиктора, выраженный в единицах стандартного отклонения, и напрямую оценить их связь с результирующей переменной — общей успеваемостью в вузе или баллом по дисциплине.

Таблица 5. **Стандартизованные коэффициенты переменных для моделей в четырех блоках**

Переменная	Std. coef. (GPA) *Модель 1.2*	95% ДИ	Std. coef. (Экзамен) *Модель 2.2*	95% ДИ
(Интерцепт)	0,17	[0,01; 0,34]	0,21	[0,08; 0,34]
ЕНТ «Математика»	0,34	[0,32; 0,35]	0,39	[0,37; 0,41]
Пол (мужской)	–0,15	[–0,16; –0,13]	–0,11	[–0,12; –0,09]
Язык (казахский)	–0,02	[–0,04; –0,01]	–0,003	[–0,02; 0,01]
Район	0,007	[–0,01; 0,02]	–0,01	[–0,03; 0,00]
Переменная	Std. coef. (GPA) *Модель 3.2*	95% ДИ	Std. coef. (Экзамен) *Модель 4.2*	95% ДИ
(Интерцепт)	0,20	[0,02; 0,37]	0,23	[0,09; 0,37]
ЕНТ (5 предметов)	0,40	[0,38; 0,41]	0,43	[0,41; 0,46]
Пол (мужской)	–0,15	[–0,16; –0,14]	–0,11	[–0,13; –0,09]
Язык (казахский)	–0,05	[–0,06; –0,04]	–0,03	[–0,05; –0,01]
Район	0,008	[0,00; 0,02]	–0,01	[–0,03; 0,01]

Судя по стандартизованным коэффициентам, наиболее сильным (среди рассмотренных) предиктором академических результатов как в модели с GPA (β = 0,40), так и в модели с экзаменом по математике в качестве итоговых переменных (β = 0,43) является общий балл ЕНТ. Балл ЕНТ только по математике тоже достаточно сильный предиктор: и для модели с GPA (β = 0,34), и для модели с экзаменом в качестве итоговых переменных (β = 0,39). Пол (мужчины = 1) имеет статистически значимый и негативный эффект во всех моделях. Влияние других демографических переменных (язык обучения и регион) близко к нулю (казахский язык = 1) или незначимо (городской район vs сельский).

Для оценки вклада каждого блока предикторов проведены также тесты отношения правдоподобия (*Likelihood Ratio Test*, LRT) (табл. 6). Добавление демографических переменных (модель .1→ модель .2) статистически значимо улучшает результаты для всех четырех блоков моделей.

Таблица 6. **Сравнение моделей по LRT**

Сравнение моделей	χ^2	*df*	*p*
GPA, ЕНТ «Математика»: Модель 1.1 vs Модель 1.2	433	3	< 0,001
Экзамен, ЕНТ «Математика»: Модель 2.1 vs Модель 2.2	149	3	< 0,001
GPA, ЕНТ 5: Модель 3.1 vs Модель 3.2	488	3	< 0,001
Экзамен, ЕНТ 5: Модель 4.1 vs Модель 4.2	167	3	< 0,001

Для оценки общего качества моделей рассчитаны коэффициенты R^2 (табл. 7). Маржинальный R^2 отражает долю дисперсии, объясняемую фиксированными эффектами; условный R^2 — долю дисперсии с учетом случайных эффектов (вузы).

Таблица 7. **Объясненная дисперсия в моделях**

Модель (итоговая переменная, основной предиктор)	Marginal R^2	Conditional R^2
Модель 1.1 (GPA, ЕНТ «Математика»)	0,132	0,515
Модель 1.2 (GPA, ЕНТ «Математика», демографические переменные)	0,165	0,536
Модель 2.1 (Экзамен, ЕНТ «Математика»)	0,146	0,341
Модель 2.2 (Экзамен, ЕНТ «Математика», демографические переменные)	0,163	0,360
Модель 3.1 (GPA, ЕНТ 5)	0,165	0,552
Модель 3.2 (GPA, ЕНТ 5, демографические переменные)	0,200	0,576
Модель 4.1 (Экзамен, ЕНТ 5)	0,169	0,382
Модель 4.2 (Экзамен, ЕНТ 5, демографические переменные)	0,186	0,401

В совокупности полученные результаты свидетельствуют о том, что балл ЕНТ по математике и общий балл ЕНТ по пяти предметам являются устойчивым и достаточно сильным предиктором академической успеваемости студентов при дальнейшем обучении в вузах Казахстана.

2.3. Дискуссия

Настоящее исследование имеет целью обосновать валидность результатов Единого национального тестирования в Казахстане на основе теории валидности С. Мессика. Согласно международно признанным Стандартам образовательного и психологического тестирования [American Educational Research Association et al., 2014] для доказательства валидности той или иной интерпретации результатов тестирования используются пять основных свидетельств валидности: по содержанию, по процессу порождения ответа на тест, по внутренней структуре, по связям с другими переменными, а также по последствиям тестирования. В рамках данной статьи мы сосредоточились на двух критически важных группах свидетельств валидности: мы подробно проанализировали внутреннее психометрическое качество экзамена ЕНТ по математике, а также оценили, насколько хорошо ЕНТ прогнозирует успешное обучение студентов в вузах Казахстана.

Качество тестовых заданий и тестов в целом играет решающую роль в обеспечении надежности и валидности измерений в экзаменах с высокими ставками. Именно поэтому при обосновании валидности измерений ЕНТ мы уделили особое внимание анализу психометрического качества теста. Для анализа случайным образом отобраны восемь вариантов теста по математике: четыре варианта на казахском языке и четыре на русском. Выбор в качестве объекта исследования именно теста по математике обусловлен тем, что, стремясь адаптировать свою образовательную политику под вызовы стремительной глобализации и цифро-

визации мира, Казахстан сделал национальным приоритетом развитие STEM-образования [Karayev, Duisenova, 2024].

Анализ проводился как в рамках классической теории тестирования, так и в рамках современной теории тестирования. Проведенное исследование показало, что все проанализированные варианты заданий имеют высокую надежность — выше 0,9, что сравнимо с надежностью тестов SAT и ACT [Rothstein, 2004, Woodruff, Wu, 2012]. Тест ЕНТ по математике является существенно одномерным и хорошо подходит целевой аудитории по уровню трудности, обеспечивает справедливое оценивание. Абсолютное большинство заданий имеют хорошее качество. Все варианты заданий близки между собой по основным характеристикам, и их можно считать параллельными. Обобщая результаты психометрического анализа теста ЕНТ по математике, можно сделать вывод, что тест является качественным измерительным инструментом. Лишь несколько заданий нуждаются в пересмотре, так как слишком трудны и провоцируют ответы наугад.

Критически важным доказательством валидности для экзаменов с высокими ставками, ориентированных на переход между средним и высшим образованием, является прогностическая валидность, так как именно она подтверждает, что экзамен выполняет свою основную функцию — предсказывает успеваемость в вузе [Kobrin et al., 2008]. Исследования старейших стандартизированных экзаменов, например SAT, показывают, что в разные годы, на разных выборках, с использованием разных статистических подходов и разных наборов предикторов этот экзамен всегда значимо предсказывает успешность студентов в колледже [Amrein-Beardsley et al., 2025; Clark, Rothstein, Schanzenbach, 2008; Fishman, Pasanella, 1960; Zwick, 2019].

Исследование показало, что баллы ЕНТ значимо и положительно коррелируют с результатами дальнейшего обучения студентов в вузах Казахстана. При этом качество используемых нами статистических моделей и объясняемая ими дисперсия оценок студентов за первый семестр во многом согласуются с результатами исследований, проведенных в других странах, включая Россию [Хавенсон, Соловьева, 2014] и США [Sanchez, 2025]. Мы исследовали четыре группы моделей: в качестве результирующей переменной мы использовали как общую среднюю успешность студентов за первый семестр обучения в вузе (GPA), так и экзамен по профильной математической дисциплине в первую сессию, а в качестве предикторов оценивали эффект как отдельно балла ЕНТ по математике, так и общего балла ЕНТ. Установлено, что предсказательная сила общего балла ЕНТ выше, и он схожим образом прогнозирует как общую успешность в вузе, так и успешность по профильному предмету.

Судя по имеющимся в литературе данным, введение дополнительных предикторов, например пола студентов, национальности или родного языка, показателя СЭС, а также средней школьной успеваемости, позволяет сильно увеличить объясняющую силу моделей. Так, комбинированные баллы ACT и GPA наиболее точно предсказывают успеваемость на первом курсе, которая, в свою очередь, является ключевым предиктором продолжения обучения [Westrick et al., 2015]. При этом социально-экономический статус показал статистически значимую, хотя слабую прогностическую силу.

Мы не располагали данными о прошлой школьной успеваемости обучающихся в Казахстане, но включили в анализ некоторые социально-демографические характеристики и показали, что они тоже значимо улучшают качество моделей. В частности, оказалось, что результаты юношей значимо ниже результатов девушек во всех моделях. В то же время переменная местности, в которой проживал ученик (сельская или городская), не имеет значимого эффекта с точки зрения его будущих результатов в вузе. Язык сдачи экзаменов (русский или казахский) либо имеет крайне малый практический эффект для первой сессии студентов (сотые доли), либо незначим.

3. Заключение

Централизованные вступительные экзамены в вузы как экзамены с высокими ставками вызывают пристальное внимание исследователей. Эти экзамены призваны обеспечивать единую метрику для оценки абитуриентов с разным образовательным опытом, позволяя приемным комиссиям вузов принимать более обоснованные решения [Geiser, Santelices, 2007]. Централизованные экзамены используются во многих странах мира, при этом могут применяться разные модели экзамена, но они должны неизменно обеспечивать валидность результатов экзамена и принимаемых на его основе решений.

В данной статье представлено первое на пространстве СНГ комплексное исследование валидности экзамена с высокими ставками, выполненное с соблюдением строгой теоретической рамки проведения валидизации. Полученные данные позволяют сделать вывод, что экзамен ЕНТ по математике обладает высоким психометрическим качеством. При этом общий балл ЕНТ и отдельно балл по математике обладают достаточно высокой прогностической валидностью, предсказывая будущую успешность обучения в вузе. Для Казахстана этот вывод очень значим, так как по результатам ЕНТ выдаются государственные гранты на получение высшего образования.

Благодарности Авторы выражают благодарность Национальному центру тестирования Республики Казахстан за финансовую поддержку в рамках внутреннего гранта.

Литература

1. Абдрасилов Б., Алтыбаева Ш., Шинетова Л. (2025) Результаты исследования системы ЕНТ, проведенного в рамках проекта Всемирного банка. *Педагогические измерения*, т. 1, № 1, сс. 6–20.
2. Смагулова А., Сатанов А., Кадирова Ф. (2025) Анализ Единого национального тестирования в рамках исследования индекса благополучия детей в Казахстане. *Педагогические измерения*, т. 1, № 1, сс. 47–63.
3. Хавенсон Т., Соловьева А. (2014) Связь результатов Единого государственного экзамена и успеваемости в вузе. *Вопросы образования / Educational Studies Moscow*, № 1, сс. 176–199. https://doi.org/10.17323/1814-9545-2014-1-176-199
4. ACT (2024) *ACT® Technical Manual*. Available at: https://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.pdf (accessed 06.11.2025).
5. Aguinis H., Culpepper S.A., Pierce C.A. (2016) Differential Prediction Generalization in College Admissions Testing. *Journal of Educational Psychology*, vol. 108, no 7, pp. 1045–1059. https://doi.org/10.1037/edu0000104
6. Algina J., Penfield R.D. (2009) Classical Test Theory. *The SAGE Handbook of Quantitative Methods in Psychology* (eds R.E. Millsap, A. Maydeu-Olivares), Los Angeles; London; New Delhi: SAGE, pp. 93–122. https://doi.org/10.4135/9780857020994.n5
7. Amangeldiyeva B. (2024) *Unveiling Gender Disparities in Kazakhstani Education: A Comprehensive Analysis using Unified National Testing Results* (Master's thesis). Vienna: Central European University.
8. American Educational Research Association et al. (2014) *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
9. Amrein-Beardsley A., Azizova Z.T., Gibbs N.P., Ikegwuonu C., Kim J., La Torre D.M. et al. (2025) A Validation Review of the SAT and ACT for College and University Admissions Decisions. *Education Policy Analysis Archives*, vol. 33, Article no 28. https://doi.org/10.14507/epaa.33.8734
10. Atkinson R.C., Geiser S. (2009) Reflections on a Century of College Admissions Tests. *Educational Researcher*, vol. 38, no 9, pp. 665–676. https://doi.org/10.3102/0013189X0935198
11. Bai C., Chi W., Qian X. (2014) Do College Entrance Examination Scores Predict Undergraduate GPAs? A Tale of Two Universities. *China Economic Review*, vol. 30, September, pp. 632–647. https://doi.org/10.1016/j.chieco.2013.08.005
12. Bethell G., Zabulionis A. (2012) The Evolution of High-Stakes Testing at the School–University Interface in the Former Republics of the USSR. *High-Stakes Testing in Education. Value, Fairness and Consequences* (eds T. Eggen, G. Stobart), London: Routledge, pp. 7–25. https://doi.org/10.1080/0969594X.2011.635591
13. Burton N.W., Ramist L. (2001) *Predicting Success in College: SAT® Studies of Classes Graduating since 1980. College Board Research Report no 2001-2.* Available at: https://files.eric.ed.gov/fulltext/ED562836.pdf (accessed 07.11.2025).
14. Clark M., Rothstein J., Schanzenbach D.W. (2008) *Selection Bias in College Admissions Test Scores. National Bureau of Economic Research Working Paper no 14265*. https://doi.org/10.3386/w14265
15. College Board (2025) *Validity of the SAT Suite of Assessments*. Available at: https://research.collegeboard.org/reports/sat-suite/validity (accessed 06.11.2025).

16. De Ayala R.J. (2018) Item Response Theory and Rasch Modeling. *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (eds G.R. Hancock, L.M. Stapleton, R.O. Mueller), New York, NY: Routledge, pp. 145–163. https://doi.org/10.4324/9780203861554
17. Dorans N.J. (1989) Two New Approaches to Assessing Differential Item Functioning: Standardization and the Mantel — Haenszel Method. *Applied Measurement in Education*, vol. 2, no 3, pp. 217–233. https://doi.org/10.1207/s15324818ame0203_3
18. Fishman J.A., Pasanella A.K. (1960) College Admission Selection Studies. *Review of Educational Research*, vol. 30, no 4, pp. 298–310.
19. Garg N., Li H., Monachou F. (2020) Dropping Standardized Testing for Admissions Trades Off Information and Access. *arXiv preprint* arXiv:2010.04396. https://doi.org/10.48550/arXiv.2010.04396
20. Geiser S., Santelices M.V. (2007) *Validity of High-School Grades in Predicting Student Success beyond the Freshman Year: High-School Record vs. Standardized Tests as Indicators of Four-Year College Outcomes. Research & Occasional Paper Series: CSHE.6.07*. Berkeley, CA: Center for Studies in Higher Education, University of California.
21. Hajar A., Abenova S. (2021) The Role of Private Tutoring in Admission to Higher Education: Evidence from a Highly Selective University in Kazakhstan. *Hungarian Educational Research Journal*, vol. 11, no 2, pp. 124–142. https://doi.org/10.1556/063.2021.00001
22. Karayev Z., Duisenova R. (2024) Transformation of the Education System Based on the STEM Approach as a Condition for Preparing Competitive Human Capital in the Modern World. Proceedings of the *8th International Scientific Conference "Scientific Results" (Rome, 2024, 7–8 November),* pp. 142–152.
23. Kobrin J.L., Patterson B.F., Shaw E.J., Mattern K.D., Barbuti S.M. (2008) *Validity of the SAT® for Predicting First-Year College Grade Point Average. College Board Research Report no 2008-5.* Available at: https://files.eric.ed.gov/fulltext/ED563202.pdf (accessed 07.11.2025).
24. Lane S. (2014) Validity Evidence Based on Testing Consequences. *Psicothema*, vol. 26, no 1, pp. 127–135. https://doi.org/10.7334/psicothema2013.258
25. Linn R.L. (2009) Considerations for College Admissions Testing. *Educational Researcher*, vol. 38, no 9, pp. 677–679. https://doi.org/10.3102/0013189X09351982
26. Marini J.P., Westrick P.A., Young L., Shaw E.J. (2020) *Validity of the SAT® for Enrollment-Related Decisions: Focus on International Students Attending College in the US.* Available at: https://research.collegeboard.org/media/pdf/validity-sat-enrollment-related-decisions.pdf (accessed 06.11.2025).
27. Messick S. (1994) The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, vol. 23, no 2, pp. 13–23. https://doi.org/10.3102/0013189X023002013
28. Messick S., Jungeblut A. (1981) Time and Method in Coaching for the SAT. *Psychological Bulletin*, vol. 89, no 2, pp. 191–216. https://doi.org/10.1037/0033-2909.89.2.191
29. Michaelides M.P. (2014) Validity Considerations Ensuing from Examinees' Perceptions about High-Stakes National Examinations in Cyprus. *Assessment in Education: Principles, Policy & Practice*, vol. 21, no 4, pp. 427–441. https://doi.org/10.1080/0969594X.2014.916655
30. Migliaretti G., Bozzaro S., Siliquini R., Stura I., Costa G., Cavallo F. (2017) Is the Admission Test for a Course in Medicine a Good Predictor of Academic Performance? A Case-Control Experience at the School of Medicine of Turin. *BMJ Open*, vol. 7, no 11, Article no e017417. https://doi.org/10.1136/bmjopen-2017-017417

31. Mingisheva N. (2023) Development and Challenges of Standardized Testing in Kazakhstan: Transition from National to International Standards. *Pedagogikalyq Gylymdar Seriiasy*, vol. 76, no 3, pp. 94–103. https://doi.org/10.26577/JES.2023.v76.i3.08
32. Morgan R. (1989) Analyses of the Predictive Validity of the SAT® and High School Grades from 1976 to 1985. *ETS Research Report Series*, no 1989 (2), pp. i–16.
33. OECD (2020) *Education at a Glance 2020: OECD Indicators*. Paris: OECD. https://doi.org/10.1787/69096873-en
34. Rothstein J.M. (2004) College Performance Predictions and the SAT. *Journal of Econometrics*, vol. 121, no 1–2, pp. 297–317. https://doi.org/10.1016/j.jeconom.2003.10.003
35. Sanchez E.I. (2025) *Predicting STEM Achievement: A Comparative Study of ACT® Scores and High School GPA. ACT Research Report no R2501*. Available at: https://www.act.org/content/dam/act/unsecured/documents/R2501-Predicting-STEM-Achievement-ACT-Scores-HSGPA-02-2025.pdf (accessed 06.11.2025).
36. Shabdenova A., Satybayeva A. (2024) Analysis of the Results of the Unified National Testing in the Context of Various Characteristics of Graduates of Schools in Kazakhstan. *The Journal of Psychology & Sociology*, vol. 88, no 1, pp. 85–97. https://doi.org/10.26577/JPsS.2024.v88.i1.07
37. Sheng Y. (2019) CTT Package in R. *Measurement: Interdisciplinary Research and Perspectives*, vol. 17, no 4, pp. 211–219. https://doi.org/10.1080/15366367.2019.1600839
38. Sireci S.G. (2021) NCME Presidential Address 2020: Valuing Educational Measurement. *Educational Measurement: Issues and Practice*, vol. 40, no 1, pp. 7–16. https://doi.org/10.1111/emip.12415
39. Sireci S., Benítez I. (2023) Evidence for Test Validation: A Guide for Practitioners. *Psicothema*, vol. 35, no 3, pp. 217–226. https://doi.org/10.7334/psicothema2022.477
40. Smith E.V., Jr. (2001) Evidence for the Reliability of Measures and Validity of Measure Interpretation: A Rasch Measurement Perspective. *Journal of Applied Measurement*, vol. 2, no 3, pp. 281–311.
41. Smith E.V., Jr. (2002) Detecting and Evaluating the Impact of Multidimensionality Using Item Fit Statistics and Principal Component Analysis of Residuals. *Journal of Applied Measurement*, vol. 3, no 2, pp. 205–231.
42. Smith R.M. (2000) Fit Analysis in Latent Trait Measurement Models. *Journal of Applied Measurement*, vol. 1, no 2, pp. 199–218.
43. Stemler S.E. (2012) What Should University Admissions Tests Predict? *Educational Psychologist*, vol. 47, no 1, pp. 5–17. https://doi.org/10.1080/00461520.2011.611444
44. Westrick P.A., Le H., Robbins S.B., Radunzel J.M., Schmidt F.L. (2015) College Performance and Retention: A Meta-Analysis of the Predictive Validities of ACT® Scores, High School Grades, and SES. *Educational Assessment*, vol. 20, no 1, pp. 23–45. https://doi.org/10.1080/10627197.2015.997614
45. Woodruff D., Wu Y.F. (2012) *Statistical Considerations in Choosing a Test Reliability Coefficient. ACT Research Report no 2012-10*. Available at: https://www.act.org/content/dam/act/unsecured/documents/ACT_RR2012-10.pdf (accessed 07.11.2025).
46. Wright B.D., Masters G.N. (1982) *Rating Scale Analysis*. Chicago, IL: Mesa Press.
47. Wright B.D., Stone M.H. (1979) *Best Test Design*. Chicago, IL: Mesa Press.
48. Zwick R. (2019) Assessment in American Higher Education: The Role of Admissions Tests. *The Annals of the American Academy of Political and Social Science*, vol. 683, no 1, pp. 130–148. https://doi.org/10.1177/0002716219843469

References

Abdrasilov B., Altybaeva Sh., Shinetova L. (2025) Results of the Study of the UNT System Conducted within the World Bank Project. *Pedagogical Measurements*, vol. 1, no 1, pp. 6–20 (In Russian).

ACT (2024) *ACT® Technical Manual*. Available at: https://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.pdf (accessed 06.11.2025).

Aguinis H., Culpepper S.A., Pierce C.A. (2016) Differential Prediction Generalization in College Admissions Testing. *Journal of Educational Psychology*, vol. 108, no 7, pp. 1045–1059. https://doi.org/10.1037/edu0000104

Algina J., Penfield R.D. (2009) Classical Test Theory. *The SAGE Handbook of Quantitative Methods in Psychology* (eds R.E. Millsap, A. Maydeu-Olivares), Los Angeles; London; New Delhi: SAGE, pp. 93–122. https://doi.org/10.4135/9780857020994.n5

Amangeldiyeva B. (2024) *Unveiling Gender Disparities in Kazakhstani Education: A Comprehensive Analysis using Unified National Testing Results* (Master's thesis). Vienna: Central European University.

American Educational Research Association et al. (2014) *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Amrein-Beardsley A., Azizova Z.T., Gibbs N.P., Ikegwuonu C., Kim J., La Torre D.M. et al. (2025) A Validation Review of the SAT and ACT for College and University Admissions Decisions. *Education Policy Analysis Archives*, vol. 33, Article no 28. https://doi.org/10.14507/epaa.33.8734

Atkinson R.C., Geiser S. (2009) Reflections on a Century of College Admissions Tests. *Educational Researcher*, vol. 38, no 9, pp. 665–676. https://doi.org/10.3102/0013189X0935198

Bai C., Chi W., Qian X. (2014) Do College Entrance Examination Scores Predict Undergraduate GPAs? A Tale of Two Universities. *China Economic Review*, vol. 30, September, pp. 632–647. https://doi.org/10.1016/j.chieco.2013.08.005

Bethell G., Zabulionis A. (2012) The Evolution of High-Stakes Testing at the School–University Interface in the Former Republics of the USSR. *High-Stakes Testing in Education. Value, Fairness and Consequences* (eds T. Eggen, G. Stobart), London: Routledge, pp. 7–25. https://doi.org/10.1080/0969594X.2011.635591

Burton N.W., Ramist L. (2001) *Predicting Success in College: SAT® Studies of Classes Graduating since 1980. College Board Research Report no 2001-2*. Available at: https://files.eric.ed.gov/fulltext/ED562836.pdf (accessed 07.11.2025).

Clark M., Rothstein J., Schanzenbach D.W. (2008) *Selection Bias in College Admissions Test Scores. National Bureau of Economic Research Working Paper no 14265.* https://doi.org/10.3386/w14265

College Board (2025) Validity of the SAT Suite of Assessments. Available at: https://research.collegeboard.org/reports/sat-suite/validity (accessed 06.11.2025).

De Ayala R.J. (2018) Item Response Theory and Rasch Modeling. *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (eds G.R. Hancock, L.M. Stapleton, R.O. Mueller), New York, NY: Routledge, pp. 145–163. https://doi.org/10.4324/9780203861554

Dorans N.J. (1989) Two New Approaches to Assessing Differential Item Functioning: Standardization and the Mantel — Haenszel Method. *Applied Measurement in Education*, vol. 2, no 3, pp. 217–233. https://doi.org/10.1207/s15324818ame0203_3

Fishman J.A., Pasanella A.K. (1960) College Admission Selection Studies. *Review of Educational Research*, vol. 30, no 4, pp. 298–310.

Garg N., Li H., Monachou F. (2020) Dropping Standardized Testing for Admissions Trades Off Information and Access. *arXiv preprint* arXiv:2010.04396. https://doi.org/10.48550/arXiv.2010.04396

Geiser S., Santelices M.V. (2007) *Validity of High-School Grades in Predicting Student Success beyond the Freshman Year: High-School Record vs. Standardized Tests as Indicators of Four-Year College Outcomes. Research & Occasional Pa-*

per Series: CSHE.6.07. Berkeley, CA: Center for Studies in Higher Education, University of California.

Hajar A., Abenova S. (2021) The Role of Private Tutoring in Admission to Higher Education: Evidence from a Highly Selective University in Kazakhstan. *Hungarian Educational Research Journal*, vol. 11, no 2, pp. 124–142. https://doi.org/10.1556/063.2021.00001

Karayev Z., Duisenova R. (2024) Transformation of the Education System Based on the STEM Approach as a Condition for Preparing Competitive Human Capital in the Modern World. Proceedings of the *8th International Scientific Conference "Scientific Results" (Rome, 2024, 7–8 November),* pp. 142–152.

Khavenson T.E., Solovyova A.A. (2014) Studying the Relation between the Unified State Exam Points and Higher Education Performance. *Voprosy obrazovaniya / Educational Studies Moscow*, no 1, pp. 176–199 (In Russian). https://doi.org/10.17323/1814-9545-2014-1-176-199.

Kobrin J.L., Patterson B.F., Shaw E.J., Mattern K.D., Barbuti S.M. (2008) *Validity of the SAT® for Predicting First-Year College Grade Point Average. College Board Research Report no 2008-5.* Available at: https://files.eric.ed.gov/fulltext/ED563202.pdf (accessed 07.11.2025).

Lane S. (2014) Validity Evidence Based on Testing Consequences. *Psicothema*, vol. 26, no 1, pp. 127–135. https://doi.org/10.7334/psicothema2013.258

Linn R.L. (2009) Considerations for College Admissions Testing. *Educational Researcher*, vol. 38, no 9, pp. 677–679. https://doi.org/10.3102/0013189X09351982

Marini J.P., Westrick P.A., Young L., Shaw E.J. (2020) *Validity of the SAT® for Enrollment-Related Decisions: Focus on International Students Attending College in the US*. Available at: https://research.collegeboard.org/media/pdf/validity-sat-enrollment-related-decisions.pdf (accessed 06.11.2025).

Messick S. (1994) The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, vol. 23, no 2, pp. 13–23. https://doi.org/10.3102/0013189X023002013

Messick S., Jungeblut A. (1981) Time and Method in Coaching for the SAT. *Psychological Bulletin*, vol. 89, no 2, pp. 191–216. https://doi.org/10.1037/0033-2909.89.2.191

Michaelides M.P. (2014) Validity Considerations Ensuing from Examinees' Perceptions about High-Stakes National Examinations in Cyprus. *Assessment in Education: Principles, Policy & Practice*, vol. 21, no 4, pp. 427–441. https://doi.org/10.1080/0969594X.2014.916655

Migliaretti G., Bozzaro S., Siliquini R., Stura I., Costa G., Cavallo F. (2017) Is the Admission Test for a Course in Medicine a Good Predictor of Academic Performance? A Case-Control Experience at the School of Medicine of Turin. *BMJ Open*, vol. 7, no 11, Article no e017417. https://doi.org/10.1136/bmjopen-2017-017417

Mingisheva N. (2023) Development and Challenges of Standardized Testing in Kazakhstan: Transition from National to International Standards. *Pedagogikalyq Gylymdar Seriiasy*, vol. 76, no 3, pp. 94–103. https://doi.org/10.26577/JES.2023.v76.i3.08

Morgan R. (1989) Analyses of the Predictive Validity of the SAT® and High School Grades from 1976 to 1985. *ETS Research Report Series*, no 1989 (2), pp. i–16.

OECD (2020) *Education at a Glance 2020: OECD Indicators*. Paris: OECD. https://doi.org/10.1787/69096873-en

Rothstein J.M. (2004) College Performance Predictions and the SAT. *Journal of Econometrics*, vol. 121, no 1–2, pp. 297–317. https://doi.org/10.1016/j.jeconom.2003.10.003

Sanchez E.I. (2025) *Predicting STEM Achievement: A Comparative Study of ACT® Scores and High School GPA. ACT Research Report no R2501*. Available at: https://www.act.org/content/dam/act/unsecured/documents/R2501-Pre

dicting-STEM-Achievement-ACT-Scores-HSGPA-02-2025.pdf (accessed 06.11.2025).

Shabdenova A., Satybayeva A. (2024) Analysis of the Results of the Unified National Testing in the Context of Various Characteristics of Graduates of Schools in Kazakhstan. *The Journal of Psychology & Sociology*, vol. 88, no 1, pp. 85–97. https://doi.org/10.26577/JPsS.2024.v88.i1.07

Sheng Y. (2019) CTT Package in R. *Measurement: Interdisciplinary Research and Perspectives*, vol. 17, no 4, pp. 211–219. https://doi.org/10.1080/15366367.2019.1600839

Sireci S.G. (2021) NCME Presidential Address 2020: Valuing Educational Measurement. *Educational Measurement: Issues and Practice*, vol. 40, no 1, pp. 7–16. https://doi.org/10.1111/emip.12415

Sireci S., Benítez I. (2023) Evidence for Test Validation: A Guide for Practitioners. *Psicothema*, vol. 35, no 3, pp. 217–226. https://doi.org/10.7334/psicothema2022.477

Smagulova A., Satanov A., Kadirova F. (2025) Analysis of the Unified National Testing within the Study of the Child Well-Being Index in Kazakhstan. *Pedagogical Measurements*, vol. 1, no 1, pp. 47–63 (In Russian).

Smith E.V., Jr. (2001) Evidence for the Reliability of Measures and Validity of Measure Interpretation: A Rasch Measurement Perspective. *Journal of Applied Measurement*, vol. 2, no 3, pp. 281–311.

Smith E.V., Jr. (2002) Detecting and Evaluating the Impact of Multidimensionality Using Item Fit Statistics and Principal Component Analysis of Residuals. *Journal of Applied Measurement*, vol. 3, no 2, pp. 205–231.

Smith R.M. (2000) Fit Analysis in Latent Trait Measurement Models. *Journal of Applied Measurement*, vol. 1, no 2, pp. 199–218.

Stemler S.E. (2012) What Should University Admissions Tests Predict? *Educational Psychologist*, vol. 47, no 1, pp. 5–17. https://doi.org/10.1080/00461520.2011.611444

Westrick P.A., Le H., Robbins S.B., Radunzel J.M., Schmidt F.L. (2015) College Performance and Retention: A Meta-Analysis of the Predictive Validities of ACT® Scores, High School Grades, and SES. *Educational Assessment*, vol. 20, no 1, pp. 23–45. https://doi.org/10.1080/10627197.2015.997614

Woodruff D., Wu Y.F. (2012) *Statistical Considerations in Choosing a Test Reliability Coefficient. ACT Research Report no 2012-10*. Available at: https://www.act.org/content/dam/act/unsecured/documents/ACT_RR2012-10.pdf (accessed 07.11.2025).

Wright B.D., Masters G.N. (1982) *Rating Scale Analysis*. Chicago, IL: Mesa Press.

Wright B.D., Stone M.H. (1979) *Best Test Design*. Chicago, IL: Mesa Press.

Zwick R. (2019) Assessment in American Higher Education: The Role of Admissions Tests. *The Annals of the American Academy of Political and Social Science*, vol. 683, no 1, pp. 130–148. https://doi.org/10.1177/0002716219843469