frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 18 September 2025
DOI 10.3389/feduc.2025.1654674

OPEN ACCESS

EDITED BY
Hope Onyinye Akaeze,
Michigan State University, United States

REVIEWED BY
Lucía Sánchez Bejerano,
Universidad Europea del Atlántico, Spain
Liamara Scortegagna,
Federal University of Juiz de Fora, Brazil

*CORRESPONDENCE
Yerbol Sarmurzin
✉ yerbol.sarmurzin@gmail.com

RECEIVED 26 June 2025
ACCEPTED 29 August 2025
PUBLISHED 18 September 2025

CITATION
Didarbekova N, Sarmurzin Y, Altybaeva S, Abdrasilov B and Karkenova A (2025) Advancing digital item development in Asian assessment systems: insights from Kazakhstan's Unified National Testing. *Front. Educ.* 10:1654674. doi: 10.3389/feduc.2025.1654674

Advancing digital item development in Asian assessment systems: insights from Kazakhstan's Unified National Testing

Nauzhan Didarbekova[1], Yerbol Sarmurzin[2]*, Shugyla Altybaeva[1], Bolatbek Abdrasilov[1] and Aina Karkenova[1]

[1]The National Testing Center of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Astana, Kazakhstan, [2]Buketov Karaganda National Research University, Karaganda, Kazakhstan

This study on the possibility of adopting a digital platform for item development in Kazakhstan's Unified National Testing (UNT) system employed a qualitative approach that incorporated both document analysis and semi-structured interviews with policymakers, item developers, and assessment experts. This study aimed to examine the current challenges, stakeholder needs, and global best practices. Thematic analysis identified issues such as fragmented workflows, limited metadata utilization, insufficient validation protocols, and security vulnerabilities in the existing system. These gaps hinder operational efficiency and raise concerns regarding test validity. The literature review and global best practices provided a hybrid framework of functional and organizational requirements for adopting digital platforms contextualized for Kazakhstan. The contributions include (1) an empirically grounded requirements specification that triangulates documentary evidence and stakeholder insights, (2) understanding the professional culture of assessment in the Central Asian context, and (3) policy recommendations for phased implementation, metadata governance, and localized platform design. The research findings and insights offer practical recommendations for policymakers and development partners involved in ecosystem assessment reform.

KEYWORDS

automated item generation, Unified National Testing, national exam, Kazakhstan, GradeMaker

Introduction

Improving national examinations is a high priority in the educational systems worldwide. While considerable work has focused on item delivery (computer-based testing, AI-based proctoring, etc.), item development is an upstream component that is less well-understood in many national assessment systems. The delivery of national examinations has greatly improved through computer-based testing and online processing systems. However, the development phase continues to depend on outdated manual processes (Abdrasilov et al., 2024; Mingisheva, 2023). Manual construction of questions requires a lot of time and effort (Kurdi et al., 2020). It is also a significant obstacle to implementing innovations that require item banks, such as adaptive testing and large-scale practice tools. The separation between modern delivery and traditional development methods has led to significant threats to testing validity, fairness, and operational performance.

International practice demonstrates how digital platforms for item authoring and management, such as GradeMaker Pro, TAO, and the Cambridge Assessment's proprietary systems, hold transformative potential. These platforms provide centralized workflows while integrating metadata and psychometric analytics and enable version control along with automated quality checks (Cambridge Assessment, 2023). Advanced tools enable Automated Item Generation (AIG), which is efficient and cost-effective and improves both assessment scalability and fairness (Gierl et al., 2022; von Davier, 2019).

According to Gierl and Lai (2016a), AIG constitutes a method in which computer algorithms produce numerous quality test items using cognitive models along with item templates. Identifying cognitive structures and designing manipulable item models constitute the first step in the process, which then proceeds to automatic item generation. The progress in AIG research encompasses improved cognitive model designs along with natural language processing integration and statistical pre-calibration methods to determine item difficulty without field testing (Embretson and Yang, 2006; Song et al., 2025). Advancements have enabled the quick generation of specialized items for various academic fields, including education and professional certification exams (Falcão et al., 2024; La Russa et al., 2025).

Globally, as international systems of assessment continue to digitize, renewed focus has emerged on the upstream processes that support item development. In this respect, artificial intelligence (AI), in the form of large language models (LLMs), has been identified as a potential force for change. Empirical findings show that LLMs such as ChatGPT are capable of generating valid multiple-choice items in pharmacotherapy (Kiyak et al., 2024) and general educational contexts (Hadzhikoleva et al., 2024). The Automatic Item Generation and Validation via Network-Integrated Evaluation (AI-GENIE) project also provides a human-AI collaborative workflow for producing test items (Russell-Lasalandra et al., 2024). While these application cases are currently primarily used in research and development settings, the increased accessibility and feasibility of such systems present a potential for future applications in large-scale testing systems to increase the efficiency and quality of their item authorship processes.

Since its inception, Kazakhstan's Unified National Testing (UNT), which functions as a pivotal exam for university admissions and state scholarships, has undergone significant reforms, including international standard integration, such as PISA, and the adoption of digital testing tools, together with AI-based proctoring methods and competency-based frameworks for reading and mathematical literacy (Jumabayeva, 2016; Mingisheva, 2023; testcenter.kz, n.d.). Following Kazakhstan's low performance in the PISA 2009 and 2012 cycles, in which students showed low average reading and mathematical literacy (Sarmurzin et al., 2021, 2025), the country embarked on a number of education reforms to align its national assessment practices with international standards (Sarmurzin et al., 2021). A key shift in policy was the decision to incorporate functional literacy components into UNT. Since 2017, mathematical literacy tasks have substituted the mathematics section, and reading literacy items have replaced traditional grammar-based questions in the Kazakh or Russian language tests, depending on the language of instruction (Abdrasilov et al., 2024). This was a fundamental shift from testing declarative knowledge to assessing higher-order thinking and real-world application skills. The authoring of UNT test items remains dependent on manual text editors and in-person workflows despite advancements in delivery technologies.

The National Testing Center (NTC), which operates under the Ministry of Science and Higher Education of Kazakhstan, serves as the central coordinator of the UNT. The NTC strategically focuses on test item quality by applying psychometric guidelines that emphasize validity, reliability, and fairness (Abdrasilov et al., 2024). Existing procedures demand excessive human and time resources and do not have systematic version control, which prevents iterative quality assurance (Mingisheva, 2023). Currently, the test is carried out manually using text editors, which creates time- and resource-consuming conditions and complicates the process of controlling the quality of the material and version (Abdrasilov et al., 2024). The lack of alignment with international standards affects Kazakhstan's capacity to maintain content integrity, because the UNT continues to grow in complexity and size (primeminister.kz, 2023). Classical methods based on the involvement of people at each stage lead to scattered work, little information security, and a lack of standardization (Frederick and Grammar, 2022). In the age of digitalization of education, an integral digital solution that ensures all stages of item creation, from authoring and reviewing to quality assurance and storage, is required for transparency, consistency, and efficiency.

Research evidence regarding the effective adaptation of digital authoring platforms for assessment systems in Central Asian regions remains minimal. Most existing research has examined the execution of computer-based testing and digital delivery systems instead of exploring the fundamental upstream stage of item creation. Kazakhstan's reform experience provides a crucial example for understanding how national examinations in transitional societies can benefit from digital platforms integrated into their workflows.

The goal of this research is to identify the state and future prospects for the development of digital UNT items. To achieve this goal, this study addresses the following research questions:

- What international best practices exist for utilizing digital platforms to develop materials for high-stake national tests?
- How is Kazakhstan's UNT system currently developing and managing test materials and what specific obstacles are encountered in this process?
- What are the functional and organizational requirements necessary for the successful implementation of a digital platform in the UNT system?

Through mixed-methods research, this study seeks to produce practical insights to shape national policy and enhance worldwide discussions on modernizing assessment methods.

Context

The Unified National Testing (UNT) in Kazakhstan

Since the period of independence, the country's educational system has undergone numerous reforms to adapt to the

requirements of integration into world education and increase the quality of education (Sarmurzin et al., 2021). One of the most significant changes was the introduction and transformation of the UNT, which has served as both a tool for admission to higher education and an indicator of secondary education quality since 2004 (Abdrasilov et al., 2024). Held annually by the National Testing Center, the UNT was designed to improve the objectivity, reliability, and transparency of student assessments and eliminate the risks of corruption and differences in the admission of students to higher educational institutions throughout the country (egov.kz, 2024; Jumabayeva, 2016).

Previously, the UNT replaced the final state certification of school graduates with entrance examinations to enter the university. Its combination was aimed at simplifying the process of moving from school to university and at the same time served as a unified measure of the level of student achievement at the end of school education (OECD, 2017).

Over the years, the UNT format, organization, and philosophy have changed significantly (Table 1). It was previously criticized as being too rigid and bureaucratic, contributing to increased stress in test-takers (Bakas uulu and Smagulov, 2016; Kirichok et al., 2025). Scholarly discussions also addressed the issue of the potential narrowing of the school curriculum and the appearance of a "hidden curriculum" due to excessive test preparation and the gap in the quality of education, especially between city and village schools (Mingisheva, 2023). In the early years of UNT, corruption and concerns about cheating were common themes (Jumabayeva, 2016).

As a result of such persistent criticism, and to improve the quality of the UNT, it was decided to move away from the UNT to international assessment models and, in particular, to the format of the Program for International Student Assessment (Sarmurzin et al., 2021, 2025). In recent years, reading and mathematical literacy sections have been added to the UNT to evaluate the competencies of reading and using the subject's information and mathematical literacy in real-life tasks (egov.kz, 2024). This change in the UNT structure proves that the government plans to support functional literacy and engage students' critical thinking skills instead of memorization (Abdrasilov et al., 2024).

Currently, the UNT consists of three mandatory subjects ("History of Kazakhstan," "Reading Literacy," and "Mathematical Literacy") and two elective subjects, one of which is chosen by the candidates depending on their chosen major (testcenter.kz, 2025). Twenty questions were for "History of Kazakhstan," 10 for "Reading Literacy," 10 for "Mathematical Literacy," and 40 for each of the two subject matters. The highest possible score is 140 points. The distribution was made in a way that gives more significance to the two elective sections since they are more related to the student's chosen study field, whereas the three compulsory sections test the student's general knowledge and core academic skills. The format of the UNT is made up of multiple-choice items consisting of single-answer questions, context-single answer questions, multiple-answer questions, and matching tasks. The test time was 4 h (240 min), with 40 min allotted for students with special educational needs (testcenter.kz, 2025). In addition to extended time, the testing system can also be set up to give students a separate room to take the test, an alternate test without charts and diagrams for visually impaired students, and an embedded function of screen magnification. Trained personnel may also be available to help students when necessary.

TABLE 1 Timeline of UNT reforms in Kazakhstan.

Year	Reform	Description
2004	Introduction of UNT	UNT launched as a unified state exam combining school graduation and university entrance. Subjects: language of instruction, mathematics, history of Kazakhstan, and one elective. Max score: 120
2008	Expansion of subjects	For Kazakh-medium schools, Russian language was added; for Russian-medium schools, Kazakh language was added. Max score increased to 125. Items per subject reduced from 30 to 25
2015	Introduction of visuals and logic items	Questions with diagrams, tables, and logical reasoning added to STEM subjects
2017	UNT format overhaul	School graduation separated from university entrance. UNT focused on university admission only. Added blocks: Reading Literacy, Math Literacy, History (20 items each) + 2 elective subjects (40 items each)
2018	English-language UNT option	Students could take the test in English for the first time
2019	Increased flexibility	UNT allowed four times a year for paid admission; grant applicants tested in June only
2020	International certificate exemption	Students with IELTS or TOEFL exempt from the foreign language block
2021	Full digitalization pilot	UNT conducted electronically via "1 computer −1 camera −1 test taker" model; two free attempts allowed. AI proctoring and personalized scheduling introduced
2023	Legalization of two attempts	Two attempts officially approved by law; best result used for grant application. Additional time granted to students with special needs. Revised item counts in History and Reading Literacy
2024	Item structure and scoring update	Fewer items in Reading and Math Literacy (−5 each); increased items in electives (+5, including matching questions)

Adapted from Abdrasilov et al. (2024).

Additionally, the format of the UNT was transformed to make it more flexible and not burdensome for test-takers. The introduction of flexibility to the UNT is ensured by one of the changes–the replacement of the annual unique administration of the test by the system according to which each student has the right to take the test four times a year (primeminister.kz, 2023). All four certificates of UNT are valid for application to tuition-based programs if the minimum score required is reached. For application to state-funded scholarships (grants), the certificate with the highest result achieved by the student during spring or summer testing is submitted. When the UNT administration was held once a year, the examination bore too much psychological burden on the student applicants, and the project was widely criticized by public opinion. The new format relieved the examination of this unnecessary

load and made it more student-oriented, as it gave students an opportunity to choose the date and location of their test session.

Moreover, digitization has become a prominent feature of UNT. The exam is now entirely computer-based, equipped with a range of innovative tools, such as artificial intelligence monitoring (e.g., "1 computer −2 cameras −1 test subject"), on-screen calculators, and instant topic-based analysis of results for rapid reaction to students' performance (primeminister.kz, 2023). It also eliminates the likelihood of manipulating results by cheating test-takers (Abdrasilov et al., 2024). To further increase test security and decrease the chances of manipulation, the NTC implemented additional security measures, including mandatory face-ID checks to detect impersonation attempts and full video surveillance of examinees during the test. The recorded video sessions were subject to review after the test was completed, and the test scores could be retroactively annulled in cases where violations were detected. For example, by 2024, several candidates had their UNT scores annulled after their videos were reviewed post-exams (Tengrinews.kz., 2024). These actions allow for the examination to be conducted more objectively, transparently, and efficiently.

Test-taking is necessary for applicants who wish to enter a higher educational institution, as the results of the UNT are the main criteria for distributing state educational grants. The test-taker must achieve a minimum score of 50 points for all subjects to be eligible for university admission. For highly competitive specialties, such as education, medicine, and law, the minimum score was higher (75 points) (egov.kz, 2024).

However, despite the above-mentioned reforms and technical improvements, the existing issues related to UNT and its influence on the educational system of the country continue to be the subject of discussion in the scholarly literature. The growing stress experienced by students before and during the test, even after multiple attempts, remains a concern for those who monitor the country's education (Kirichok et al., 2025).

Additionally, the effectiveness of UNT in preparing applicants for the international education market and real-world workforce remains questionable. In recent years, an increasing number of Kazakhstani school graduates have been studying abroad, especially in countries where admission is carried out through alternative assessment systems or international English language proficiency tests, such as IELTS (Anuarbekova and Sultan, 2025). This shows that while the UNT aims to standardize the assessment system within the country, some students and their parents believe that there are better chances for a more comprehensive assessment in foreign educational systems. Additionally, Kazakhstani schoolchildren's growing interest in international exams, especially the English part of the UNT, confirms that there is a changing assessment and recognition space in Kazakhstan (Anuarbekova and Sultan, 2025).

English is an elective subject in the UNT system. The majority of the test takers are applicants who want to get into English- related majors, such as English language teaching, English philology, or international relations. Students with internationally recognized language certificates, such as IELTS or TOEFL, can convert their results to UNT scores (Abdrasilov et al., 2024). Since 2018, as part of Kazakhstan's trilingual education policy (Amanzhol et al., 2024; Sarmurzin et al., 2024), test-takers have been allowed to sit the entire UNT in English to accommodate students who have graduated from schools with English as the medium of instruction.

The process of creating materials for UNT is regulated by the NTC. It includes determining test goals, writing test items, reviewing subject matter experts, assembling test items, and distributing them. However, the creation of test items is currently carried out manually and on standard text editing tools divided by topic (Abdrasilov et al., 2024). The system is unable to deal with the increasing demand for high-quality, secure, and diverse materials, particularly because the number of UNT test-takers continues to grow every year (primeminister.kz, 2023). The approach also does not fully comply with modern requirements for security, as evidenced by strict control over rule violations, where test-takers' results are canceled if the rules are violated. Sixty-eight applicants did not receive the right to take the test because they brought forbidden objects, and 74 were expelled from the testing rooms to violate the rules during the test (Yermaganbetova, 2025). Although this percentage is very small (only 0.1%), it still clearly indicates the continuing need to further develop preventive measures, rather than focusing exclusively on repressive steps.

Thus, as an educational tool, UNT has long affected the lives of the younger generation and continues to do so. However, the aim of the research presented here is not to present the positive or negative aspects of the process, but to highlight the points of improvement in the UNT and consider the possibilities of the study subject from a scientific point of view.

Literature review

Automated Item Generation (AIG)

In recent years, the emergence of Automated Item Generation (AIG) has offered a new approach for large-scale assessment design by leveraging cognitive models and computational algorithms to rapidly produce test items at scale (Gierl et al., 2022; Gierl and Lai, 2016a). While traditional test development relies on manual authoring guided by psychometric principles such as validity, reliability, fairness, and standardization (APA, 2014; Kane, 2013), AIG provides a systematic means to operationalize these principles in digital assessment environments.

Although the conceptual roots of AIG can be traced back to Bormuth's (1970) early ideas, its practical development gained momentum only decades later. The initial implementations were based on static instructional objectives (Roid and Haladyna, 1978), but this was not until the work of Drasgow et al. (2006) that a coherent theoretical and computational framework was formalized. Since then, AIG has evolved into a robust methodology for generating test items automatically and at scale, enabling institutions to reduce development costs, accelerate production cycles, and maintain item security and psychometric consistency (Drasgow et al., 2006; Pugh et al., 2016).

In education, AIG aims to produce multiple unique items targeting the same construct, enhancing test security and content breadth (Pugh et al., 2016). Despite these advantages, the practical application of AIG remains limited in many national testing systems, necessitating critical evaluation of its feasibility, item

quality, and scalability. The AIG enhances reliability by ensuring a consistent item structure and controlled variation, thereby reducing human error in item authoring. It supports validity by using domain-specific models to ensure item alignment with targeted constructs (Gierl et al., 2021). Fairness is addressed through standardized templates that minimize unintended bias, whereas standardization is improved through automated item formatting and quality checks.

AIG systems typically include three components: (1) item models that define item structure and cognitive processes; (2) algorithms to generate item variants; and (3) validation procedures to screen for psychometric quality (Gierl and Lai, 2016a). These innovations allow scalable production of test items, especially for high-stakes testing programs where item security and psychometric consistency are crucial.

Facilitating the production of consistently high-quality items at scale has been one of the primary long-standing challenges in the AIG literature (Chan et al., 2025). Recent AIG research has established that it is now possible for automatically generated items to achieve a quality that is on par with that of manually developed items (Falcão et al., 2023). Concurrently, substantial efforts have been directed toward improving the validity, reliability, item discrimination, and difficulty of AIG-based assessments by adopting more methodologically sound approaches (Tan et al., 2024).

It is believed that there is a growing sophistication of AIG systems and their practical feasibility in education. For instance, Circi et al. (2023) discussed how template-based AIG, when designed through domain modeling and validated by experts, can significantly enhance item diversity, while ensuring construct validity and content representativeness. Their findings also highlighted the role of subject matter experts in calibrating item difficulty, an area that is often considered a limitation of automated approaches.

Despite their potential, Circi et al. (2023) underscores persistent challenges such as generating higher-order cognitive items and ensuring pedagogical alignment across domains. They advocated hybrid models that combined human validation with automated generation cycles, striking a balance between scale and quality.

Moreover, platforms such as GradeMaker, TAO, and AIGen provide not only high item output but also diagnostic feedback, metadata tagging, and cognitive complexity calibration, streamlining both formative and summative assessments. However, AIG requires interdisciplinary collaboration combining domain expertise, psychometric validation, and software engineering. Key challenges include maintaining content diversity, detecting subtle errors in the generated output, and gaining acceptance from teachers and stakeholders (Gierl et al., 2022).

Manual vs. automated item authoring

Automatic testing systems have been developed to overcome several issues that affect the manual creation of tests. They make the test creation process less monotonous, more time-efficient, less subject to duplicates, and significantly more secure (Klammer and Ramler, 2017). These features of automatic testing help eliminate unnecessary repetition of items from previous exams and manage various constraints such as test difficulty, score range, question type, and alignment with course learning outcomes (Gangar et al., 2017). The automatic testing system is connected to an extensive item bank organized by scores, difficulty, question type (knowledge, memory, logic, and application), and alignment with course learning outcomes.

The security of an automatic testing system can be enhanced by minimizing the risk of paper leaking prior to an exam. Features such as real-time generation (which can generate test minutes before an exam), two-factor authentication, limited access, and audit logs have significantly improved security (Pauli and Ferrell, 2020). In addition, they provide increased efficiency by eliminating the need for manual effort, save time, and are capable of automating formatting and PDF output (Gangar et al., 2017). These systems also have the capability to create assignments for students with special needs, such as Braille, and translate the test into another language, along with regular tests (Frederick and Grammar, 2022).

GradeMaker Pro is a commercial system with a variety of features, including streamlined test authoring, test item banks, security, quality control, and multi-format test publishing (Frederick and Grammar, 2022; Pauli and Ferrell, 2020). The test item bank provides secure storage and management of the test items and materials that can be reused in the future. By using this feature, it is possible to create multiple test forms and practice papers (Pauli and Ferrell, 2020).

Quality control is a term used in digital settings to refer to questions and their accuracy, validity and reliability. This term refers to Bloom's taxonomy, which measures the cognitive scope of questions (Gangar et al., 2017; Stevens et al., 2007). Another method to control quality is to use Bloom's taxonomy and ensure that the questions align with course learning outcomes. In addition, more advanced techniques such as Item Response Theory (IRT) can be applied to determine item difficulty and item discrimination to be more accurate and enable comparison between the knowledge domains (Avanesov, 2005; Schmucker and Moore, 2025).

Automation plays a key role in the enhancement of assessment quality and integrity, as the transition from manual to automated test creation mitigates several issues affecting the quality (duplication, bias) and integrity (leaks) of tests (Frederick and Grammar, 2022; Gangar et al., 2017). The automatic test creation system systematizes item selection, randomizes questions, and implements strong security features to make the entire process fair and reliable, especially in the case of high-stake exams.

Another key feature of item centralization in digital systems is the possibility of aggregating performance data on a large scale. These data can be analyzed using advanced psychometric models, such as the IRT, and the parameters of each item in the item bank (difficulty and discrimination) can be adjusted. Thus, the item bank maintains a high-quality collection of items that contribute to a valid and reliable test. This also leads to a feedback loop that continues to improve assessment tools. Table 2 summarizes the comparison between manual and automated test creation following Gangar et al. (2017).

TABLE 2 Manual vs. automated test creation.

Characteristic/ aspect	Manual test generation	Automated test generation
Human effort	Requires significant human effort; monotonous and labor-intensive	Automated process, reduces human effort
Patterns/repetition	Patterns or question repetition may occur	Fully random and impartial process, prevents duplication
Security	Low security; high risk of test leakage	High security; minimal risk of test leakage (generated minutes before the exam)
Speed	Slow due to manual labor involvement	Faster due to computer automation
Question variety	Limited variety of question types	Wide variety of question types
Storage	Limited storage capacity; vulnerability to damage; document transport issues	Questions stored in an extensive, classified database; easily modified and expanded
Modification	Difficult to modify paper documents	Convenient question modification
Systematic procedure	Lack of systematic quality assurance procedure	Implements a modern, evolving process that manages multiple constraints and ensures quality
Bias	Potential for human bias	Fully impartial process

Adapted from Gangar et al. (2017).

Methodology

Research design

This research followed a qualitative approach, utilizing an embedded case study design (Yin, 2014). As test item development is both a complex and layered institutional and technological process, it is important to explore it in detail, which case study methodology is well-suited to do. Case studies are one of the strategies for inquiry that works best when the investigator has a contemporary set of phenomena to study in its real-life context when the boundaries between phenomenon and context are not well-defined (Creswell and Poth, 2016; Yin, 2014). The case here is the process of test development with its institutional logic, which is part of the national assessment reform conducted at the NTC in a digital environment.

Yin (2014) suggested that an embedded case study design be composed of one unit of analysis and several sub-units of analysis. The unit of analysis in the current study is the institutional process of test development, and the sub-units of analysis are the practices of using the platform, item generation workflows, and quality assurance in the NTC. Case studies have been recognized as an effective approach for collecting rich data across different stakeholders in educational research (Yazan, 2015). Data for this study were collected using two techniques (document review and semi-structured interviews) that triangulated and complemented each one.

Data collection

Data were collected from February to May 2025. It consists of two data streams: document analysis and semi-structured interviews. A document analysis included a review of peer-reviewed literature, policy, and other public documents, reports, technical manuals, and other descriptive resources related to digital platforms (e.g., GradeMaker Pro, Cambridge Assessment). Document analysis focused on information regarding platform features and functionalities related to item banking and workflow configuration as well as AIG tools (in the context of automated item generation) embedded in such systems.

We selected semi-structured interviews as a method for their structure that is rigorous and consistent across participants, while simultaneously offering flexibility and opportunities for in-depth and open discussion of emerging ideas and issues (Ruslin et al., 2022). Interviews are a common choice in exploratory case study research, particularly in the case of a study seeking detailed input from a relatively small set of participants, who may differ in their roles and levels of expertise (Naz et al., 2022). An interview protocol was created based on the literature on digital assessment and themes that emerged from the document analysis. The protocol included open-ended questions and was organized around key themes such as platform features, item writing workflow, AIG tools, and policy-related challenges. A pilot interview with one of the NTC experts was conducted to ensure the clarity and flow of the questions.

A purposeful stratified sampling approach was employed to achieve diversity in expertise and institutional roles. Patton (2002; p. 240) stated that a stratified purposeful sample could "capture major variations rather than identify a common core, although the latter may also emerge in the analysis. Each of the strata constitutes a fairly homogeneous sample." Inclusion criteria required that participants be directly involved in the UNT process in Kazakhstan. A total of 12 participants were recruited, including six test developers from the National Testing Center (NTC), representing various subject domains; four NTC experts in quality assurance and platform design; two Ministry of Education officials overseeing assessment modernization; and external IT experts with experience in large-scale educational technology implementation. Each interview lasted approximately 45–60 min and was conducted using Zoom video conferencing. With the participants' consent, all sessions were audio-recorded and transcribed verbatim. The interview protocol included scenario-based prompts regarding platform use, functional requirements, and challenges. Manual verification and anonymization of the transcripts were performed. Thematic coding was used iteratively (with the help of NVivo) to identify patterns in the data that reflected participants' experiences and views.

For RQ3, which aimed at determining the functional and organizational requirements for adopting a digital platform in the UNT system, we integrated findings from the interview data (RQ1) and the literature review (RQ2). The interview transcripts were

TABLE 3 Sample of interview data.

Theme	Sub-theme	Quotation
Fragmented and inefficient item development workflows	Lack of centralized item tracking	"Sometimes we write duplicates by accident because we can't access the bank or a previous version of the test."—Expert 7
	Isolated work practices	"Actually, we can create item together, but it is time consuming. Every expert attempt to finish as quickly as possible. So usually everyone works in their own bubble."—Expert 9

coded by themes related to interviewees' perceptions of barriers and expectations regarding digital assessment. These themes were triangulated with the best practices and recommendations identified in the literature on international digital assessment systems. This approach enabled us to abstract a set of context-specific requirements for Kazakhstan's UNT while simultaneously anchoring each requirement in theory and evidence.

Data analysis

Data analysis was conducted concurrently and iteratively with data collection, with findings from one stream informing the other.

Document analysis

Qualitative content analysis was applied to identify key themes related to digital platform functionality, quality assurance mechanisms, implementation factors, and challenges. These themes directly informed the interview protocol and questions to explore international practices from a locally relevant perspective.

Interview data

A six-phase thematic analysis (Braun and Clarke, 2006) was employed: familiarization, initial coding, theme identification, review, definition/naming, and final write-up. To enhance analytical rigor, transcripts were independently coded by two researchers and intercoder reliability was checked against each other's coding to resolve discrepancies. Thirty-six initial codes were generated from the dataset, representing the range and richness of the experts' responses. The initial codes were reviewed and subsequently collapsed into three main themes and 11 sub-themes to maintain analytical coherence and prevent fragmentation. To increase transparency and ensure the trustworthiness of the analysis, Table 3 provides an example of one main theme and its subthemes with illustrative quotes from the interviews. This presentation shows how the extracted interview passages contributed to the building of the themes, and how the key ideas were inductively developed.

A shared coding framework was employed across both data types to facilitate the triangulation. Convergences and divergences between international practices and local perspectives were identified to analyze adaptation opportunities and potential implementation barriers. When themes from the document analysis converged with interview findings (e.g., fragmented workflows, metadata gaps, and the need for version control), they were considered triangulated findings. Unique themes from each stream were identified and flagged for analysis. This approach ensures that both global benchmarks and local constraints inform the final requirement specification and contextual discussion.

Ethics

All procedures involving human participants in this study were conducted in accordance with the Ethical Code of Researchers in Education (Kazakhstan Educational Research Association, 2020). Before data collection, all interviewees received a detailed informed consent form that outlined the study's purpose, procedures, potential risks, and their rights as participants, including the right to withdraw at any point without any consequences.

Written consent was obtained before each interview session. All participants were assured of strict confidentiality and personal identifiers were removed from the transcriptions to maintain privacy. Audio recordings and transcripts were stored on encrypted password-protected servers, which could only be accessed by the research team.

No financial or material incentives were provided for participation. The study did not include minors or vulnerable populations, no deception or covert data collection methods were employed, and all findings were reported with a commitment to accuracy, transparency, and the protection of stakeholder interests.

Findings

Thematic analysis of data, integrating documentary analysis of international best practices, semi-structured interviews, and internal documents from the NTC, identified four key themes regarding the modernization of item development processes for Kazakhstan's UNT. They are directly aligned with our three research questions: Theme 1, provides an answer to RQ1 in terms of distilling the body of international best practice on digital assessment; Themes 2 and 3 are our response to RQ2 in terms of the immediate problems and priorities that manifest in the system of UNT; and Theme 4 is our response to RQ3 in terms of the functional and organizational requirements of an appropriate platform choice.

Theme 1 (RQ1): international case studies on digital assessment platforms

In 2017, Zimbabwe School Examinations Council (ZIMSEC) switched to GradeMaker to modernize its formerly manual, laborious, and inefficient process of developing questions for national exams. Challenges in the previous system included fragmented workflow, lack of version control, insecure sharing, lack of ability to use high-quality questions in other tests, and high costs. These issues were addressed by providing a centralized

system with authoring tools, secure item bank, item tracking, and controlled access. This provides improved transparency, flexibility, and security. By 2021, the platform is expected to support 37 ordinary subjects. This resulted in increased efficiency, quality, and cost-effectiveness, but success also depended on the ability to build digital literacy and to adopt the new process into a new digital workflow (Frederick and Grammar, 2022).

Moreover, the move to GradeMaker enabled ZIMSEC to move from full paper authoring to testing the assembly from a bank of vetted questions. This resulted in reduced predictability, plagiarism, or leakage, enhanced standardization, and improved the psychometric quality of the test forms. The system also includes built-in auditing and more advanced user rights management, which improves information security. Additionally, GradeMaker facilitated the development of adapted versions of examinations for learners with special educational needs (Frederick and Grammar, 2022).

In the United Kingdom, the National Examination and Qualifications Alliance (AQA), which administers over half of all GCSE and A-level exams, has successfully implemented GradeMaker Pro for large-scale exam development. By 2023, over two-thirds of AQA's GCSE and A-level exam papers had been authored using GradeMaker Pro. With AQA's complete acquisition of GradeMaker, the platform will be further developed to support on-screen assessment and broader usage across a range of qualification types. This highlights how national-level assessment organizations use advanced authoring platforms to modernize item banking and improve exam paper security (Leigh, 2023).

Guyana's education improvement program, led by the Ministry of Education and the World Bank, involves the implementation of GradeMaker Pro for national exam authoring and item banking. Starting in 2018, the program was established to transform the examination process using new software for item authoring, secure item banks, and capacity building for test writers. This has also led to new scanning processes for multiple-choice questions and performance reporting by topic, thus creating Guyana's first national longitudinal assessment database. These changes were designed to increase test security, support the country's new competency-based curriculum, and allow for better school-level feedback and planning. By incorporating technology into the country's national assessment system, the project facilitated standardization, increased transparency, and improved data-driven decisions in education in Guyana (Haggie, 2019).

Another unique aspect of the Guyana project is its blended approach to building capacity in test-writing. Training started with face-to-face workshops, followed by continuous in-system mentoring online through the GradeMaker Pro authoring system. It offers targeted feedback on live items, particularly higher-order thinking questions, and eliminates the need for frequent visits. Embedding training in the authoring process institutionalizes new practices in item writing, increases quality control, and promotes ongoing professional growth for local assessment experts (Haggie, 2019).

The Botswana Examinations Council (BEC) transitioned to the GradeMaker Pro system, which enabled the optimization of the exam paper creation process through secure remote access and efficient workflow management. This solution provides timely quality control, continuity of operations, and substantial reduction in costs related to organizing test development (Leigh, 2022).

Through the introduction of GradeMaker Pro, BEC has successfully transformed its entire test development cycle into a fully digital workflow. Since its introduction, the BEC has converted all its assessment cycles from paper-based assessments to complete digital test delivery and development. In addition to automating version control, user permissions and 2 factor authentication to provide a secure authoring environment, BEC has automated its workflows and milestones. The BEC experienced efficiency gains and a reduction in time delay for item development. The BEC will now explore the use of GradeMaker Pro's item bank functionality to improve and scale its assessments (Leigh, 2022).

Theme 2 (RQ2): fragmented and inefficient item development workflows

The primary theme that emerges from integrated data analysis is that the item-writing process is inefficient and disorganized. Despite the significance of NTC exams as high-stakes tests, there is little process in place to develop high-quality items.

Currently, item authors are most often selected by the NCT from a pool of secondary school teachers and university faculty. For a 10–14-day period, the authors were brought to the NTC or affiliated university centers. At the centers, the authors left their phones at the front desk and worked with the printed books and reference materials. The computers they work on are provided by the center and are not connected to the internet or a centralized item bank.

"Usually I come for 14 days, leave phones at the front door, and work only with printed books. It's exhausting and mistakes happen"—Expert 4.

Because item authors do not have access to an existing pool of items or a digital database, they sometimes create duplicates without knowing.

"Sometimes we write duplicates by accident because we can't access the bank or a previous version of the test,"—Expert 7.

Test authors are required to abide by the Center's confidentiality rules enforced by the Information Security Department. Only computers provided by the center are allowed, and no personal devices or Internet connections are permitted. Participants usually worked eight hours a day.

The authors are divided into 10–90 person teams that work in computer labs. The study was conducted both alone and in a group setting. Item authors developed questions in four formats: single-answer multiple choice, context-based single-answer, multiple answer, and matching questions. The authors are required to create 300–400 items per campaign, depending on the subject matter and their quota.

All authors were required to participate in a 36-h training program, with 13 h dedicated to theoretical foundations and 23 h

dedicated to item development. The author selection criteria included an appropriate academic degree, subject matter expertise, knowledge of assessment, and at least 3–5 to years of teaching experience. The lists of authors and subject matter experts were updated each year, in conjunction with regional educational authorities and higher education institutions.

The written items underwent two stages of moderation, both of which were manual. The reviews were conducted using word documents. Owing to the lack of data and metadata, the item history and item quality are difficult to trace.

'

"We receive documents with comments, but we can't see the entire history or any statistics."—Expert 3.

Test development experts also described how the work is conducted in silos.

'

"Actually, we can create item together, but it is time consuming. Every expert attempt to finish as quickly as possible. So usually everyone works in their own bubble."—Expert 9.

At present, items are written in static format and converted to HTML for computer-based testing. This step is manual and subject to errors and delays.

Documentary analysis supports these findings. For example, GradeMaker Pro and Cambridge Assessment Authoring include a centralized item bank, collaborative item writing environment, and real-time tracking features. These tools improve the quality assurance and help prevent duplications.

The lack of such tools in the NTC leads to inefficiency, additional workload, and inconsistent item quality. It also leads to decreased transparency and real-time feedback, which are necessary for reliable and valid high-stake tests.

The findings indicate that a centralized system with features, such as centralized authoring, item version control, metadata tagging, and collaborative authoring, results in better resource allocation, reduced error, improved item quality, and test fairness. Subsequent themes will explore the additional organizational, technical, and pedagogical issues needed to implement a new system.

Theme 3 (RQ2): emerging priorities for quality assurance and security

The second major theme that emerges from integrated data analysis is the emerging recognition within the NTC of the need to enhance quality assurance (QA) and security throughout the item development process. This theme is based on interviews and documentary data. The NTC is aware of the shortcomings of the current system, but is also open to adopting a more robust, technology-enabled solution that would be in line with international practice.

A key issue is the disjuncture between quality assurance and item development. As currently constructed, psychometric analysis occurs *post hoc* and separates item development and moderation workflows. Therefore, the authors and moderators do not have access to performance data during the authoring process.

'

"The UNT system has psychometric analysis, but it happens afterwards. Authors and moderators can't see item performance data during the development process."—Expert 5.

Another quality assurance expert said,

'

"We don't systematically track metadata like item difficulty or discrimination—it's done manually after administration."—Expert 9.

This state of affairs stands in stark contrast to what has been described in the international literature. International assessment programs with sophisticated platforms, such as GradeMaker Pro and TAO, allow psychometric data and metadata to be integrated into the authoring environment so that QA can occur in real time during item construction. Not only does this allow for higher-quality items, it also makes QA more efficient.

Interviewees also mentioned that manual workflow poses security vulnerabilities. The handling of Word files and the lack of audit trails expose the system to security breaches and access, which should not be allowed.

'

"Manual file handling, no audit trails—this is a big vulnerability."—Expert 8.

In addition, decentralized access control makes it difficult to ensure the confidentiality of high-stake exams. Security is at the heart of most modern exam assessment platforms that use features such as role-based access control, comprehensive audit trails, and encrypted data storage. The Botswana Examinations Council uses GradeMaker Pro, which controls who can see papers through role-based access control; authors can only see documents related to their role and the review phase. All revisions were tracked to prevent the generation of parallel test versions, and two-factor authentication and printless workflows-maintained item integrity (Botswana Examination Council, 2022).

Beyond quality and security concerns in the near term, NTC experts also expressed concerns regarding the future sustainability of the UNT system. Recent reforms have expanded the domains covered by the test and added new item formats such as components of adaptive testing. It is clear that manual processes will not be able to meet the needs of the UNT program.

'

"We need much larger and more flexible item banks. Without automation, we can't meet the demand."—Expert 2.

International literature supports this view. Gierl et al. (2021) pointed out that the need to support adaptive testing and frequent test revisions calls for item banks that can scale, which in turn requires platforms with an AIG and metadata-driven test assembly. Given this demand, the NTC's current workflow for manual item writing and static moderation will not work.

The other issues mentioned in the interviews were expert fatigue and inconsistency. As discussed above, item authors work in silos with very limited access to NTC archives and no access to performance data, which is a strain on the cognitive load and quality.

'

"We come here and work long hours with no access to prior materials or psychometric feedback. It's exhausting and frustrating."—Expert 3.

Such conditions do not allow for iterations and introduce inconsistencies between the test forms. Thus, the findings suggest that in addition to technological upgrades, the organization of processes around item development needs to change. Making the QA part of the item development workflow, strengthening security, and managing an item bank at scale are steps toward improving the UNT system.

Theme 4 (RQ3): functional and organizational requirements for platform adoption

According to the analysis of international experience, the main principles of modern measurement theory, the research results in domestic practice, and the main elements of the test item development process by the NCT, the main requirements for developing a digital platform for managing the lifecycle of examination materials were determined. They include both functional features directly related to the process of item development, validation, and expert review as well as non-functional features that ensure the reliability, security, and user-friendliness of the system.

Digital workflow

According to the results of studies by Cen et al. (2010) and Hegde et al. (2018), automating and structuring the test development process leads to increased efficiency and a reduction in the burden on people. The practical experience of using the GradeMaker system (Frederick and Grammar, 2022) demonstrates the possibility of creating a full-cycle digital platform. In addition, the need to switch to a digital platform is determined by the need to move from a scattered development process to a unified and controlled environment. Simultaneously, a key requirement is to implement a scalable digital workflow that includes all stages of the test material life cycle, starting from the initial creation, continuing with multi-level editing, expert review by different profiles, pilot testing with the target audience, and centralized management. The workflow process should be structured to strictly regulate the participation of all parties involved in the test development process using a role-based control system and differentiated access management. The possibility of integrating documentation accompanying the tests (test specifications and methodological guidelines) and functionality for managing deadlines and subsequent progress is also an important condition.

To determine the requirements for a platform, Sagindikov et al. (2022) was considered, emphasizing the need to unambiguously and clearly formulate the expected behavior of the test-taker when interacting with the item. This should be performed as clearly and simply as possible, avoiding additional information that would reduce the probability of ambiguity in understanding and interpreting the conditions of the task (Sagindikov et al., 2022). Moreover, the implementation of a mechanism for remote interaction between developers and experts is considered to be a strategically important element that provides flexibility and efficiency in the development process.

Test item development functionality

The functional architecture of the platform must support the development of test items of various types and levels of complexity including integral test instruments and discrete items (Circi et al., 2023; Irvine and Kyllonen, 2002). The development subsystem must contain specialized tools for integrating and visualizing mathematical and scientific formulas (with support for rendering digital ones to avoid distortion in display), as well as functionality for assigning structured taxonomies.

Simultaneously, the platform must support the creation and storage of such taxonomies if it creates and stores them. Development of assessment materials in different formats is a basic requirement (test booklets, scoring guides, inserts, and confidential digital instructions). The ability to change the order of questions in the test and create structured items (e.g., multi-component tasks used in the final certification) is also necessary. The user interface must contain standard text-editing tools (text formatting and adding special characters) and table tools (creation, editing structure, and properties). Support for the development and storage of scoring schemes directly connected to the content of the item, as well as the ability to add notes and glossaries to the developed materials and answers in the scoring schemes, will ensure the transparency and standardization of the scoring process (1EdTech, 2022; APA, 2014). The platform must support the development of various types of items within both single and structured tasks, the development of scoring scheme content (data and points), and basic functionality for changing visual appearance (layout, fonts). The integration of advanced tools (e.g., for creating interactive elements) with a preview function in a computer-rendered format is critically important. Support for the compatibility of developed materials with the IMS Question and Test Interoperability (QTI) standard and the development of all item types and their variations included in the standard and applicable to the system (including special types for computer-based testing) are key requirements for ensuring interoperability and compliance with international standards (1EdTech, 2022).

Validation

The integration of an automated validation subsystem is an important requirement to ensure that the developed materials comply with the established quality criteria (Falcão et al., 2023). The platform must validate developed tests and individual items to comply with the agreed content coverage rules (e.g., distribution of scores by topic and skill). Within the digital workflow, the system should control the completion of tasks in accordance with established rules. The availability of information for users regarding their degree of compliance with item development rules in real time will help increase the efficiency and quality of work. In this way, the automatic system needs to be related to the main sources of validity evidence: (a) test content, in terms of providing appropriate sampling of curricular domains; (b) response processes, in terms of monitoring items as elicitors of the cognitive operations intended

by the test developer; (c) internal structure, in terms of congruence among items and scales; (d) relations of test scores to other variables, such as external criteria or past achievement; and (e) consequences of testing, in terms of the general impact of the quality of items on fairness and decision making (APA, 2014).

Item templates

To optimize the development process, especially for complex and multimedia items, and ensure compliance with international standards, the platform must support the creation and use of item templates aligned with the IMS QTI standard (1EdTech, 2022). Functional requirements include the ability to create customizable templates for various item types (including multimedia), configure and filter metadata and taxonomies available for assignment during development, assign agreed taxonomies to content in advance, set font style and size parameters for rendering templates, set allowable question types and score ranges, and set content coverage rules based on taxonomies for automated content validation (Gierl and Haladyna, 2013; Hedden, 2018).

Media resources and copyright management

If item development involves the use of multimedia elements, the platform must provide developers with controlled access to an integrated media-resource library. Functional requirements include mandatory specification of copyright information for all images and resources used, automated verification of such information, logging of copyright declarations, prevention of publishing materials without proper authority, and the ability to track copyright information for resources stored in the stock image library.

Expert review process

To ensure high-quality examination materials, the platform must support a multi-stage review process (Khafagy et al., 2016). Functional requirements include the ability to review both the entire test and individual items, tools for adding comments and notes to developed materials, specialized functionality for checking the correct display of mathematical formulas and scientific symbols in a digital format, review of materials in different formats (test booklets and scoring guides), and review of structured items used in the final certification.

Item bank management and test assembly

Storage of items in a structured item bank is an important requirement (Gierl and Haladyna, 2013). The platform must support the assembly of tests from materials entered into the item bank, according to the agreed rules (test specifications). Functional requirements include the ability to search for items using metadata and content coverage rules, automated validation of the assembled tests, automatic monitoring of the life cycle status of the item, monitoring of change history, support for semi-automation (pre-selection of items), fully automated test assembly, monitoring of the progress of the assembly process, filtration of search results by metadata, item preview before including them in the test, addition/removal and reordering of items, preview of the assembled tests and scoring guides, generation of content reports for assembled tests based on agreed taxonomies, and a user-friendly interface for assembling tests.

Rules for item bank usage

To effectively and safely use an item bank, the platform must support a flexible system to manage item reuse rules. Functional requirements include editing and defining the ability to reuse items, automatic placement of already used items for reuse, defining reuse rules for items in tests, authorization for reuse and change of materials for other purposes, automated control of reuse permissions or multiple inclusion of items in a test, the ability to retrieve and store information about the content of the assessment materials of other systems, provision of developers with information about used items, tracking item usage, exchange of items between tests/modules, and cloning of existing materials.

Item bank management

In addition to the specified requirements, the item bank management functionality includes the definition of "incompatible" and "linked" items, editing of content containing cultural characteristics and related taxonomies, editing of search keywords, sending items into a new workflow directly from the search interface, displaying the test elements entered into the item bank, and analyzing the item bank's readiness and compliance with requirements.

Scoring, analysis, and evaluation

The platform must support integration with external systems for the scoring, analysis, and evaluation of test results. Functional requirements include the export of scoring schemes and test structures to appropriate applications, configuration of automated scoring and reference data for manual scoring, and the export of item usage statistics.

Non-functional requirements

In addition to functional capabilities, the platform must meet a range of non-functional requirements to ensure reliable, safe, and user-friendly operations (Kong et al., 2022). They include:

Security: two-factor authentication.
Performance: fast response time and high data volume support.
Reliability: stability of the system and data storage.
Usability: user-friendly interface and ergonomic design.
Scalability: expandable functionality and user capacity.
Data volume: archiving functions accessibility: end-user support and documentation availability.

In conclusion, the transition from traditional methods to digital solutions should be based on the real needs of the academic environment and consider the infrastructure, personnel, and regulatory peculiarities of the higher education system in Kazakhstan. Moreover, the participants emphasized the

importance of securing strong institutional leadership support and implementing a staged change management approach to facilitate acceptance and minimize resistance.

'

"Without proper change management and training, a new platform will fail—people need to see the benefits and feel supported."—Expert 12

The findings demonstrate that, while the NTC faces significant technical and organizational challenges, there is a strong institutional commitment to updating item development practices. Aligning the UNT system with international best practices will require the implementation of an advanced digital platform and the cultivation of the organizational capacities needed to sustain its use.

Discussion

This study reveals that item development for the UNT falls short of international standards of quality, efficiency, and security in design and practice. Although recent improvements have enabled a digital approach to test administration and supervision, test development remains a laborious process. This disparity undermines the broader objective of establishing a transparent, reliable and equitable national assessment system (Abdrasilov et al., 2024).

Analyses of digital assessment transformations that use GradeMaker platforms in a number of international contexts reveal a common pattern where successful efforts have typically combined elements of secure workflow and item banking systems, along with investments in staff capacity building. For example, centralized item banks and automated workflows have been important for countries like Zimbabwe and Botswana to improve security and speed of delivery in response to exam leakage and prolonged processing delays (Frederick and Grammar, 2022; Leigh, 2022). Professional development interventions that are embedded in the assessment authoring workflow have also proven vital for supporting digital transitions, as shown in the Guyana case (Haggie, 2019). An example of the scale and long-term platform maturation that can be realized in this space is also described in the UK context (AQA) (Leigh, 2023). These examples broadly align with the perspectives of stakeholders in the UNT case, who also expressed significant concerns related to security and efficiency alongside limited staff readiness for these processes. The UNT experience can thus both inform and be informed by broader global experiences. In addition to technology adoption, Kazakhstan's digital shift will likely need to be supported by investments in quality assurance and staff professional development that are institutionalized in policy and practice over the long term.

Our analysis explains why the extremely fragmented and inefficient item development process at NTC matters in practice. Isolated authoring, the absence of a centralized item bank, and the absence of a shared database of metadata and psychometric information characterize the current state of item development at the NTC. This finding is consistent with prior research that emphasizes the role of digitally authoring environments in facilitating consistency, minimizing redundancy, and supporting collaborative item development (Gierl et al., 2021, 2022; Gierl and Lai, 2016a,b). Following Kane's (2013) call for greater focus on systematic quality assurance, the UNT case illustrates how the structural fragmentation can adversely affect the trustworthiness of high-stakes testing systems.

The three key contributions of this study are as follows: (1) the first empirical account of test development processes in Kazakhstan's national testing system undergoing digital modernization, an understudied setting in the AIG and assessment modernization literature; (2) a hybrid requirement framework based on both documentary and field data that integrates global standards with local considerations; and (3) a stakeholder-informed roadmap for implementation that recognizes the interdependence between technological design, organizational capacity, and professional culture.

As indicated earlier, another important policy implication of this research is the additional workload for item authors in their isolated and restricted work conditions. This leads to expert burnout, inefficiency, and the potential for item duplication, which are all issues widely documented in the literature as impediments to high-quality assessment development (Mohd Noor et al., 2019; Redecker and Johannessen, 2013). In addition, the lack of access to psychometric indicators during authoring and moderation limits item developers' ability to continually improve assessment quality, which has been observed in studies of best practices in formative item evaluation (Cen et al., 2010; Gierl et al., 2022). This outcome points to the necessity for digital platforms for the UNT to be equipped with analytics and feedback mechanisms that support, rather than further burden, the work of item developers instead of merely going digital with existing processes.

Security also emerges as a key policy concern. Manual file management and the absence of role-based access controls or audit trials render the assessment process vulnerable to integrity issues. This vulnerability has been observed in other national testing systems, transitioning from paper-based to digital assessments, where improved security is achieved through encryption, authentication protocols, and change histories (Cambridge Assessment, 2023; Frederick and Grammar, 2022; Sychev et al., 2024). In the case of Kazakhstan, exam security needs to be understood not just as a technical issue, but as a credibility-enhancing project that underlies public confidence in the UNT.

There is institutional awareness at the NTC about these shortcomings and a clear understanding that reforms are necessary, but the successful implementation of a digital item development platform requires more than just technological procurement. Organizational capacity, which includes staff training, leadership commitment, and governance clarity, is equally important. These findings are consistent with empirical studies that show that staff training and clear governance are critical success factors in digital transformation projects (Leigh, 2023; Sychev et al., 2024). It follows that policy makers need to time reforms judiciously so that financial investment in human capital and institutional architecture keep pace with the introduction of technology.

Finally, the platform developed by NTC must be functionally flexible to support a range of item types, including multimedia and adaptive formats, while maintaining compatibility with international interoperability standards (Abdrasilov et al., 2024; Sychev et al., 2024). Otherwise, the system will become obsolete

as international testing standards evolve. This highlights the importance of future-proofing reforms by anticipating not only current needs but also emerging global trends in assessment design.

When comparing the interview results by stakeholder groups, no major differences were found. However, some minor differences in focus were identified. For example, test developers were more prone to point out issues and shortfalls of the current item development system, such as its inefficiency and lack of quality control, and to emphasize the need for improvement of the workflow and working conditions. At the same time, interviewees from the Ministry of Education were more likely to mention the progress made so far and the benefits of the overall system modernization. Such nuances require reform strategies that accommodate the concerns of implementers and the priorities of policymakers so that both of their perspectives can shape further digital transformation's next steps.

To conclude, this study reinforces the notion that digital transformation in assessment is a holistic endeavor that requires coordinated reform across technological, organizational, and pedagogical domains. By drawing on international best practices and the lived experience of Kazakhstani test developers, this study contributes to global discussion on enhancing the integrity and efficiency of large-scale national assessments through digital innovation.

Limitations

The study's limitations include its qualitative design, purposive sampling of NTC-affiliated stakeholders that may not reflect other educational authorities or end-users in schools, and the platform's lack of development at the time of research. In addition, there is a need for future research to examine the platform's impact on item quality, workflow efficiency, and user satisfaction after its implementation. Further, a comparative study with other Central Asian or developing countries that have undergone similar national assessment reforms would provide contextual insights into the digital innovation of national testing in Kazakhstan.

Conclusion

This study aimed to explore the validity of using a digital platform for item development to meet the operational needs of Kazakhstan's Unified National Testing. Document review and semi-structured interviews with NTC stakeholders were conducted to identify existing inefficiencies, quality assurance gaps, and critical considerations for platform adoption.

The research found that the current item development workflow is fragmented, paper-based, and lacks automated tracking of psychometric feedback and metadata. No strong security or version control mechanisms are in place for item development. These inadequacies create barriers to Kazakhstan's efforts to achieve international best practices in terms of fairness, validity, and scalability of national assessments.

The study also highlighted that the successful implementation and use of the platform are dependent not only on technical feasibility but also on other important factors such as organizational capacity, staff training, change management, and governance structures. The results show that a move toward a digital platform for item development is highly encouraged for national assessments in Kazakhstan.

The study suggests that the Ministry of Science and Higher Education of Kazakhstan begin with a gradual nd phased pilot approach in selecting subject domains and regions, while also developing clear guidelines and policies on metadata standards and training requirements for item authors. Also, the future platform development should consider localization language support, adherence to Question and Test Interoperability (QTI) standards, and integration with a psychometric dashboard for long-term scalability and impact.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the Ethic Committee of Karaganda Buketov University. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

ND: Conceptualization, Funding acquisition, Supervision, Writing – review & editing. YS: Writing – original draft, Formal analysis, Methodology, Investigation, Writing – review & editing. SA: Writing – original draft, Investigation, Data curation. BA: Writing – review & editing, Project administration, Validation. AK: Validation, Methodology, Writing – original draft.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

1EdTech (2022). *Question and Test Interoperability (QTI) v3.0: Final Release.* Available online at: https://www.imsglobal.org/spec/qti/v3p0/oview#contributors (Accessed August 26, 2025).

Abdrasilov, B., Niyazov, T., Baizhaov, N., Altybaeva, Sh., Iskakova, A., and Umerbaeva, G. (2024). *Unified National Testing: History, Experience, Perspective*. Astana: National Test Center.

Amanzhol, N., Amanova, A., Kerimbekova, B., Zholmakhanova, A., and Sarmurzin, Y. (2024). "My expectation did not meet reality": challenges of undergraduate students in English-medium instruction in Kazakhstan. *Asian Educ. Dev. Stud.* 13, 31–44. doi: 10.1108/AEDS-06-2023-0062

Anuarbekova, A., and Sultan, A. (2025). Kazakhstani high school students' preference for ielts over the English exam on unified national testing. *Universum: Psychol. Educ.* 131:19818. doi: 10.32743/UniPsy.2025.131.5.19818

APA, AERA, and NCME (2014). *Standards for Educational and Psychological Testing (2014 Edition)*. Available online at: https://www.aera.net/publications/books/standards-for-educational-psychological-testing-2014-edition (Accessed August 26, 2025).

Avanesov, V. S. (2005). *Theory and Methods of Educational Measurement*. Available online at: https://charko.narod.ru/tekst/biblio/Avanesov_Teoriya_i_metod_ped_izmer.pdf (Accessed June 19, 2025).

Bakas uulu, B., and Smagulov, Y. (2016). Analysis of dynamics of high school graduates who participated in the unified national test Kazakhstan. *Int. Electron. J. Math. Educ.* 11, 3176–3186. Retrieved from: https://www.iejme.com/article/analysis-of-dynamics-of-high-school-graduates-who-participated-in-the-unified-national-test

Bormuth, J. R. (1970). *On the Theory of Achievement Test Items.* Chicago, IL: University of Chicago Press.

Botswana Examination Council (2022). *Annual Report 2021/22.* Retrieved from: https://www.bec.co.bw/images/2023/annual-reports/annual-report-03-11-22.pdf

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 77–101. doi: 10.1191/1478088706qp063oa

Cambridge Assessment (2023). *Annual Report*. Retrieved from: https://www.cambridge.org/sites/default/files/media/documents/Annual%20Report%202023-24.pdf

Cen, G., Dong, Y., Gao, W., Yu, L., See, S., Wang, Q., et al. (2010). A implementation of an automatic examination paper generation system. *Math. Comput. Model.* 51, 1339–1342. doi: 10.1016/j.mcm.2009.11.010

Chan, K. W., Ali, F., Park, J., Sham, K. S. B., Tan, E. Y. T., Chong, F. W. C., et al. (2025). Automatic item generation in various STEM subjects using large language model prompting. *Comput. Educ. Artif. Intell.* 8:100344. doi: 10.1016/j.caeai.2024.100344

Circi, R., Hicks, J., and Sikali, E. (2023). Automatic item generation: foundations and machine learning-based approaches for assessments. *Front. Educ.* 8:858273. doi: 10.3389/feduc.2023.858273

Creswell, J. W., and Poth, C. N. (2016). *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*. London: Sage publications.

Drasgow, F., Luecht, R. M., and Bennett, R. E. (2006). "Technology and testing," in *Educational Measurement*, ed. R. L. Brennan (London: Praeger Publishers), 471–516.

egov.kz. (2024). *SNT in Kazakhstan: Preparation and Order of Conduction | Electronic Government of the Republic of Kazakhstan*. Available online at: https://egov.kz/cms/en/articles/about_ent (Accessed June 16, 2025).

Embretson, S., and Yang, X. (2006). 23 Automatic item generation and cognitive psychology. *Handb. Stat.* 26, 747–768. doi: 10.1016/S0169-7161(06)26023-1

Falcão, F., Pereira, D. M., Gonçalves, N., De Champlain, A., Costa, P., and Pêgo, J. M. (2023). A suggestive approach for assessing item quality, usability and validity of Automatic Item Generation. *Adv. Health Sci. Educ.* 28, 1441–1465. doi: 10.1007/s10459-023-10225-y

Falcão, F. M. V., Pereira, D. S. M., Pêgo, J. M., and Costa, P. (2024). Progress is impossible without change: implementing automatic item generation in medical knowledge progress testing. *Educ. Inf. Technol.* 29, 4505–4530. doi: 10.1007/s10639-023-12014-x

Frederick, M., and Grammar, C. (2022). *Moving Towards a Digital Question Paper Development Process: The Zimbabwe School Examinations Council Experience with The GradeMaker*. Livingstone: Annual AEAA Conference.

Gangar, F. K., Gori, H. G., and Dalvi, A. (2017). Automatic question paper generator system. *Int. J. Comput. Appl.* 66, 42–47. doi: 10.5120/ijca20179014138

Gierl, M., Lai, H., and Tanygin, V. (2021). "Advanced methods in automatic item generation," in *Advanced Methods in Automatic Item Generation*, eds. M. J. Gierl, H. Lai, and V. Tanygin (New York, NY: Routledge), 1–233. doi: 10.4324/9781003025634

Gierl, M., Shin, J., Firoozi, T., and Lai, H. (2022). Using content coding and automatic item generation to improve test security. *Front. Educ.* 7:853578. doi: 10.3389/feduc.2022.853578

Gierl, M. J., and Haladyna, T. (2013). *Automatic Item Generation: Theory and Practice*. Availabl online at: https://www.routledge.com/Automatic-Item-Generation-Theory-and-Practice/Gierl-Haladyna/p/book/9780415897518?srsltid=AfmBOopf7CPi0HqPH_SG_3i2oOsO1OyISHcYu06M0iklXcc6LRR4xcXD (Accessed June 17, 2025).

Gierl, M. J., and Lai, H. (2016a). "Automatic item generation," in *Handbook of Test Development*, 2nd Edn., eds. S. Lane, M. Raymond, and T. Haladyna (London: Routledge), 410–429.

Gierl, M. J., and Lai, H. (2016b). "The role of cognitive models in automatic item generation," in *The Handbook of Cognition and Assessment*, eds. A. A. Rupp, and J. P. Leighton (Hoboken, NJ: Wiley), 124–145. doi: 10.1002/9781118956588.ch6

Hadzhikoleva, S., Rachovski, T., Ivanov, I., Hadzhikolev, E., and Dimitrov, G. (2024). Automated test creation using large language models: a practical application. *Appl. Sci.* 14:9125. doi: 10.3390/app14199125

Haggie, D. (2019). "Technology and education reform-the case study of Guyana," in *45th Annual Conference* (Baku).

Hedden, H. (2018). Taxonomies and metadata for digital asset management. *J. Digit. Media Manag.* 6:380. doi: 10.69554/OFDP6137

Hegde, V., Rao, L. V., and Shivali, B. S. (2018). The framework for web-based automated online question paper generator through JEE. *Int. J. Eng. Technol.* 7, 1415–1419. doi: 10.14419/ijet.v7i3.13573

Irvine, S. H., and Kyllonen, P. C. (2002). *Item Generation for Test Development*. Routledge. Available online at: https://www.routledge.com/Item-Generation-for-Test-Development/Irvine-Kyllonen/p/book/9781138973473?srsltid=AfmBOopMf8DvMeDD6Xom4hKTdRbzo7kqmB_M5WaqXDxIqaKsMoj4O8PJ (Accessed August 26, 2025).

Jumabayeva, Z. (2016). The key drivers of the Unified National Test in Kazakhstan: a critical analysis of its impact on school leavers. *NUGSE Res. Educ.* 1, 16–20. Retrieved from: http://nur.nu.edu.kz/handle/123456789/2088

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Meas.* 50, 1–73. doi: 10.1111/jedm.12000

Kazakhstan Educational Research Association (2020). *An Ethical Code of Researchers in Education.* Astana: Kazakhstan Educational Research Association.

Khafagy, G., Ahmed, M., and Saad, N. (2016). Stepping up of MCQs' quality through a multi-stage reviewing process. *Educ. Prim. Care 27*, 299–303. doi: 10.1080/14739879.2016.1194363

Kirichok, O., Amankeldiyeva, S., and Nussenov, Z. (2025). Students' perceptions of unified national test scores, state scholarships, and academic performance. *World J. Educ. Technol. Curr. Issues* 17, 1–13. doi: 10.18844/wjet.v17i1.7695

Kiyak, Y. S., Coşkun, Ö., Budakoglu, I. I., and Uluoglu, C. (2024). ChatGPT for generating multiple-choice questions: Evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *Eur. J. Clin. Pharmacol.* 80, 729–735. doi: 10.1007/s00228-024-03649-x

Klammer, C., and Ramler, R. (2017). "A journey from manual testing to automated test generation in an industry project," in *Proceedings - 2017 IEEE International Conference on Software Quality, Reliability and Security Companion, QRS-C* (Prague: IEEE), 591–592. doi: 10.1109/QRS-C.2017.108

Kong, L. Y., Zeng, Q. T., Zhu, X. F., and Lu, L. K. (2022). "Design and realization of an online examination system based on maven framework," in *Proceedings - 2022 3rd International Conference on Computer Science and Management*

Technology, ICCSMT (Shanghai: IEEE), 56–59. doi: 10.1109/ICCSMT58129.2022.00019

Kurdi, G., Leo, J., Parsia, B., Sattler, U., and Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *Int. J. Artif. Intell. Educ.* 30, 121–204. doi: 10.1007/s40593-019-00186-y

La Russa, F., Marzoli, R., Mastrogiovanni, A., and Mattei, A. (2025). "Automated item generation approaches in educational testing: a systematic review protocol and preliminary findings," in *EDULEARN25 Proceedings, 2007—2007* (Palma). doi: 10.21125/edulearn.2025.0581

Leigh, A. (2022). *Botswana Examinations Council (BEC)*. Available online at: https://www.grademaker.com/customer-stories/botswana-examinations-council-bec/?utm_source=chatgpt.com (Accessed June 20, 2025).

Leigh, A. (2023). *Educational Charity AQA Acquires GradeMaker*. Available online at: https://www.grademaker.com/news/educational-charity-aqa-acquires-grademaker/?utm_source=chatgpt.com (Accessed June 20, 2025).

Mingisheva, N. (2023). Development and challenges of standardized testing in Kazakhstan: transition from national to international standards. *KazNU Bull. Pedag. Series* 76, 94–103. doi: 10.26577/JES.2023.v76.i3.08

Mohd Noor, N., Mohd Napi, N., Farzana, I., and Amin, I. (2019). The development of autonomous examination paper application: a case study in UiTM cawangan perlis. *J. Comput. Res. Innov.* 4, 21–30. doi: 10.24191/jcrinn.v4i2.105

Naz, N., Gulab, F., and Aslam, M. (2022). Development of qualitative semi-structured interview guide for case study research. *Competitive Soc. Sci. Res. J.* 3, 42–52. Retrieved from: https://cssrjournal.com/index.php/cssrjournal/article/view/170

OECD (2017). *Higher Education in Kazakhstan 2017, Reviews of National Policies for Education*. Paris: OECD

Patton, M. Q. (2002). *Qualitative Research and Evaluation Methods*, 3rd Edn. London: Sage.

Pauli, M., and Ferrell, G. (2020). *The Future of Assessment: Five Principles, Five Targets for 2025*. Retrieved from: https://repository.jisc.ac.uk/7733/1/the-future-of-assessment-report.pdf

primeminister.kz (2023). *162 thousand entrants take Unified National Testing*. Available online at: https://primeminister.kz/en/news/reviews/162-thousand-entrants-take-unified-national-testing-24760 (Accessed June 18, 2025).

Pugh, D., De Champlain, A., Gierl, M., Lai, H., and Touchie, C. (2016). Using cognitive models to develop quality multiple-choice questions. *Med. Teach.* 38, 838–843. doi: 10.3109/0142159X.2016.1150989

Redecker, C., and Johannessen, Ø. (2013). Changing assessment - towards a new assessment paradigm using ICT. *Eur. J. Educ.* 48, 79–96. doi: 10.1111/ejed.12018

Roid, G. H., and Haladyna, T. M. (1978). A Comparison of objective-based and modified-bormuth item writing techniques. *Educ. Psychol. Meas.* 38, 19–28. doi: 10.1177/001316447803800104

Ruslin, Mashuri, S., Rasak, M., Alhabsyi, F., and Syam, H. (2022). Semi-structured interview: a methodological reflection on the development of a qualitative research instrument in educational studies Ruslin. *J. Res. Method Educ.* 12, 22–29. Retrieved from: https://www.iosrjournals.org/iosr-jrme/papers/Vol-12%20Issue-1/Ser-5/E1201052229.pdf

Russell-Lasalandra, L. L., Christensen, A. P., and Golino, H. (2024). *Generative Psychometrics via AI-GENIE: Automatic Item Generation and Validation via Network-Integrated Evaluation*. OSF. doi: 10.31234/osf.io/fgbj4

Sagindikov, I., Zhumazhanova, S., Auezkhanova, A., and Tasbulatova, M. (2022). *Improvement and Stimulation*. Astana: Altynsarin National Academy of Education.

Sarmurzin, Y., Amanzhol, N., Toleubayeva, K., Zhunusova, M., and Amanova, A. (2021). The impact of OECD research on the education system of Kazakhstan. *Asia Pac. Educ. Rev.* 22, 757–766. doi: 10.1007/s12564-021-09715-8

Sarmurzin, Y., Amanzhol, N., Toleubayeva, K., Zhunusova, M., Amanova, A., and Abiyr, A. (2024). Challenging aspects of Kazakhstan's trilingual education policy: evidence from a literature review. *Asia Pac. Educ. Rev.* 25, 801–811. doi: 10.1007/s12564-023-09823-7

Sarmurzin, Y., Kerimbekova, B., Toleubaeva, K., Zhunusova, M., Amanzhol, N., and Zhumagulov, A. (2025). "Forced us to adopt alternative pedagogical approaches": Kazakhstani reading literacy improvement initiatives. *Res. Comp. Int. Educ.* 20, 397–421. doi: 10.1177/17454999251342142

Schmucker, R., and Moore, S. (2025). The impact of item-writing flaws on difficulty and discrimination in item response theory. *arXiv*. https://arxiv.org/pdf/2503.10533

Song, Y., Du, J., and Zheng, Q. (2025). Automatic item generation for educational assessments: a systematic literature review. *Interact. Learn. Environ.* 1–20. doi: 10.1080/10494820.2025.2482588

Stevens, S., Shin, N., Delgado, C., Krajcik, J., and Pellegrino, J. (2007). *Using Learning Progressions to Inform Curriculum, Instruction and Assessment Design*. New Orleans, LA: NARST Annual Conference.

Sychev, O., Prokudin, A., and Denisov M. (2024). Automatic generation of tasks based on the results of analysis of the use of their bank in an intelligent training system. *Softw. Syst.* 37, 201–212. doi: 10.15827/0236-235X.146.201-212

Tan, B., Armoush, N., Mazzullo, E., Bulut, O., and Gierl, M. J. (2024). *A Review of Automatic Item Generation Techniques Leveraging Large Language Models*. doi: 10.35542/osf.io/6d8tj

Tengrinews.kz. (2024). *UNT Results of Several Applicants Annulled After Video Review*. Available online at: https://tengrinews.kz/newseducation/rezultatyi-ent-neskolkih-abiturientov-annulirovali-proverki-538427/

testcenter.kz (n.d.). *Unified National Testing*. Available online at: https://testcenter.kz/en/shkolnikam/ent/edinoe-natsionalnoe-testirovanie-ent/ (accessed June 16, 2025).

testcenter.kz. (2025). *The Main UNT of 2025 Began*. Available online at: https://testcenter.kzru/press-tsentr/novosti/detail.php?ID=7640

von Davier, M. (2019). Training optimus prime, M.D.: generating medical certification items by fine-tuning OpenAI's gpt2 transformer model. *arXiv*. https://arxiv.org/pdf/1908.08594 (Accessed June 8, 2025).

Yazan, B. (2015). Three approaches to case study methods in education: Yin, Merriam, and Stake. *Qual. Rep.* 20, 134–152. doi: 10.46743/2160-3715/2015.2102

Yermaganbetova, D. (2025). *UNT Results Cancelled for 142 School Graduates*. Available online at: Kursiv.Kz. https://kz.kursiv.media/2025-06-02/dnrm_test_ent/ (Accessed June 16, 2025).

Yin, R. (2014). *Case Study Research Design and Methods*, 5th Edn. London: Sage.