International Journal of Evaluation and Research in Education (IJERE)
Vol. 15, No. 5, October 2026, pp. 4068~4079
ISSN: 2252-8822, DOI: 10.11591/ijere.v15i5.39221

Evaluating integrity and governance efficiency in blockchain-supported online assessment

Aivar Sakhipov[1], Laura Smagulova[2], Aigul Yelepbergenova[2], Alexey Fedenko[1]
[1]School of Software Engineering, Astana IT University, Astana, Kazakhstan
[2]Department of Information Technologies and Artificial Intelligence, I. Zhansugurov Zhetysu University, Taldykorgan, Kazakhstan

Article Info

Article history:

Received Feb 24, 2026
Revised Jun 5, 2026
Accepted Aug 5, 2026

Keywords:

Auditability
Blockchain
Governance efficiency
Online assessment
Procedural fairness
Trust

ABSTRACT

With the increasing deployment of remote proctoring systems, there is a noticeable gap in online testing governance due to the lack of credible evidence, which might have adverse implications in terms of trust between the testing platform and its participants. This study, therefore, examines the potential impact of blockchain-powered evidence management on auditability and audit process efficiency in the context of online examinations. In our study, we used an experimental research method and examined the impact of tamper-proof distributed ledger technology on auditability and audit time using a sample of 160 exam sessions divided equally into two groups: a control group (CG, n=80) and an experimental group (EG, n=80). The findings suggest a positive and statistically significant influence of blockchain adoption on auditability (12.5% higher), reaching 96.3% compared to 83.8% in the CG, as well as a 37% reduction in audit time. Furthermore, the inter-rater agreement level reached 0.846. Notably, based on our surveys, test-takers find the process more credible and legitimate without any increased privacy concerns.

Corresponding Author:

Laura Smagulova
Department of Information Technologies and Artificial Intelligence, I. Zhansugurov Zhetysu University
187a, Ilyas Zhansugurov Street, Taldykorgan, Kazakhstan
Email: smagulovalaura272@gmail.com

1. INTRODUCTION

Online assessments have now become a fundamental tool for assessing learning outcomes, certifying students, and providing flexible options for students' educational pathways across educational institutions and digital learning environments [1]. In addition, with increasing numbers of high-stakes assessments being administered remotely; there is an increasing trend toward using remote proctoring and automated monitoring systems to maintain the pedagogical value and social recognition of academic certifications [2], [3]. Although education research indicates that the credibility of online assessments does depend on the ability to detect cheating or other dishonest behavior, it also indicates that credible online assessments depend on assessment governance, which includes clear procedural rules, reviewable evidence, and defensible processes for making decisions regarding students [4]. As such, credible online assessments support achieving both of the United Nations' sustainable development goals (SDG) 4–quality education by ensuring fair and reliable evaluations of students, and SDG 16–peace, justice, and strong institutions by promoting transparent and accountable decision-making processes within educational institutions [5].

The tension between protecting the integrity of remote proctoring (and therefore the results) and providing an optimal learning environment for students has been at the center of most research on remote proctoring [6], [7]. Studies have shown that while remote proctoring technology can help prevent cheating and provide institutions with ways to scale their online assessments, these same studies have reported issues related to how students perceive the level of surveillance they receive during remote proctored exams [8], [9]; privacy; anxiety; and whether or not they feel the exam is legitimate. Furthermore, many of the alerts generated by automated flagging mechanisms used in remote proctoring systems are based on what may appear to be anomalies in a student's behavior that could occur in actual testing environments–e.g., connection loss or variations in environmental conditions. These factors can lead to unnecessary reviews of false positives and therefore potentially harm confidence in decision-making processes within educational institutions [10], [11]. Therefore, fair and transparent practices need to be considered top-level regulatory outcomes rather than merely secondary technological characteristics [12].

Procedural justice has been a growing area of study within the use of digital assessment systems [13], [14] with the degree to which an individual perceives the assessment as fair being influenced by how easily one can understand the rules of the assessment, how explainable decisions are, and whether the review process is consistent, objective and based on evidence [15], [16]. However, most of these digital assessments use log files from central platforms for collecting data which can be difficult to verify independently, leaving students vulnerable in cases of technical errors or disputed findings. This lack of verifiable evidence creates a 'black box' in assessment governance, where the power imbalance between the institution and the learner is most evident. Consequently, the potential for incomplete or unverifiable data in appeal or audit situations could significantly reduce the ability to provide sufficient evidential support for decision making at the institutional level [17]. Therefore, the core education problem goes far beyond identifying suspicious activity, and will require developing tamper-evident and auditable workflows to enable credible reviews and conflict resolutions [18].

Previous studies on remote proctoring and online assessment have primarily focused on detection accuracy, cheating prevention, privacy concerns, student anxiety, and perceptions of surveillance. At the same time, blockchain-based approaches in education have mainly emphasized credential verification, certificate authenticity, and secure academic recordkeeping [19], [20]. In parallel, existing blockchain audit-trail frameworks demonstrate the technical feasibility of tamper-evident logging and verifiable evidence preservation [21], [22]. However, these studies do not evaluate how tamper-evident evidence workflows affect governance-related outcomes within online assessment environments. In particular, empirical evidence remains limited regarding whether verifiable evidence chains improve auditability, reduce audit effort, support reviewer consistency, and increase perceived legitimacy of assessment procedures under controlled experimental conditions [23].

To address this gap, this study evaluates a blockchain-supported evidence workflow integrated into an online assessment system using a controlled comparative design. The blockchain layer serves as an integrity mechanism for verifying and auditing assessment evidence while applying data-minimization principles through the use of pseudonymous identifiers, timestamps, and cryptographic hashes of predefined critical evidence items [21]. The study employs a pilot comparison between a baseline logging procedure and a blockchain-supported tamper-evident ledger (TEL) workflow using de-identified examination-session data. The study focuses on governance-related measures: i) examination integrity, operationalized through auditability, evidence completeness, and hash-based verification; ii) fairness of assessment decisions, measured through false positives, reversals, and audit efficiency; and iii) trust and acceptance, assessed through perceived procedural fairness, transparency, trust in assessment, and intention to use.

Unlike previous studies, this research provides the first controlled comparative evaluation of a blockchain-supported evidence workflow from an educational governance perspective. Specifically, the research evaluates how cryptographically verifiable evidence chains influence auditability, audit efficiency, reviewer consistency, and learner perceptions in remote assessment settings. The fundamental scientific question addressed in this study is whether the introduction of a decentralized, tamper-evident evidence layer can bridge the transparency gap inherent in automated proctoring systems and restore the perceived legitimacy of online assessment governance. To investigate this, the following research questions (RQs) were formulated:

- How does the integration of a blockchain-supported evidence chain impact the verifiability and auditability of assessment sessions compared to traditional logging methods? (RQ1)
- To what extent can verifiable evidence workflows improve institutional governance efficiency by mitigating fairness risks and streamlining the audit process? (RQ2)
- Do perceived procedural fairness and transparency predict trust in online assessment, and does trust predict acceptance and intention to use the procedure? (RQ3)

2. METHOD

2.1. Research design and setting

In this study, a two-condition experiment was employed to determine the influence of an evidence governance layer on the quality, efficiency, and legitimacy of remote assessment review. This experiment was conducted on the online assessment system BlockchainStudy.kz within the context of an exam administered in a test course. The exam did not affect students' grades or progress and had no bearing on any academic decisions. In the control condition, the evidence processing mechanism provided by the platform was applied, while in the experimental condition, the blockchain-based evidence workflow was employed. The logic of automated flagging and reviewer decision guidelines remained unchanged [24].

Sessions were allocated through simple randomization, with the software using a 1:1 allocation ratio. The reviewers used the same criteria for violations and documentation in both groups. Blinding was not possible because the interface and integrity markers differed between the two conditions. To reduce procedural bias, the reviewers applied the same violation criteria, audit rubric, decision thresholds, and evidence-review procedures across both groups. In addition, the automated flagging logic remained unchanged between conditions, allowing differences in outcomes to be attributed primarily to the evidence-governance workflow rather than to changes in detection mechanisms.

Figure 1 presents the overall research design, including participant allocation, workflow structure, and the comparison between the control and experimental evidence-governance conditions. This schematic diagram represents the assessment process lifecycle, depicting the progressive shift from the automatic flagging process stage to the final human audit process stage. By identifying the specific points at which blockchain integrity measures are incorporated, the difference between ordinary monitoring and governance is made clear.

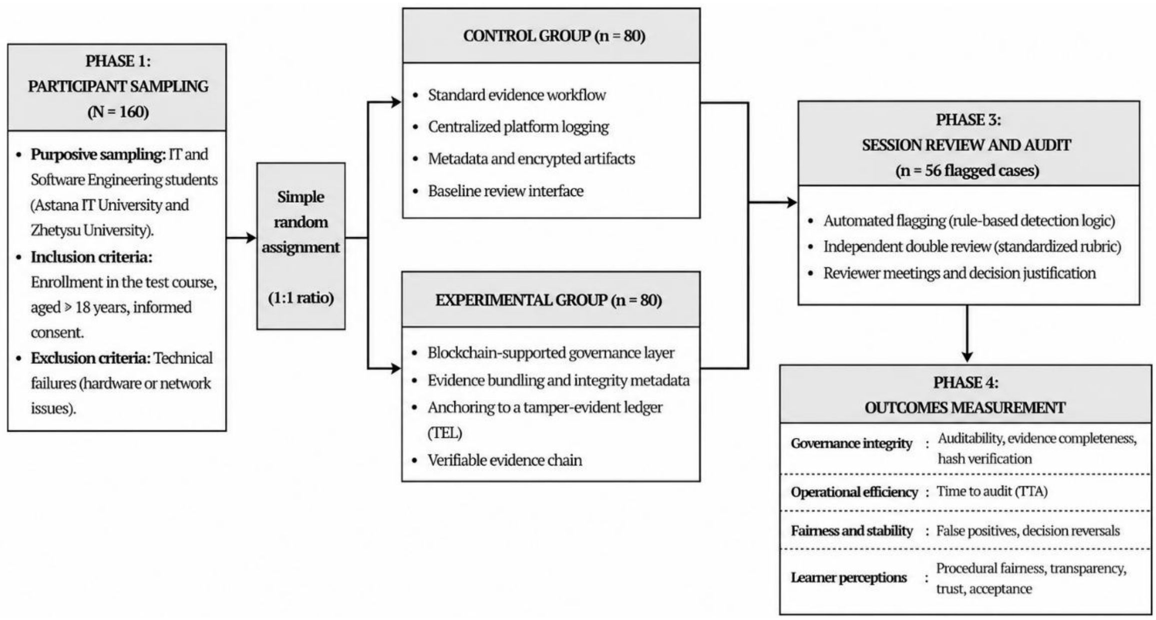

Figure 1. Research design schematic and comparative evidence governance workflows

2.2. Data sources, participants, and procedure

Participants were recruited among the students enrolled in Information Technology and Software Engineering courses offered at Astana IT University and I. Zhansugurov Zhetysu University. Participants were sampled using a purposive sampling strategy, where the intention was to recruit individuals with adequate digital literacy capabilities. Random allocation to the conditions occurred at the session level. Participants were eligible to participate in the study if: i) they were enrolled in the identified course; ii) they were 18 years old or older; and iii) they provided informed consent. Sessions that had hardware problems or failed due to network disruptions attributable to internet service providers (ISPs) were excluded from the analysis. The sample size (n=160 sessions) was determined to be sufficient for detecting medium effect sizes (d=0.5) with a statistical power of 0.80 under a two-group comparative design. This pilot-scale sample was considered appropriate for evaluating governance-related operational and perceptual outcomes under controlled experimental conditions.

Two datasets were collected. First, a dataset comprising 160 examination sessions was created, where 80 sessions were assigned to the control condition, while the other 80 sessions were allocated to the experimental condition. This dataset contained information regarding each examination session. In cases where suspicious activity was recorded based on the platform rules, an encrypted screen capture was taken. Session-level information was extracted after each session [25].

The second dataset consisted of post-test surveys gathered from 138 subjects (68 from the control group (CG) and 70 from the experimental group (EG)). Participation in the study was completely voluntary and abided by the institution's rules regarding the piloting of new initiatives. Confidentiality was maintained by ensuring that all collected information was anonymized. Upon completing the test, all subjects filled out a questionnaire. Sessions that qualified under the automated rules were audited according to the same standards across both groups.

2.3. Evidence workflows and audit procedure

In the control condition, reviewers accessed the platform's standard evidence export and interface. In the experimental condition, evidence was organized into a structured bundle aligned with a predefined critical evidence set. The workflow generated integrity metadata for each evidence bundle and anchored this metadata to an append-only ledger [21], [22]. Integrity metadata consisted of SHA-256 hashes, timestamps, and a pseudonymous session identifier. Only integrity metadata were anchored, and no personally identifying information and no content artifacts were stored on-chain. From the educational governance standpoint, this procedure ensures that the organization implements a process of 'verifiable evidence chain.' Cryptographic linking of the students' performance data creates a tamper-proof record. This makes it possible for the organization to conduct any future audit or appeal using a set of records that will be absolutely identical to the originals. All evidence artifacts, including encrypted screenshots triggered by suspicious events, were kept off-chain, and access was provided only to authorized auditors and researchers. During the integrity checking process, a hash verification protocol was followed where the hash value of the evidence package was recalculated during the audit and checked against the anchored hash. If the recalculated hash and the anchored hash matched, the session was marked as a hash verification pass. Once the evidence packages were extracted, the sessions were audited according to the automated rules set by the platform.

When reviewers differed on whether to flag a session based on the flagged-session decisions, the reviewer would have a meeting and come to an agreement about how to decide based on the rubric that was used as the basis for making decisions on the audit cases. In addition to coming to an agreement, reviewers would document the reasoning behind their decision in a structured way so that the decision-making process could be traced by documenting what evidence was used to make each decision and why each decision was made. To further add structure to the decision-making process and to provide evidence of when each decision was made, reviewers were required to point to the specific evidence reviewed and include a date and time stamp with their final decision in the audit workflow. Although these processes remained consistent between conditions, the ways in which the data was packaged and whether integrity markers existed during the experimental workflow varied. These similarities allow researchers to attribute potential differences in findings to the governance features that were different (i.e., evidence packaging and integrity markers), and not to differences in the detection criteria or the review criteria.

2.4. Measures and qualitative coding

The results obtained at the session level were defined beforehand. Flagging involved checking whether a session was flagged in the rule-based system. Auditability was determined by whether the required evidence set for auditing was present. Sessions where all the necessary information was available for review and sufficient for reconstruction were considered auditable. Evidence completeness indicated whether all required data fields for the evidence were collected during the export process. Time to audit was calculated as the time taken between the beginning and end of auditing, in minutes. Hash verification pass was noted for experimental sessions where the recomputed hash matched the original. Outcomes related to fairness and stability were assessed for the flagged sessions with respect to the audit results. False positives referred to sessions that were flagged but not found to be violations after auditing. Reversal after audit referred to the overturning of initial judgments after auditing. Agreement among raters on flagged sessions was assessed using Cohen's kappa coefficient with a 95% confidence interval (CI) [26].

The five constructs used to measure governance and acceptance in the survey were acceptance, procedural justice, trust, transparency, and privacy concerns. These constructs were rated using items on a Likert scale of 1–5, and the scale scores were averaged. The survey instrument was developed based on governance and procedural-justice constructs commonly used in prior educational technology and assessment research. Content validity was established through alignment between questionnaire items, the study constructs, and the research questions. Internal consistency reliability was evaluated using Cronbach's alpha coefficients. Qualitative data included structured rationale statements generated during the review process for all flagged sessions. These statements were coded using a directed coding approach consistent with the governance rubric and the structured evidence package.

The coding scheme included consistency of the timeline sequence, correspondence in logs, references to evidence, timestamped statements, and integrity criteria such as hashing or ledger checks. In addition, linkage between cited evidence and the final decision was explicitly coded. Coding was conducted to understand how decisions were justified and how documentation varied between conditions. Qualitative data were used to interpret quantitative outcomes on a mechanistic basis. All coding categories were defined prior to data collection based on the governance constructs measured quantitatively.

2.5. Statistical analysis and ethics

Binary session results were analyzed for differences between groups via Pearson chi-square tests. When comparing flagged subsets where cell counts were expected to be low, Fisher's exact tests were employed along with odds ratios with 95% CI. The time-to-audit (TTA) variable was examined with the Mann-Whitney U test since its values are skewed, a natural characteristic of such duration variables; the rank-biserial correlation statistic was reported as the measure of effect size. Survey differences between groups were compared by independent samples t-tests using Cohen's d. Two regression analyses were conducted. The first model examined the effects of procedural fairness, transparency, and privacy concern on trust, while the second model examined the effects of trust, procedural fairness, and transparency on acceptance.

Standardized coefficients and model R-squared values were reported, and multicollinearity was assessed using variance inflation factors. The study was conducted in a low-stakes test-course exam context, and all participants provided informed consent, participation was voluntary, and withdrawal was possible at any time without consequences. Session logs and metadata contained no direct identifiers, and encrypted screenshots were stored off-chain with restricted access and were used only for audit purposes. Only non-identifying integrity metadata were anchored on-chain, and results are reported in aggregate. These safeguards were implemented to minimize privacy risk while enabling verifiable evidence handling and audit transparency. Since the information obtained through the research was de-identified and conducted under low-stakes conditions where grades were not affected, the study was classified as an internal educational assessment, and formal institutional ethics approval was not required under applicable guidelines.

3. RESULTS

3.1. Quantitative results

A total of 160 mock-exam sessions were analyzed (CG=80, EG=80). The automated workflow flagged 29 sessions in CG (36.3%) and 27 sessions in EG (33.8%), with no statistically significant difference in flagging rates ($\chi^2=0.027$, $p=0.868$). In the experimental workflow, ledger anchoring was applied to all sessions and hash-based verification succeeded in 79/80 sessions (98.8%), with one mismatch recorded. These initial session-level integrity and efficiency outcomes, which serve as the foundation for subsequent auditability analysis, are summarized in Table 1.

Auditability (a full, reconstructable evidence set) was 83.8% in CG and 96.3% in the EG, showing a statistically significant difference between groups ($\chi^2=6.944$, $p=0.008$; relative risk (RR)=1.15; Cramér's V=0.208). The proportion of evidence that was complete (had all fields as defined in the study protocol) was 86.3% in the CG and 93.8% in the EG and there were no group differences in this measure ($\chi^2=2.500$, $p=0.11$; RR=1.09; V=0.125). The findings underscore the high level of reliability of the blockchain-enabled system in preserving a reconstructive set of evidence. Figure 2 presents the comparative auditability and evidence-completeness outcomes between the CG and EG. This visualization confirms that the EG achieved superior results in maintaining a verifiable evidence chain compared to the baseline logging method.

The time to audit (in minutes) was defined as an operational measure for governance efficiency. Time to audit was evaluated at the session level (n=160), with a median time to audit being 18.6 minutes for the CG and 11.7 minutes for the EG (means=19.40 min vs 12.43 min). An analysis using the Mann-Whitney U-test indicated that the average audit time reduced significantly when using the experimental workflow ($p<0.001$; $U=5849$; $r(rb)=-0.828$). Figure 3 illustrates the distribution of audit duration across both groups, showing consistently shorter review times in the experimental condition.

Table 1. Session-level integrity and efficiency outcomes (N=160)

Outcome	CG (n=80)	EG (n=80)	Statistic	p	Effect
Flagged sessions	29 (36.3%)	27 (33.8%)	$\chi^2=0.027$	0.868	V=0.026
Auditability	67 (83.8%)	77 (96.3%)	$\chi^2=6.944$	0.008	RR=1.15; V=0.208
Evidence completeness	69 (86.3%)	75 (93.8%)	$\chi^2=2.500$	0.114	RR=1.09; V=0.125
Hash verification pass (EG only)	—	79 (98.8%)	—	—	—
TTA (min)	Median 18.6; mean 19.40	Median 11.7; mean 12.43	U=5849	<0.001	r(rb)=−0.828

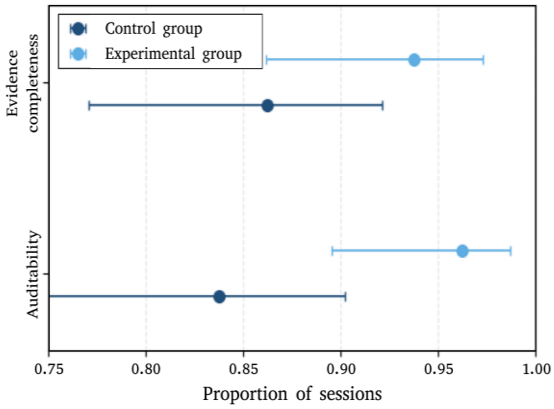

Figure 2. Auditability and evidence completeness by group

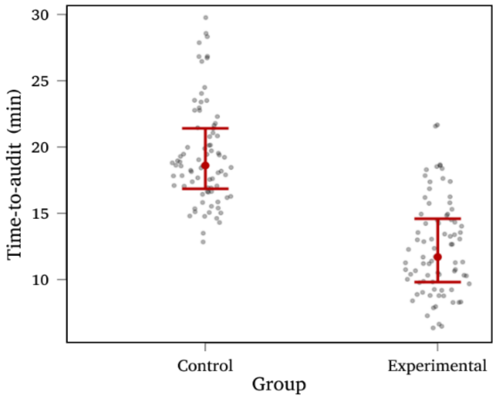

Figure 3. Distribution of TTA by group

The evaluation of fairness indicators involved examining the flagged subgroup (n=56; CG=29, EG=27), where a third-party reviewer played the role of ground truth to determine whether a flag had been correctly applied. In the CG, the false positive rate was 24.1%, whereas in the EG, it was 14.8%. Fisher's exact test failed to detect any statistically significant differences between the groups (p-value=0.506, odds ratio=1.83, 95% CI [0.47, 7.13]). Additionally, reversals of fairness decisions after auditing happened in an equal number of sessions for both groups (CG=3.45% vs. EG=3.70%), as confirmed by Fisher's exact test (p=1.000). Furthermore, inter-rater reliability was very high when making decisions regarding flagged sessions (Cohen's κ=0.846, 95% CI [0.679, 1.000]). Ultimately, these stable metrics confirm that the reviewers evaluated both groups with a high degree of consistency and objectivity.

Of utmost importance is that these findings show that the improvement in auditability of the experimental procedure did not lead to any compromises in the reliability of the evaluation process. Fairness of outcomes was similar for both groups; however, there was a considerable range in the odds ratios obtained in the analysis, suggesting that the sample size used did not afford sufficient precision, and that small-to-moderate discrepancies in the frequency of false-positive ratings cannot be excluded. Overall, this analysis suggests that the use of the experimental procedure did not create any observable instabilities in the reviewers' decision-making. Detailed fairness and stability analysis results can be seen in Table 2.

Table 2. Fairness and decision stability (flagged subset, n=56)

Metric	CG (n=29)	EG (n=27)	Test	p	Effect
False positives	7 (24.1%)	4 (14.8%)	Fisher	0.506	OR=1.83; 95% CI [0.47, 7.13]
Reversal after audit	1 (3.45%)	1 (3.70%)	Fisher	1.000	—
Inter-rater agreement	—	—	κ=0.846	—	95% CI [0.679, 1.000]

Survey responses were obtained from 138 participants (CG=68, EG=70). Internal consistency was acceptable to excellent across constructs (Cronbach's α range 0.776–0.901), supporting the reliability of the scale scores used in subsequent analyses. The group comparisons are summarized in Table 3, which reports descriptive statistics and independent-samples t-tests for procedural fairness, transparency, trust, acceptance, and privacy concern.

As shown in Table 3, participants in EG reported significantly higher procedural fairness (t=−4.847, p<0.001), transparency (t=−3.111, p=0.002), trust (t=−4.963, p<0.001), and acceptance (t=−4.505, p<0.001) compared with CG. The magnitudes of the mean differences were practically meaningful: the EG means exceeded CG by 0.64–1.01 points on a 1–5 scale across these constructs, indicating a broad shift in perceived governance quality rather than an isolated improvement in a single dimension. In contrast, privacy concern did not differ significantly between conditions (t=1.391, p=0.167), suggesting that increasing evidence traceability and audit structure did not increase perceived privacy risk in this low-stakes setting. To examine the proposed governance pathway, relationships among the perception constructs were evaluated, and regression models were estimated linking fairness and transparency to trust, and trust to acceptance. Regression results are reported in Table 4, which presents standardized coefficients and model fit indices for both models. Regression models were estimated on the survey sample (N=138) and are reported with standardized coefficients.

Table 3. Reliability and group differences in education-relevant perceptions (Likert scale 1–5)

Construct	α	CG (n=68) M±SD	EG (n=70) M±SD	t	p
Procedural fairness	0.852	2.82±0.97	3.66±1.07	−4.847	<0.001
Transparency	0.836	2.80±1.20	3.44±1.21	−3.111	0.002
Trust	0.885	2.67±1.20	3.68±1.20	−4.963	<0.001
Acceptance	0.901	2.53±1.43	3.58±1.31	−4.505	<0.001
Privacy concern	0.776	3.01±1.02	2.77±1.04	1.391	0.167

Table 4. Regression models (standardized β)

Model DV	Predictor	β	p	R²
Trust	Procedural fairness	0.413	<0.001	0.362
	Transparency	0.357	<0.001	
	Privacy concern	0.062	0.381	
Acceptance	Trust	0.562	<0.001	0.436
	Procedural fairness	0.143	0.058	
	Transparency	0.035	0.637	

3.2. Qualitative results

Qualitative data consisted of structured reviewer rationale notes recorded during audits of all flagged sessions (n=56; CG=29, EG=27). Notes were coded using a pre-specified scheme aligned with the governance rubric and evidence package structure, covering: timeline coherence, log alignment, explicit evidence traceability, timestamp anchoring, integrity-marker references (hash/ledger), and explicit linkage between evidence and final decision. Coding outcomes are summarized in Table 5 and provide additional qualitative insight into how reviewers justified audit decisions, referenced evidence artifacts, and documented procedural reasoning across the two governance conditions.

Table 5. Coded reviewer rationale features by group (flagged subset, n=56)

Feature in reviewer notes	CG (n=29)	EG (n=27)
Timeline coherence referenced	26	26
Log alignment referenced	24	25
Explicit evidence traceability	16	25
Timestamp anchoring referenced	0	23
Hash/ledger integrity referenced	0	21
Explicit evidence-to-decision linkage	16	26

Across both groups, reviewers consistently relied on reconstructing a chronological narrative of learner activity and cross-checking that narrative against system logs, indicating that timeline coherence and log alignment were foundational decision cues. However, notes from EG were more frequently expressed in explicitly evidentiary terms (e.g., referring to traceable artifacts or anchored timestamps), reflecting the availability of structured evidence packaging in the experimental workflow. This pattern is consistent with the higher auditability observed in EG and offers a plausible qualitative explanation for the efficiency gains.

Rationales in CG, tended to describe probable misconduct at a general descriptive level (i.e., unusual behaviors; inconsistencies) whereas, in EG rationales more often provided explicit references to particular pieces of evidence and a more direct progression from evidence to interpretation to final determination. As is consistent with the documentation pattern used within the EG, there is a greater frequency of reference to the evidence cited, and the determination made based upon that evidence, in the EG; thus, suggesting that the workflow process used in the EG supported both a more rapid review process as well as more transparent justification processes for the reviewer's decisions. The reviewer comments also indicated a different qualitative approach to addressing ambiguity. The ambiguity in cases reviewed in CG was more frequently resolved through an additional manual search of available resources, and/or a back-and-forth process with respect to all available resources. Ambiguity in cases reviewed in the EG, was more frequently resolved through the use of integrity markers and the structured evidence package, which facilitated expedited confirmation of whether the necessary evidence artifacts existed, were consistent, and complete. These findings alone do not provide causality, however, when combined with the quantitative results, they support the conclusion that the primary mechanism for achieving governance enhancements through the EG was enhanced evidence organization and verifiability, rather than any change in the detection threshold.

4. DISCUSSION

Unlike other studies that have primarily addressed the "detection accuracy" aspect of the remote proctoring debate, this study addresses a broader scientific question: whether a verifiable governance layer can bridge the transparency gap and restore institutional legitimacy. The findings confirm that legitimacy is created not by more surveillance, but through auditable processes and defensible decision-making. By strengthening the evidence pipeline between automated flagging and evaluation, the intervention successfully creates a defensible evidence base without altering the detection threshold. These results also resonate with the United Nations' SDG, especially SDG 4 on quality education and SDG 16 on peace, justice, and strong institutions, through the promotion of ethical and transparent assessment practices. Through greater auditability and transparency, the suggested system will contribute to increasing institutional legitimacy and conducting unbiased assessments.

RQ1 examined whether blockchain-supported assessment increases auditability and enables tamper-evident verification compared with baseline logging. The results were affirmative and analytically important for assessment governance. First, improved auditability indicates that sessions in the EG more often met reviewability standards, meaning an external party could reconstruct the full evidence set without relying solely on platform trust. This is important because critiques of remote proctoring often highlight that legitimacy breaks down not at the moment of flagging, but during disputes, appeals, and escalations, when decisions must be clearly explained and defended. Second, the near-universal hash verification success rate in the EG demonstrates the practical value of a TEL: the integrity of evidence bundles can be verified after the fact using cryptographic methods, without storing sensitive content on-chain. Together, improved auditability and TEL functionality reduce reliance on the platform as the single source of truth by adding an independent, verifiable integrity layer to the audit process.

RQ2 examined whether governance efficiency improves without increasing fairness risks. The strongest finding was improved efficiency. The substantial reduction in TTA shows that the evidence workflow reduced reviewer effort required for evidence retrieval, reconciliation, and resolving uncertainty. In governance terms, the intervention appears to reduce procedural workload during the audit phase. Importantly, this reduction reflects a consistent shift across cases rather than improvement in only a few extreme instances, suggesting that the workflow improved routine reviewer processes rather than simply accelerating easier reviews. Fairness outcomes within the flagged subset did not show statistically significant differences [27]. However, the wide CI indicates limited precision, meaning the correct interpretation is not that fairness is identical, but that no fairness degradation was detected at the current sample size, and small-to-moderate effects remain possible. This aligns with assessment governance research, where absence of evidence does not confirm absence of risk, especially when false positive rates are evaluated only within flagged cases [28]. From a risk-management perspective, stable flagging rates, strong reviewer agreement, and no increase in false positives suggest that improving evidence governance can enhance efficiency without destabilizing decision outcomes. This efficiency gain is important because institutions can conduct scalable audit oversight without proportionally increasing administrative workload.

RQ3 examined whether procedural fairness and transparency predict trust, and whether trust predicts acceptance and intention to use. Results support a pathway in which fairness and transparency in the EG strengthen trust in the assessment process, which in turn increases acceptance. Regression findings show that fairness and transparency predict trust, and trust predicts acceptance, while fairness and transparency lose direct predictive power once trust is included, indicating that trust mediates acceptance. This is consistent with prior research showing that perceived legitimacy and procedural quality drive acceptance of educational technologies. The absence of significant differences in privacy concern suggests that governance improvements based on verifiable evidence and data minimization can increase legitimacy without increasing privacy risks [29].

Qualitative findings clarify the mechanism. EG reviewers more often referenced specific evidence artifacts, integrity markers, and explicit evidence-to-decision linkage, while CG rationales were more descriptive and required additional manual verification. This indicates that structured evidence packaging and TEL cues reduce cognitive effort, improve verification, and support clearer justification of decisions. Tamper-evident features therefore function not only as security controls but also as governance tools that improve traceability and consistency.

The broader implication of these findings is that remote proctoring systems should be treated as governance systems rather than surveillance tools. These findings reinforce that legitimacy in online assessment depends on verifiable, auditable decision processes rather than increased monitoring. Providing a verifiable evidence chain reduces asymmetry between institution and student and addresses the "black box" problem in assessment governance. From a theoretical perspective, the study operationalizes governance as measurable system properties: auditability, traceability, and verifiability. Extending prior work on remote proctoring and student experience [6], [30], the findings demonstrate that improvements in procedural fairness and transparency can be achieved through evidence design rather than increased control. The regression results further clarify that trust mediates the relationship between governance quality and acceptance. From a practical standpoint, the experimental workflow reduced TTA by 37% while maintaining stable fairness indicators and not increasing privacy concern. This indicates that structured evidence and integrity markers improve efficiency without introducing additional risks and without increasing administrative overhead. In line with prior research by Maphalala and Nkosi [31], the findings suggest that institutions can scale online assessment while improving legitimacy, without intensifying surveillance.

The study contributes in three concrete ways. First, it operationalizes governance outcomes that are often discussed but rarely measured in a single design, including auditability, evidence completeness, tamper-evident verification, false positive risk, reversal stability, reviewer agreement, TTA, and governance perceptions. Second, it provides a controlled pilot template that isolates evidence governance from detection changes by holding the flagging logic and rubric constant. Third, it offers a design implication for practice: institutions can treat evidence packaging, integrity anchoring, and traceable decision documentation as a governance subsystem of remote proctoring, rather than treating remote proctoring purely as a surveillance subsystem.

Several limitations should be explicitly acknowledged. First, the study was conducted in a low-stakes assessment context, which may limit the generalizability of the findings to high-stakes environments where incentives for academic misconduct are stronger and audit cases may be more complex. Second, the sample consisted of students from IT and software engineering programs, who are likely to possess higher levels of digital literacy; this may have influenced both their interaction with the system and their perception of fairness and transparency. Third, reviewers were not blinded to experimental conditions due to interface differences, which introduces the possibility of procedural bias in audit behavior. Fourth, fairness outcomes were evaluated only within the flagged subset, and the relatively small number of cases resulted in wide CI, limiting the precision of these estimates. Finally, the study employed a pilot-scale sample, which constrains the statistical power for detecting smaller effects, particularly in fairness-related measures.

Future research should address these limitations by replicating the study in high-stakes assessment settings and across more diverse student populations, including disciplines with lower levels of technical proficiency. Larger-scale studies are needed to improve the precision of fairness estimates and to examine potential effects on false positives and appeal outcomes. Longitudinal designs would be particularly valuable for assessing whether verifiable evidence workflows reduce institutional dispute rates and sustain efficiency gains over time. In addition, future work should explore how governance-oriented system design can be integrated with existing proctoring infrastructures and how different configurations of evidence granularity and transparency affect user trust and institutional accountability.

5. CONCLUSION

Based on the findings of the current research, the implementation of a blockchain-based evidence management system in online tests significantly enhances assessment governance. From the results generated in response to the research questions, the study concludes that structured governance systems can address the lack of transparency in conventional monitoring approaches. In answering RQ1 and RQ2, the introduction of this solution led to improvements in session auditability by 12.5% (to 96.3%) and a reduction in the duration required for audits by 37%. These results indicate that a process-driven, evidence-governed, and cryptographically validated auditing system can enhance governance efficiency by reducing the cognitive load on auditors without necessitating additional monitoring measures. Regarding RQ3, the findings demonstrate a clear link between verifiability and institutional legitimacy. Participants in the EG reported better experiences with procedural justice and fairness. Thus, an accountable digital education solution is expected to reduce student distrust toward online testing. The significance of these results lies in the demonstration that transparency in proctoring is not a technical impossibility, but a policy choice.

It is imperative to identify several limitations that should be taken into consideration, as they help to set the limits of this research's applicability. First, the study was conducted in the context of low-stakes testing, in which motivations for academic fraud are less pronounced compared to high-stakes scenarios. Thus, the 37% reduction in audit time achieved through the experimental software might not sufficiently address more complex cases in other contexts. Second, the participants come from IT and software engineering backgrounds and thus have higher digital literacy, which allowed them to adopt the proposed solution more quickly, potentially limiting the generalizability to groups with lower technical proficiency.

Future research should focus on replicating this paradigm in high-stakes settings and among diverse student populations to confirm its cross-disciplinary viability. Additionally, longitudinal studies are required to investigate the long-term impact of verifiable audit trails on institutional appeal rates and the sustainability of administrative workloads. Overall, this study offers an innovative, cryptographically defensible model for educational institutions seeking to evolve from passive proctoring systems to active, accountable assessment governance mechanisms.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Aivar Sakhipov	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓
Laura Smagulova	✓	✓		✓		✓	✓	✓	✓	✓		✓	✓	
Aigul Yelepbergenova		✓		✓		✓	✓	✓		✓	✓			
Alexey Fedenko	✓	✓	✓	✓	✓				✓	✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

The study was conducted in accordance with applicable institutional policies and relevant national regulations. The research was carried out in a low-stakes test-course context using de-identified platform data. Given that the study involved de-identified data and was conducted in a low-stakes context without impact on academic outcomes, formal institutional ethics approval was not required under applicable guidelines. All participants provided informed consent, participation was voluntary, and withdrawal was possible at any time. Data were processed and reported in aggregate form with restricted access to audit artifacts.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [LS], upon reasonable request. The dataset is not publicly available because it includes de-identified platform logs and encrypted audit artifacts that may still carry a risk of re-identification and could reveal system security-relevant details if disclosed.

REFERENCES

[1] K. Butler-Henderson and J. Crawford, "A systematic review of online examinations: a pedagogical innovation for scalable authentication and integrity," *Computers and Education*, vol. 159, p. 104024, Dec. 2020, doi: 10.1016/j.compedu.2020.104024.
[2] A. Nigam, R. Pasricha, T. Singh, and P. Churi, "A systematic review on AI-based proctoring systems: past, present and future," *Education and Information Technologies*, vol. 26, no. 5, pp. 6421–6445, Sep. 2021, doi: 10.1007/s10639-021-10597-x.
[3] S. Han, S. Nikou, and W. Y. Ayele, "Digital proctoring in higher education: a systematic literature review," *International Journal of Educational Management*, vol. 38, no. 1, pp. 265–285, Jan. 2024, doi: 10.1108/IJEM-12-2022-0522.
[4] M. Główczewski and S. Burdziej, "(In)justice in academia: procedural fairness, students' academic identification, and perceived legitimacy of university authorities," *Higher Education*, vol. 86, no. 1, pp. 163–184, Jul. 2023, doi: 10.1007/s10734-022-00907-8.
[5] M. S. Wanasinghe-Pasqual, "Achieving sustainable development goals 4 and 16 in former war-torn countries through peace education," *Colombo Journal of Multi-Disciplinary Research*, vol. 5, no. 1–2, pp. 121–148, Dec. 2020, doi: 10.4038/cjmr.v5i1-2.57.
[6] E. Marano, P. M. Newton, Z. Birch, M. Croombs, C. Gilbert, and M. J. Draper, "What is the student experience of remote proctoring? A pragmatic scoping review," *Higher Education Quarterly*, vol. 78, no. 3, pp. 1031–1047, Jul. 2024, doi: 10.1111/hequ.12506.
[7] A. Terpstra, A. de Rooij, and A. Schouten, "Online proctoring: privacy invasion or study alleviation?" in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Apr. 2023, pp. 1–20, doi: 10.1145/3544548.3581181.

[8] C. Mutimukwe, O. Viberg, C. McGrath, and T. Cerratto-Pargman, “Privacy in online proctoring systems in higher education: stakeholders’ perceptions, awareness and responsibility,” *Journal of Computing in Higher Education*, vol. 38, no. 2, pp. 732–761, Jun. 2026, doi: 10.1007/s12528-025-09461-5.

[9] Y. Cherif, C. Sayffaerth, F. Chiossi, and L. Schütz, “Stress by design? The influence of online exam interfaces on student anxiety,” in *Proceedings of the Mensch und Computer 2025*, Aug. 2025, pp. 705–710, doi: 10.1145/3743049.3748538.

[10] M. E. Eltahir *et al.*, “Students’ experiences of fairness in online assessment: a phenomenological study in a higher education institution context,” *SAGE Open*, vol. 13, no. 4, pp. 1–13, Oct. 2023, doi: 10.1177/21582440231209816.

[11] D. Yao, “Examining the subjective fairness of at-home and online tests: taking Duolingo English test as an example,” *PLoS ONE*, vol. 18, no. 9, p. e0291629, Sep. 2023, doi: 10.1371/journal.pone.0291629.

[12] J. Wanner, L.-V. Herm, K. Heinrich, and C. Janiesch, “The effect of transparency and trust on intelligent system acceptance: evidence from a user-based study,” *Electronic Markets*, vol. 32, no. 4, pp. 2079–2102, Dec. 2022, doi: 10.1007/s12525-022-00593-5.

[13] J. A. Colquitt, “On the dimensionality of organizational justice: a construct validation of a measure.,” *Journal of Applied Psychology*, vol. 86, no. 3, pp. 386–400, Jun. 2001, doi: 10.1037/0021-9010.86.3.386.

[14] A. Mitev, L. Koloszár, and Á. Wimmer, “A model of constructive procedural justice through the example of a student academic competition,” *The International Journal of Management Education*, vol. 23, no. 2, p. 101176, Jul. 2025, doi: 10.1016/j.ijme.2025.101176.

[15] D. Playfoot, R. Horry, and A. E. Pink, “Fairly useful feedback: characteristics of feedback comments perceived as fair by students,” *Assessment and Evaluation in Higher Education*, vol. 51, no. 4, pp. 665–682, May 2026, doi: 10.1080/02602938.2025.2586836.

[16] A. Al Fraidan, “Procedural transparency and legal accountability to sustain AI-mediated language assessment in Saudi Arabia,” *SAGE Open*, vol. 15, no. 4, pp. 1–13, Oct. 2025, doi: 10.1177/21582440251396113.

[17] T. Fawns and S. Schaepkens, “A matter of trust: online proctored exams and the integration of technologies of assessment in medical education,” *Teaching and Learning in Medicine*, vol. 34, no. 4, pp. 444–453, Aug. 2022, doi: 10.1080/10401334.2022.2048832.

[18] D. R. Isbell, B. Kremmel, and J. Kim, “Remote proctoring in language testing: implications for fairness and justice,” *Language Assessment Quarterly*, vol. 20, no. 4–5, pp. 469–487, Oct. 2023, doi: 10.1080/15434303.2023.2288251.

[19] Bhawna, P. Gupta, and P. Rai, “Can blockchain revolutionize educational practices? An in-depth analysis of applications and challenges,” *Sustainable Futures*, vol. 10, p. 101171, Dec. 2025, doi: 10.1016/j.sftr.2025.101171.

[20] C. Regueiro and B. Urquizu, “Blockchain-based evidence trustworthiness system in certification,” *Journal of Cybersecurity and Privacy*, vol. 5, no. 1, p. 1, Dec. 2024, doi: 10.3390/jcp5010001.

[21] A. Ahmad, M. Saad, M. Al Ghamdi, D. Nyang, and D. Mohaisen, “BlockTrail: a service for secure and transparent blockchain-driven audit trails,” *IEEE Systems Journal*, vol. 16, no. 1, pp. 1367–1378, Mar. 2022, doi: 10.1109/JSYST.2021.3097744.

[22] C. Regueiro, I. Seco, I. Gutiérrez-Agüero, B. Urquizu, and J. Mansell, “A blockchain-based audit trail mechanism: design and implementation,” *Algorithms*, vol. 14, no. 12, p. 341, Nov. 2021, doi: 10.3390/a14120341.

[23] E. Heinrich, “A systematic-narrative review of online proctoring systems and a case for open standards,” *Open Praxis*, vol. 17, no. 3, pp. 485–499, Aug. 2025, doi: 10.55982/openpraxis.17.3.836.

[24] A. Tweissi, W. Al Etaiwi, and D. Al Eisawi, “The accuracy of AI-based automatic proctoring in online exams,” *Electronic Journal of e-Learning*, vol. 20, no. 4, pp. 419–435, Oct. 2022, doi: 10.34190/ejel.20.4.2600.

[25] S. Mukherjee, V. Distler, G. Lenzini, and P. Cardoso-Leite, “Balancing the perception of cheating detection, privacy and fairness: a mixed-methods study of visual data obfuscation in remote proctoring,” in *Proceedings of the 2024 European Symposium on Usable Security*, Sep. 2024, pp. 337–353, doi: 10.1145/3688459.3688474.

[26] D. Woldeab and T. Brothen, “Exploring the impact: how online exam proctoring reduces cheating and enhances course legitimacy,” *International Journal of E-Learning and Distance Education/Revue internationale du e-learning et la formation à distance*, vol. 39, no. 1, p. n1, Aug. 2024, doi: 10.55667/ijede.2024.v39.i1.1327.

[27] A. Sakhipov, I. Omirzak, and A. Fedenko, “Beyond face recognition: a multi-layered approach to academic integrity in online exams,” *Electronic Journal of e-Learning*, vol. 23, no. 1, pp. 81–95, Feb. 2025, doi: 10.34190/ejel.23.1.3896.

[28] H. Choung, P. David, H. Mahmud, and S. Norcutt, “Fairness and trust in AI decision-making: the role of human involvement and outcome favorability,” *International Journal of Human–Computer Interaction*, vol. 42, no. 12, pp. 9350–9370, Jun. 2026, doi: 10.1080/10447318.2025.2576634.

[29] M. Henderson *et al.*, “Online examinations: factors that impact student experience and perceptions of academic performance,” *Australasian Journal of Educational Technology*, vol. 40, no. 4, pp. 73–89, Oct. 2024, doi: 10.14742/ajet.9412.

[30] J. Oeding, T. Gunn, and J. Seitz, “The mixed-bag impact of online proctoring software in undergraduate courses,” *Open Praxis*, vol. 16, no. 1, pp. 82–93, Feb. 2024, doi: 10.55982/openpraxis.16.1.585.

[31] M. C. Maphalala and N. Nkosi, “Proctoring online assessments: enhancing security and academic integrity in open distance eLearning,” *Open Praxis*, vol. 17, no. 4, pp. 680–690, Nov. 2025, doi: 10.55982/openpraxis.17.4.911.

BIOGRAPHIES OF AUTHORS

Aivar Sakhipov is an assistant professor at Astana IT University, Ph.D. in Computer Sciences. His academic and professional career combines teaching, research, and the development of innovative digital solutions for education. His professional and research interests include: implementation of blockchain, artificial intelligence, and data-driven technologies in education; development of digital platforms for certification, distance learning, and academic integrity; application of machine learning, databases, and data analysis in educational and real-world systems. He can be contacted at email: aivar.sakhipov@astanait.edu.kz.

Laura Smagulova is a lecturer at I. Zhansugurov Zhetysu University, candidate of Pedagogical Sciences. She is an experienced and highly qualified specialist with over 32 years of research and teaching experience in higher education. Her professional and research interests include: implementation of innovative technologies and digital resources in the educational process; main directions and challenges in the professional preparation of information technology specialists. Her research and publications focus on key areas, including the application of digital technologies in the educational process, as well as the improvement of teaching methodologies for programming languages. She can be contacted at email: smagulovalaura272@gmail.com.

Aigul Yelepbergenova is a lecturer at the Department of Information Technologies and Artificial Intelligence of I. Zhansugurov Zhetysu University. In the educational process, priority is given to the effective use of modern digital technologies, programming languages, and elements of artificial intelligence and machine learning. Research activities cover mobile application design, artificial intelligence, data analysis, digitalization of the educational process, and the development of intelligent systems. Publications have appeared in both international and national scientific journals. Active efforts are devoted to implementing innovative teaching methods aimed at enhancing students' professional competencies, developing research skills, and fostering creative potential. She can be contacted at email: yelepbergenovaaigul@outlook.com.

Alexey Fedenko holds a master's degree (2024–2026) in Technical Sciences, majoring in Computer Science and Engineering, from Astana IT University, Astana, Kazakhstan, and a bachelor's degree (2021–2024) in Information and Communication Technologies, majoring in Software Engineering, from the same university. He is the author and co-author of scientific papers in the fields of blockchain technologies and academic integrity systems, with a focus on building secure and transparent digital infrastructures for education. His professional and research interests include: software engineering; blockchain technologies and decentralized systems; machine learning and artificial intelligence; data-driven approaches to academic integrity; and the development of secure digital platforms for online assessment and certification. He can be contacted at email: supp.fed@gmail.com.