Original Article
IRSTI 14.35.09 10.51889/2960-1649.2025.65.4.001

BAIZHANOV NURSEIT[1], ABDRASILOV BOLATBEK[1], HAO JINGANG[2], MAKHMUTOVA ALFIRA[3]*

[1]National Testing Centre, Ministry of Science and Higher Education (Astana, Kazakhstan)
[2]Educational Testing Service (Princeton, New Jersey, USA)
[3]New Uzbekistan University (Tashkent, Uzbekistan)

*Address of correspondence: Nurseit Baizhanov, First Deputy Director, National Testing Centre, Ministry of Science and Higher Education, Rodnikovaya 1/1 St., Astana, Kazakhstan; https://orcid.org/0009-0008-5302-9858, E-mail address: nurseit.baizhanov@testcenter.kz, nurbaizhanov@gmail.com, Tel.: +77020003063

Harnessing the Potential of AI Technologies in the National Educational Assessments in Kazakhstan

Abstract

Introduction: Artificial intelligence (AI) is being rapidly adopted in education, including the evaluation of student performance and the assessment of teacher professional capabilities. The following presents a discussion of AI's potential applications in each stage of Kazakhstan's national educational assessments, from test item development to piloting, adaptive administration, automated scoring, and results interpretation. *Methodology and Methods:* A literature review of academic reports and practical applications reveals that AI has significant potential to enhance the efficiency, objectivity, and analytical depth of assessments. *Results*: These AI solutions enable the development of multi-modal tasks, the identification and addressing of anomalies in piloting, intelligent proctoring during administration, transparent scoring of open-ended responses, and deep learning analytics. At the same time, challenges related to algorithmic bias, data privacy, ethical responsibility, and algorithmic reliability must be addressed to deploy AI successfully. *Scientific Novelty:* This study presents one of the first systematic analyses regarding how AI can be integrated into all stages of Kazakhstan's national assessment cycle and offers a unified conceptual model linking multimodal task development with adaptive delivery, automated scoring, and deep-learning-based diagnostic analytics. *Practical Significance:* It also recognizes the need to upgrade infrastructure, create regulations, train workers, and introduce robust quality assurance systems. When used systematically and progressively, AI can help make assessments more transparent, equitable, and efficient. This would turn extensive evaluation into a flexible instrument for enhancing instruction and learning.

Keywords: Artificial intelligence in national assessment system, national assessment system, test item generation, automated scoring, adaptive testing, learning analytics, AI-based proctoring.

Introduction. Artificial intelligence (AI) technology has developed rapidly in recent years, bringing a substantial impact to the field of education. Research has shown that education has greatly benefited from AI, and AI has already been extensively applied to the learning, teaching, and assessment of students (Chen et al., 2020). The advent of advanced generative models has increased the opportunity for productivity and efficiency. However, it has created a number of problems relating to ethical challenges and different ways of thinking about our assessment of students (Chiu et al., 2023; Cotton et al., 2024; Hao et al., 2024). The system of national educational assessments (surveys), a crucial cornerstone in the quality assurance of schooling, has not been unaffected by these trends.

The main national educational assessments (survey) in Kazakhstan is diversified: the Unified National Test (UNT) as main entrance exam for university; the Teacher Knowledge Test aimed at the assessment of teachers' theoretical professional knowledge (competences); the introductory exam for obtaining master's and doctoral degree; the Monitoring of Educational

Achievements of Students (MODO), a monitoring assessment system for the quality of education in secondary schools. MODO is an autonomous system of surveillance that integrates the phases of preparation, administration, processing, and analysis of results, as well as methodological support, in schools (Csapó & Molnár, 2019). These are helpful tools for all types of education, from secondary to college, and they would all benefit from modern AI technology.

This study aims to explore the potential applications of AI at each level of a national testing system, with a focus on Kazakhstan. AI applications are utilized for test item development, piloting, test administration, and result scoring, as well as for subsequent analysis at various levels of aggregation (class, school, region, and country). This paper discusses recent advancements in AI, including multimodal models, generative models, explainable AI (XAI), federated learning, adaptive testing, and cognitive/emotional analytics. It also provides guidelines on how to integrate these innovations into the operation of national assessment systems.

Materials and Methods. This study is a literature review examining the potential applications of AI in national educational assessments, with a focus on Kazakhstan. We collected data through a systematic search of peer-reviewed sources, including top Scopus and Web of Science journals, as well as thematic databases such as IEEE Xplore and Google Scholar. We searched for phrases related to "artificial intelligence in assessment", "AI in education", "automated scoring", "adaptive testing", and "national examinations". We focused on research studies published from 2015 through 2025, concentrating on pertinent research from the past decade and capturing the most recent studies and debates in the field (Attali & Burstein, 2006; Chiu et al., 2023; Wang et al., 2024). To consider an international perspective, we examined local sources that could lend credence, including regulatory documents, official publications, and ministerial reports, on the Unified National Test (UNT) and the Monitoring of Educational Achievements of Students. We captured current developments through public announcements and news on AI pilots in Kazakhstan. Internal analysis by co-authors and expert opinions also contributed to the study, especially regarding AI's role in reporting education quality (U.S. Department of Education, 2023).

We processed the collected data using qualitative content analysis. We coded and organized the information thematically into five stages of the testing cycle: test construction, piloting, administration, scoring, and results analysis (Martínez-Comesaña et al., 2023). This procedure involved a systematic identification of AI applications within each phase of the assessment process. To ensure trustworthiness, we compared information from international evidence, gathered from the literature, with government documentation and expert review in a triangulation exercise (Perrotta & Selwyn, 2020; Zawacki-Richter et al., 2019). The act of triangulation strengthens the validation process, in turn adding strength to the findings and reducing bias.

The Results section outlines our findings, providing examples from each phase of the testing cycle, along with specific examples of AI applications. Examples include models that generate items, models that make predictions for piloting purposes, models that support proctoring in the administration phase (often referred to as "AI-based proctoring"), natural language processing systems that automate scoring, and advanced learning analytics that are used to analyze results (Cotton et al., 2024; Perrotta & Selwyn, 2020; Martínez-Comesaña et al., 2023). The discussion also addresses potential challenges, such as algorithmic bias, data privacy, and infrastructure limitations, giving a clear view of both opportunities and risks (Perrotta & Selwyn, 2020; Zawacki-Richter et al., 2019).

Results. *Test Item Development.* The development of tests is a time-consuming and intellectually demanding process. Generating thousands of questions across various areas (meeting the right competencies, matching curricula and difficulty levels, and facilitating translation into multiple languages) is a

significant amount of work. The efficiency with which this process can be carried out can be transformed by AI technologies. In Kazakhstan, some initial measures are already being taken in this direction. The Ministry of Science and Higher Education has announced a pilot project that will utilize AI to develop questions during the UNT process (Mukanov, 2025; Nikitin, 2025). A national AI platform is being developed, which will load all textbooks. Draft questions will then be generated by the generative model and be subject to expert review and piloting before being conducted (Mukanov, 2025; Nikitin, 2025)

At this stage of the assessment cycle, the incorporation of AI technologies offers several key benefits. Itembank expansion: Automated generation can produce a large number of distinct questions, preventing candidates from memorizing those from previous years and thereby making the exam fairer. Developers estimate that the number of test items might increase exponentially by cloning and rephrasing stem questions into different forms (Nikitin, 2025).

Decreased error: AI algorithms can produce results aligned with their intended purpose without accounting for arbitrary human factors (random error or writers' personal bias). This also ensures exam material is developed fairly. Reduced labor and time: manual inputs from content experts are removed through automated item creation, freeing up time for them to focus on creative and methodological review. This reduces the time and costs of test preparation (Nikitin, 2025).

There are additional efficiencies to be gained by using more AI and optimizing the quality of the questions themselves. Contemporary AI models can generate multimodal and context-enriched questions that assess a broader range of skills. For instance, an AI could automatically generate questions associated with an image (or graph or video clip) (U.S. Department of Education, 2023).

One test task had a model watch a brief video about an environmental issue (an ocean oil spill) and then perform a set of relevant tasks (describing what caused the spill, selecting the relevant consequences from a list, and suggesting an environmental fix). Appropriate visualization aids were presented for each task (U.S. Department of Education, 2023).

Scenario exercises with an emphasis on soft skills could be an interesting idea that might work. The AI can place the candidate in a situation and then ask the candidate how or why they chose a particular course of action. Then the system takes into account not only the answer, but also the applicant's reasoning and emotional reaction. For example, an AI-written scenario could read, "You are a volunteer at a camp, and one of the participants breaks a safety rule," while the student views a video of the scenario. The student must then select an action and justify their answer. Based on students' responses, the system can evaluate soft skills, including decision quality, justification, and emotional stability, among others (U.S. Department of Education, 2023). Authentic, complex tasks like this are challenging to develop manually and are valuable for a more accurate measure of proficiency.

AI also allows us to construct diagram and figure-based questions. Generative image models, combined with computer vision methods, can also automatically generate illustrations (function graphs or diagrams) and associated questions (U.S. Department of Education, 2023). For instance, a system could draw a speed–time graph based on the physics syllabus and then challenge students with questions that test the concepts and applications of this theoretical knowledge, such as identifying trends, interpreting acceleration, and predicting outcomes under various conditions (U.S. Department of Education, 2023).

Test Piloting. Test questions are generally piloted before they are added to high-stakes tests to ensure they work and produce reliable results. This pilot phase can significantly benefit from AI. Intelligent algorithms can predict how people taking the test will perform and analyze pilot data to identify outliers, culling them before they can negatively impact the results. Notable instances of AI-enabled improvements at this stage include identifying the most frequently answered questions that students get

wrong (or those with the most variable answers), detecting potential biases (such as items that are answered correctly only by a particular group of students), or predicting which items might not be functioning as expected. AI makes sure that only authorized content is live on the exam for this kind of analysis. That decreases the chances of having questions that are inaccurate or discriminatory. AI models can also forecast the difficulty and discrimination of an item through the role of "test taker" that they play in the item descriptions. AI in the pilot is ultimately included in the improved item pool.

Test Administration and Security. Additionally, AI tools can enhance the governance of exams. One of these areas is cheating detection, which many AI-driven proctoring systems aim to improve by leveraging real-time test-taker monitoring via face recognition, gaze tracking, and other computer vision technologies that help a human proctor identify cheating or suspicious behavior. Identity verification can be verified through biometric systems, and AI surveillance may be used to control online examinations to ensure that the rules are followed (preventing unauthorized supports, such as audio or paper) (Perrotta & Selwyn, 2020). Another application is adaptive testing algorithms, which allow the difficulty or order of questions to be adjusted online based on the test taker. This would make the tests more efficient while also providing each student with a personalized experience that maintains the appropriate level of challenge. AI can also be used in tests to make them less stressful and more informative by reducing anxiety and ambiguity. For example, chatbots can help applicants or answer questions they may have about the application process. In aggregate, these AI-driven mechanisms strengthen the reliability and equity of the administrative process and are more pleasant for users. Automated assistants help event organizers keep tabs and identify problems quickly, making things easier for human monitors.

Scoring and Evaluation. Once testing is complete and feedback is gathered, the appraisal process begins. This stage is crucial for open-ended tasks that cannot be scored using correct-answer keys. Most test questions in Kazakhstan's national tests (such as the UNT) have been multiple-choice because they could be graded by computer. There is, however, an increasing demand to incorporate more open-ended tasks, such as essays, short constructed responses, oral responses (in language tests), and applied tasks. The scoring process involves not only handing out marks but also verifying that they have been awarded correctly and responding to challenges and re-marks. AI offers several methods to expedite scoring, reduce costs, and increase objectivity at this stage.

Automatic scoring of constructed responses is a big opportunity. Natural language processing (NLP) has made tremendous progress to the point where it can grade essays, short answers, and even free-form problem solutions with considerable accuracy. In practice, nonhuman scorers of AI-generated essays (such as ETS's E-rater, used in the TOEFL) are already in use in high-stakes standardized testing (Attali & Burstein, 2006). Modern neural network models can be trained on student response datasets graded by human scorers, and then assigned scores to new responses that are highly correlated with human ratings. Adopting such a system in Kazakhstan might help address the issue of smudging written answers in exams (for instance, the essay section of a language exam) or in other national exams that contain open-ended questions. Research suggests that AI may do so faster and more reliably than humans, accelerating scoring turnaround time and allowing human graders to concentrate on cases that game finds contentious or too close to call (Martínez-Comesaña et al., 2023). Automated scoring can also be transparent by applying explainable AI approaches. This is crucial for convincing stakeholders to trust the AI system, as it can explain the reasoning behind its scoring decisions. Grading in the future could be done in a variety of ways, with the AI handling initial scoring and humans reviewing disputed answers or samples. In this way, algorithms can learn from human input and improve their performance - a good strategy if future tests are more diverse in task types.

Another element is the deployment of AI for quality control in scoring. AI may be able

to uncover errors or inconsistencies within the grading process, such as when one group of answers is scored differently from the rest (or, in the worst-case scenarios, when there are signs of human bias in the form of scores). It can also help flag unusual or potentially errant occurrences (such as when one batch of answers appears to have been graded inconsistently) and assist in identifying potential bias. AI can also support the automation of the appeals procedure; for instance, when a student contests a score assigned by AI, an XAI system can explain the features of the answer that affected the score and help human judges determine the final verdict. Overall, AI-based testing solutions maintain the integrity of scores by accurately and equitably grading exams (Perrotta & Selwyn, 2020; Hao et al., 2024; Csapó & Molnár, 2019; U.S. Ministry of Education of the Republic of Kazakhstan, 2023; Martínez-Comesaña et al., 2023; Liang et al., 2025).

Reporting and Feedback. A step that is frequently overlooked when test scores are given is analyzing the test results to improve educational services. AI methods in learning and educational data analytics can transform raw test data into actionable insights at multiple levels. Such AI-based analysis could provide each student with detailed feedback on their work, identifying individual strengths and weaknesses, as well as areas that require improvement. Teachers and schools may also adjust their regimens or teaching styles in response to aggregated analysis that indicates which subjects or skills students struggled with. At the system level (district, region, or country), AI can identify larger patterns and uncover unfairness or bias. For example, AI might identify that students in certain areas struggle with algebra, or that specific demographic groups perform differently, patterns that may not be apparent from overall scores alone. These analyses provide a roadmap for specific actions to improve the quality and equity of education.

Applying AI to large-scale testing data can also help identify "hidden factors" that influence achievement, which regular statistical methods may not capture. AI systems can forecast future results, categorize schools or students based on similar performance profiles, and show how a particular change (such as the introduction of a new curriculum or a teaching intervention) is likely to affect test scores. Policymakers and school officials can draw on the data from these studies to make wise decisions for the future. They are used to inform the allocation of resources and interventions based on evidence, identifying gaps in competency in specific content areas or areas that could benefit from support to reduce inequalities. Conceptually, AI can take national assessments from merely ranking students to diagnosing the overall "educational health" of the system. As recent research suggests, using AI-assisted analytics in conjunction with assessment data can provide multi-level feedback: personalized suggestions for students and teachers on the one hand, and database guidance for system-level improvements on the other. Ultimately, this helps to form the closed loop between learning and assessment by transforming test scores into opportunities to improve teaching and learning.

Discussion. The examination of AI applications reported shows considerable promise for updating the national teaching and learning cycle. Relying on AI can enhance every aspect of the process, from test development to score reporting. Testing fairness and objectivity can improve, save resources, and yield more information from all stakeholders. AI helps tests shift the emphasis from purely recording scores to learning more about how students learn (like their cognitive strategies, learning gaps, and other difficult-to-observe things) (Perrotta & Selwyn, 2020).

However, doing so has several potential challenges and restrictions. For one thing, there are concerns about the reliability and validity of AI solutions. Items must also be highly aligned with the curriculum and at the appropriate difficulty level for automatic item generation. If an algorithm goes wrong, a bad question winds up on an exam. Therefore, what is needed is a multi-tiered quality control system that combines machine efficiency with expert human judgment. Similarly, AI-generated scoring models must be validated to ensure they accurately measure the right things and

yield scores comparable to those of traditional grading. AI tools will have to be tested and calibrated frequently to ensure assessments are valid.

The second question concerns moral responsibility. If an AI system makes a mistake, for example, if it gives the wrong score, or if an algorithm accidentally spits out sensitive information, who is responsible? Guidelines or regulations with clear definitions and the assignment of responsibility are necessary, including the authors of the algorithms, the agency that implements them (such as the Ministry or the National Testing Centre), and, perhaps, the education community as a whole. Matters such as the right of appeal in cases involving AI decisions (a student challenging an essay grade awarded by an algorithm) will need to be explicitly addressed in policy (Perrotta & Selwyn, 2020).

Also, data privacy is a crucial aspect. National exams are based on data from hundreds of thousands of students, and proposals to use more multimodal sensors (such as video, audio, and biometrics from proctoring) raise legitimate privacy concerns. Strong measures should be taken to safeguard personal information, such as encrypting sensitive data, de-identifying data, and potentially employing naïve federated learning methods, which means that personal data should not be consolidated in one place (Perrotta & Selwyn, 2020). Adopting "privacy by design" practices will be crucial from the initial system development stage to ensure the secure storage of student data.

We should also be mindful of the risks associated with bias and discrimination in AI systems. AI models are trained on historical data that may be biased. For example, assume that certain groups - say, students from rural areas - have always performed worse because they went to worse schools. A basic AI model could assume that those are the patterns it was "meant for" and overlook the actual problems. We need special rules for vetting algorithms and curbing bias to ensure that AI does not exacerbate already unfair circumstances. Based on the literature, AI testing should include a diverse range of individuals, and AI should be easily explainable for auditing purposes to prevent unfair disadvantage to any students (Zawacki-Richter et al., 2019). In a school with students from diverse cultures and socioeconomic statuses, trust and fairness must rule. That is why AI systems need to be transparent about the reasons behind their decisions.

Another set of issues concerns people's skills and infrastructure. Introducing AI that works at a national scale in testing will require expertise in both technology (such as AI development and data science) and assessment (test development and psychometrics) to be embedded within the same professionals. Training will be necessary for staff and teachers to use the new systems effectively, ensure they can access AI-generated analytics reports, and understand how to utilize the information. The technical standardization of the infrastructure must be enhanced. All testing sites require fast, reliable internet service, sufficient computers, and secure servers capable of storing large amounts of data simultaneously. Strong cybersecurity is an essential component of the digital assessment environment's security. Without equal infrastructure development, AI solutions may not work as planned or could even worsen inequalities (by which schools with more advanced technology might get a leg up over those with outdated equipment).

Regulations and laws will also have to change in response to these new technologies. It may require revising the rules and regulations governing exams to permit and regulate the use of AI officially. Therefore, the Ministries of Education, Science, and Higher Education may need to establish guidelines outlining the limitations and capabilities of automated systems in testing, the validation process for AI-generated content, procedures for handling appeals against AI-graded results, and data protection standards. Such a sound legal basis will legitimize the procedure and provide clarity for all parties when AI is used in high-stakes testing.

Social acceptance of AI in testing is also a factor to consider. Prospective teacher candidates, parents, and teachers may also show skepticism about the reliability of "machine" grading and content production.

There will be an educational piece required to develop familiarity and understanding among stakeholders to be transparent about how the AI systems work, their usefulness, as well as the safety protocols." Demonstrating, for example, that an automated scoring system can perform as well as human graders, or introducing examples of AI-generated questions that human experts have approved, can help people view AI as a helpful tool rather than a menace.

Kazakhstan's initiatives are part of a broader global interest in utilizing AI in education. Many countries are innovating in their schools by using adaptive testing, automated grading, and big data analysis. Kazakhstan has already initiated several pilot projects, including an AI-powered question-generation system for the UNT and a chatbot to assist university applicants. It appears justified to incorporate AI-driven data analytics into the MODO system and to test automated essay scoring on a multilingual corpus of texts from Kazakhstani learners (Uzbek and Russian). To date, few national assessment systems have implemented AI at scale, positioning Kazakhstan as a potential early adopter that could develop a national assessment with AI at its core; if implemented effectively, Kazakhstan may become one of the early international adopters of such an integrated AI-based assessment model. This demonstrates how to integrate new technology with a human-centered evaluation process. This approach aimed to maximize the utilization of AI technology while keeping humans in the loop to review and explain the results (US. Department of Education, Office of Educational Technology, 2023).

It is also necessary to consider how poorly constructed an AI assessment might be, especially in light of the new challenges that AI (particularly generative AI) creates for essential and routine testing. As has been said before, students can cheat or evade real test problem-solving more readily now that they have tools like ChatGPT at their disposal. This can render conventional tests less helpful. To retain the fidelity of knowledge assessment, even in an era of disruption, assessment formats may also need to adapt. One direction increasingly highlighted in contemporary assessment research is that schools should teach more skills and tasks less prone to automation, such as creativity, collaboration, and vocational skills. If automated intelligence can solve "procedural" fundamental factual problems, it does not make sense to require kids to reproduce those answers on a test. In the future, national assessment systems may need to transition to include more performance-based tasks, project work, or authentic experimental problem-solving involving human creativity or physical interaction, domains where AI tools can assist teachers in judgment, rather than deceiving students. In summary, there are two key providers of AI: it assists us in better measuring, and it challenges us to rethink what we measure and how we do it, so that tests remain fit for purpose in the AI age. In the future, artificial intelligence (AI) could help evaluate this system more accurately and fairly. It could even help students master the higher-order skills education is supposed to teach.

Simultaneously, it is essential to understand AI integration not just as an improvement in technology, but as a shift in the strategic function of assessment. National assessments can shift from a ranking-based approach to directly supporting teaching and learning by providing customized feedback to students and guidance for educational policy. Together with expert and moderated human judgment, AI can enhance the assessment process to be more adaptive, inclusive, and competency-based in the twenty-first century.

Conclusion. This research suggests the potential to integrate artificial intelligence (AI) into Kazakhstan's national assessment system to create a more holistic, equitable, and efficient system. This study demonstrates how and where AI can be integrated within the testing process, from item generation and piloting stages to administration, reporting, and analysis. These integrations could facilitate the development of valid and reliable assessment instruments, enhance transparency within the assessment process, and provide more comprehensive feedback for candidates, teachers, and

policymakers. The primary contribution of this study is to contextualize and systematize the perspective on integrating AI technologies into national educational assessment processes. By synthesizing the expansive international literature on AI within a national context, this study contributes to the scientific development of understanding the role of AI in large-scale assessment. This also provides a conceptual foundation for developing a further empirical research agenda and underpins a pathway to modernize assessment systems that are better aligned with the formative competencies needed for the twenty-first century.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3). http://ejournals.bc.edu/index.php/jtla/article/view/1650

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, *8*, 75264-75278. https://ieeexplore.ieee.org/abstract/document/9069875/

Chiu, T. K. F., Lin, T. J., & Lonka, K. (2023). Repositioning higher education in an era of AI: Perspectives on ChatGPT. *Journal of Applied Learning & Teaching, 6*(1), 1–22. https://journal.alt.ac.uk/index.php/jalt/article/view/2932

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2), 228-239. https://www.tandfonline.com/doi/abs/10.1080/14703297.2023.2190148

Csapó, B., & Molnár, G. (2019). Online diagnostic assessment in support of personalized teaching and learning: The eDia system. *Frontiers in Psychology*, *10*, 1522. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.01522/full

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice*, *43*(2), 16-29. https://onlinelibrary.wiley.com/doi/abs/10.1111/emip.12602

Liang, J., Stephens, J. M., & Brown, G. T. (2025). A systematic review of the early impact of artificial intelligence on higher education curriculum, instruction, and assessment. In *Frontiers in Education, 10*, 1522841. https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2025.1522841/full

Martinez-Comesana, M., Rigueira-Díaz, X., Larranaga-Janeiro, A., Martínez-Torres, J., Ocarranza-Prado, I., & Kreibel, D. (2023). Impact of artificial intelligence on assessment methods in primary and secondary education: Systematic literature review. *Revista de Psicodidactica (English ed.)*, *28*(2), 93-103. https://www.sciencedirect.com/science/article/pii/S2530380523000072

Ministry of Education of the Republic of Kazakhstan. (2023). In Kazakhstan, the development of artificial intelligence for automatic checking of students' work has begun. https://www.gov.kz/memleket/entities/edu/press/news/details/537190?lang=ru

Mukanov, B. (2025). How artificial intelligence will be used when passing the UNT. *Zakon.kz*. https://www.zakon.kz/nauka/6468208-kak-iskusstvennyy-intellekt-budut-ispolzovat-pri-sdache-ent.html

Nikitin, S. (2025). Questions for the final UNT exam for schoolchildren will be generated by: The number of questions will increase exponentially. *DigitalBusiness.kz.* https://digitalbusiness.kz/2025-02-24/voprosi-vipusknogo-ent-dlya-shkolnikov-budet-generirovat-ii-kolichestvo-zadaniy-kratno-vozrastet/

Perrotta, C., & Selwyn, N. (2020). Deep learning goes to school: Toward a relational understanding of AI in education. *Learning, Media and Technology*, *45*(3), 251-269. https://www.tandfonline.com/doi/abs/10.1080/17439884.2020.1686017

U.S. Department of Education, Office of Educational Technology. (2023). Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations. https://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf

Wang, S., Wang, F., Zhu, Z., Zhao, X., & Hu, Y. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications, 252*(Part A), Article 124167. https://doi.org/10.1016/j.eswa.2024.124167

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators?. *International Journal of Educational Technology in Higher Education*, *16*(1), 1-27. https://link.springer.com/article/10.1186/S41239-019-0171-0

Information about authors:

Baizhanov Nurseit, Candidate of Legal Sciences, First Deputy Director, National Testing Center, Ministry of Science and Higher Education of the Republic of Kazakhstan. ORCID ID: 0009-0008-5302-9858, email: nurbaizhanov@gmail.com

Abdrasilov Bolatbek, Doctor of Biological Sciences, PhD in Physical and Mathematical Sciences, Corresponding Member of the National Academy of Sciences of the Republic of Kazakhstan, Chairman of the National Testing Center, Ministry of Science and Higher Education of the Republic of Kazakhstan. ORCID ID: 0009-0002-1371-6211, email: info@testcenter.kz

Jiangang Hao, PhD, Research Director, ETS Research Institute (USA), Associate Editor of Frontiers in Psychology: Quantitative Psychology and Measurement, and a regular reviewer for over ten high-impact journals, including Psychometrika, Journal of Educational Measurement, and Computers & Education. ORCID ID: https://orcid.org/0000-0003-0502-7571, email: jhao@ets.org

Makhmutova Alfira, Assistant Professor, New Uzbekistan University, Tashkent/Republic of Uzbekistan. ORCID ID: https://orcid.org/0000-0002-8597-7667, email: alfira2002@gmail.com