МРНТИ 14.01.85 https://doi.org/10.63597/UTO3105-4161.2026.5.1.006

К.В. Тарасова
Центр психометрики и измерений в образовании, г. Москва, Российская Федерация
e-mail: ktarasova@hse.ru

ORCID: 0000-0002-3915-3165

ПСИХОМЕТРИКА И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: НОВАЯ ПАРАДИГМА ОБРАЗОВАТЕЛЬНОГО ОЦЕНИВАНИЯ

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

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

Введение

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

Расширение сфер применения ИИ, от промышленности и медицины до экономики, искусства и повседневных сервисов, неизбежно отражается и на образовании. В образовательной среде ИИ уже используется в интеллектуальных обучающих системах, адаптивном обучении, рекомендательных системах, прогнозировании образовательных и карьерных траекторий, автоматизированном оценивании и предоставлении персонализированной обратной связи. При этом оценивание и измерение становятся одной

из ключевых областей, где ИИ не только позволяет оптимизировать отдельные процедуры (например, автоматическую проверку заданий с открытым ответом), но и радикально пересматривает саму архитектуру инструментов тестирования, от проектирования конструкта до анализа результатов.

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

Материалы и методы исследования

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

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

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

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

вариантов текстов, после чего проводится многоступенчатый отбор на основе показателей сложности и экспертной оценки содержания [6].

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

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

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

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

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

Результаты и обсуждение

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

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

Исследования автоскоринга коротких ответов демонстрируют высокую вариативность точности — от 65% до 99% [20, 21]. Автоматическая оценка эссе остаётся более сложной задачей. Традиционные модели, использующие статистические, стилевые и контекстные признаки, достигают умеренной точности [22, 23]. Современные LLM лучше работают с «сырыми» текстами, но пока уступают специализированным моделям, особенно по критериям аргументации [24].

Таким образом, ИИ способен существенно поддерживать проверку открытых ответов, однако полностью автономный автоскоринг без участия экспертов остаётся недостижимым [25].

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

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

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

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

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

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

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

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

Заключение

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

Важным условием развития этой области становится международное сотрудничество. Центр психометрики НИУ ВШЭ ведет активное взаимодействие с Шанхайским институтом

искусственного интеллекта в образовании, Нанкинским университетом, Сианьским университетом Цзяотун–Ливерпуль (XJTLU) и Национальным центром тестирования Республики Казахстан. Эти партнерства позволяют сопоставлять модели в разных контекстах, проводить совместные исследования и укреплять научные основания для внедрения ИИ в оценивание.

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

Список литературы

1. Brown T. B. et al. Language models are few-shot learners // arXiv preprint. - 2020. - arXiv:2005.14165. - https://doi.org/10.48550/arXiv.2005.14165
2. Devlin J. BERT: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint. - 2018. - arXiv:1810.04805.
3. Radford A. et al. Language models are unsupervised multitask learners // OpenAI blog. - 2019. - Vol. 1, no. 8. - P. 9.
4. Darwish S. M., Mohamed S. K. Automated essay evaluation based on fusion of fuzzy ontology and latent semantic analysis // The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019). - Springer International Publishing, 2020. - P. 566-575. - https://doi.org/10.1007/978-3-030-14118-9_57
5. Attali Y. et al. The interactive reading task: Transformer-based automatic item generation // Frontiers in Artificial Intelligence. - 2022. - Vol. 5. - P. 934124. - https://doi.org/10.3389/frai.2022.903077
6. Bezirhan U., von Davier M. Automated reading passage generation with OpenAI's large language model // Computers and Education: Artificial Intelligence. - 2023. - Vol. 5. - P. 100161. - https://doi.org/10.1016/j.caeai.2023.100161
7. Bulut O. et al. The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges // arXiv preprint. - 2024. - arXiv:2406.18900. https://doi.org/10.48550/arXiv.2406.18900
8. Sharma K., Giannakos M. Multimodal data capabilities for learning: What can multimodal data tell us about learning? // British Journal of Educational Technology. - 2020. - Vol. 51, no. 5. - P. 1450-1484. - https://doi.org/10.1111/bjet.12993
9. Banville F. et al. Validation of a sorting task implemented in the virtual multitasking task-2 and effect of aging // Human Interface and the Management of Information. Information in Applications and Services: 20th International Conference, HIMI 2018. Proceedings, Part II. — Springer International Publishing, 2018. - P. 41-54. - https://doi.org/10.1007/978-3-319-92046-7_4
10. Al-Ansi A. M. et al. Analyzing augmented reality (AR) and virtual reality (VR) recent development in education // Social Sciences & Humanities Open. - 2023. - Vol. 8, no. 1. - P. 100532. - https://doi.org/10.1016/j.ssaho.2023.100532
11. Al Balushi J. S. G. et al. Incorporating Artificial Intelligence Powered Immersive Realities to Improve Learning using Virtual Reality (VR) and Augmented Reality (AR) Technology // 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC). - IEEE, 2024. - P. 760-765. - https://doi.org/10.1109/ICAAIC60222.2024.10575046
12. Von Davier M. Automated item generation with recurrent neural networks // Psychometrika. - 2018. - Vol. 83, no. 4. - P. 847-857. - https://doi.org/10.1007/s11336-018-9608-y
13. Tan B. et al. A Review of Automatic Item Generation Techniques Leveraging Large Language Models // OSF Preprints. - 2024. - https://doi.org/10.35542/osf.io/6d8tj
14. Omopekunola M. O., Kardanova E. Y. Automatic generation of physics items with Large Language Models (LLMs) // REID (Research and Evaluation in Education). - 2024. - Vol. 10, no. 2. - P. 168-185. - https://doi.org/10.21831/reid.v10i2.76864
15. Bozkurt A., Sharma R. C. Generative AI and prompt engineering: The art of whispering to let the genie out of the algorithmic world // Asian Journal of Distance Education. - 2023. - Vol. 18, no. 2. - P. i–vii. - https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/749

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

33. Panfilova A. S., Valueva E. A., Ilyin I. Y. The application of explainable artificial intelligence methods to models for automatic creativity assessment // Frontiers in Artificial Intelligence. - 2024. - Vol. 7. - P. 1310518. - https://doi.org/10.3389/frai.2024.1310518

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

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

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

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

К.В. Тарасова

ПСИХОМЕТРИКА ЖӘНЕ ЖАСАНДЫ ИНТЕЛЛЕКТ: БІЛІМ БЕРУДІ БАҒАЛАУДЫҢ ЖАҢА ПАРАДИГМАСЫ

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

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

K.V. Tarasova

PSYCHOMETRICS AND ARTIFICIAL INTELLIGENCE: A NEW PARADIGM OF EDUCATIONAL ASSESSMENT

This article examines how artificial intelligence is reshaping educational assessment and argues that these developments require a reconsideration of the psychometric foundations of measurement. The paper starts from a clear tension: while generative AI is rapidly being incorporated into item development, automated scoring, adaptive testing, and multimodal assessment environments, the measurement logic underlying these practices often remains insufficiently articulated. The aim of the article is to conceptualize this emerging agenda and to demonstrate that AI should not be treated merely as an external technical tool, but as a force that reconfigures assessment architecture from construct definition to score interpretation. The paper is conceptual and analytical in nature and is organized around three core domains: automated generation of assessment stimuli and test items with large language models; AI-based scoring of open-ended textual and graphical responses; and the assessment of AI literacy and other complex competencies in

digitally mediated environments. The analysis shows that AI substantially expands the possibilities of scalable content generation, multimodal task design, process-data capture, and operational efficiency. At the same time, these advantages do not in themselves ensure validity, fairness, interpretability, or theoretical coherence. The article therefore argues for a hybrid assessment model in which AI supports and extends assessment operations, while psychometric expertise, validation procedures, and human oversight remain indispensable. Particular attention is given to Evidence-Centered Design as a methodological framework for moving from declarative and self-report-based approaches toward behavior-based assessment of learners' actions in interaction with AI systems. In this way, the article contributes to the emerging methodology of AI-mediated assessment and offers a theoretically grounded perspective for integrating technological innovation with psychometric rigor in contemporary educational measurement.

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

References

1. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint*. https://doi.org/10.48550/arXiv.2005.14165

2. Devlin, J. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*. https://arxiv.org/abs/1810.04805

3. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

4. Darwish, S. M., & Mohamed, S. K. (2020). Automated essay evaluation based on fusion of fuzzy ontology and latent semantic analysis. In *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019)* (pp. 566–575). Springer International Publishing. https://doi.org/10.1007/978-3-030-14118-9_57

5. Attali, Y., Davier, M. V., Moore, C., Steier, M., & Burstein, J. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*, Article 934124. https://doi.org/10.3389/frai.2022.903077

6. Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence*, *5*, Article 100161. https://doi.org/10.1016/j.caeai.2023.100161

7. Bulut, O., Erhelfer, E. S., von Davier, M., & Kardanova, E. Y. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. *arXiv preprint*. https://doi.org/10.48550/arXiv.2406.18900

8. Sharma, K., & Giannakos, M. (2020). Multimodal data capabilities for learning: What can multimodal data tell us about learning? *British Journal of Educational Technology*, *51*(5), 1450–1484. https://doi.org/10.1111/bjet.12993

9. Banville, F., Nolin, P., & Labrosse, F. (2018). Validation of a sorting task implemented in the virtual multitasking task-2 and effect of aging. In *Human Interface and the Management of Information. Information in Applications and Services: 20th International Conference, HIMI 2018. Proceedings, Part II* (pp. 41–54). Springer International Publishing. https://doi.org/10.1007/978-3-319-92046-7_4

10. Al-Ansi, A. M., Jaboob, M., Akour, M., & Al-Ansi, A. (2023). Analyzing augmented reality (AR) and virtual reality (VR) recent development in education. *Social Sciences & Humanities Open*, *8*(1), Article 100532. https://doi.org/10.1016/j.ssaho.2023.100532

11. Al Balushi, J. S. G., Al-Ansi, A. M., Akour, M., & Al-Ansi, A. (2024). Incorporating artificial intelligence powered immersive realities to improve learning using virtual reality (VR) and augmented reality (AR) technology. In *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)* (pp. 760–765). IEEE. https://doi.org/10.1109/ICAAIC60222.2024.10575046

12. Von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika*, *83*(4), 847–857. https://doi.org/10.1007/s11336-018-9608-y

13. Tan, B., Liu, L., & Liu, G. (2024). A review of automatic item generation techniques leveraging large language models. *OSF Preprints*. https://doi.org/10.35542/osf.io/6d8tj

14. Omopekunola, M. O., & Kardanova, E. Y. (2024). Automatic generation of physics items with Large Language Models (LLMs). *REID (Research and Evaluation in Education)*, *10*(2), 168–185. https://doi.org/10.21831/reid.v10i2.76864

15. Bozkurt, A., & Sharma, R. C. (2023). Generative AI and prompt engineering: The art of whispering to let the genie out of the algorithmic world. *Asian Journal of Distance Education*, *18*(2), i–vii. https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/749

16. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35. https://doi.org/10.1145/3560815

17. Haley, D. T., Thomas, P., de Roeck, A., & Petre, M. (2007). Measuring improvement in latent semantic analysis-based marking systems: using a computer to mark questions about HTML. In *Proceedings of the 9th Australasian Conference on Computing Education (ACE 2007)* (Vol. 66, pp. 35–42). Australian Computer Society.

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

19. Higgins, D., Burstein, J., Marcu, D., & Gentile, C. (2004). Evaluating multiple aspects of coherence in student essays. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004* (pp. 185–192). https://aclanthology.org/N04-1024/

20. Ariely, M., Nazaretsky, T., & Alexandron, G. (2023). Machine learning and Hebrew NLP for automated assessment of open-ended questions in biology. *International Journal of Artificial Intelligence in Education*, *33*(1), 1–34. https://doi.org/10.1007/s40593-021-00283-x

21. Raz, T., Nazaretsky, T., & Alexandron, G. (2024). Automated scoring of open-ended question complexity: A large language model approach. *arXiv preprint*. https://doi.org/10.21203/rs.3.rs-3890828/v1

22. Burstein, J., LaFlair, G. T., Kunnan, A. J., & von Davier, M. (2021). A theoretical assessment ecosystem for a digital-first assessment — The Duolingo English Test. *Duolingo Research Report DRR-21-04*. https://duolingo-papers.s3.amazonaws.com/reports/DRR-21-04.pdf

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

24. Masikisiki, B., Marivate, V., & Hlophe, Y. (2023). Investigating the efficacy of large language models in reflective assessment methods through chain of thought prompting. In *Proceedings of the 4th African Human Computer Interaction Conference (Africhi '23)* (pp. 44–49). https://doi.org/10.48550/arXiv.2310.00272

25. Kostic, M., Stevanovic, A., & Jovanovic, J. (2024). LLMs in automated essay evaluation: A case study. *Proceedings of the AAAI Symposium Series*, *3*(1), 143–147. https://doi.org/10.1609/aaaiss.v3i1.31193

26. Schneider, J., Di Mitri, D., Limongelli, C., & Drachsler, H. (2023). Towards LLM-based autograding for short textual answers. *arXiv preprint*. https://doi.org/10.48550/arXiv.2309.11508

27. Von Davier, M., Tyack, L., & Khorramdel, L. (2023). Scoring graphical responses in TIMSS 2019 using artificial neural networks. *Educational and Psychological Measurement*, *83*(3), 556–585. https://doi.org/10.1177/00131644221098021

28. Tyack, L., Khorramdel, L., & von Davier, M. (2024). Using convolutional neural networks to automatically score eight TIMSS 2019 graphical response items. *Computers and Education: Artificial Intelligence*, *6*, Article 100249. https://doi.org/10.1016/j.caeai.2024.100249

29. Youn, Y. C., Kim, J., Ryu, J. H., Park, J. H., & Kim, H. J. (2021). Use of the clock drawing test and the Rey–Osterrieth complex figure test-copy with convolutional neural networks to predict cognitive impairment. *Alzheimer's Research & Therapy*, *13*, 1–7. https://doi.org/10.1186/s13195-021-00821-8

30. Ruengchaijatuporn, N., Phutrakool, P., Hemrungrojn, S., & Suchato, A. (2022). An explainable self-attention deep neural network for detecting mild cognitive impairment using multi-input digital drawing tasks. *Alzheimer's Research & Therapy*, *14(1)*, Article 111. https://doi.org/10.1186/s13195-022-01043-2

31. Raksasat, R., Phutrakool, P., Ruengchaijatuporn, N., Hemrungrojn, S., & Suchato, A. (2023). Attentive pairwise interaction network for AI-assisted clock drawing test assessment of early visuospatial deficits. *Scientific Reports*, *13*(1), Article 18113. https://doi.org/10.1038/s41598-023-44723-1

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

33. Panfilova, A. S., Valueva, E. A., & Ilyin, I. Y. (2024). The application of explainable artificial intelligence methods to models for automatic creativity assessment. *Frontiers in Artificial Intelligence*, *7*, Article 1310518. https://doi.org/10.3389/frai.2024.1310518

34. Uglanova, I. L., Gelver, E. S., & Tarasov, S. V. (2021). Otsenivanie kreativnosti na osnove analiza izobrazheniy s pomoshchyu neyronnykh setey [Creativity assessment based on image analysis using neural networks]. *Iskusstvennyy intellekt i prinyatie resheniy* [Artificial Intelligence and Decision Making], (1), 86–97. https://doi.org/10.14357/20718594210108

35. Lintner, T. (2024). A systematic review of AI literacy scales. *npj Science of Learning*, *9*(1), Article 50. https://doi.org/10.1038/s41539-024-00264-4

36. Ivanova, A. E., Tarasova, K. V., & Talov, D. P. (2025). Mezhdu interesom i umeniem: kak studenty vosprinimayut i primenyayut II [Between interest and skill: How students perceive and use AI]. *Vysshee obrazovanie v Rossii* [Higher Education in Russia], *34*(8-9), 9–32. https://doi.org/10.31992/0869-3617-2025-34-8-9-9-32

37. Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Авторлар туралы мәлімет:

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

Сведения об авторах:

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

Information about authors:

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