

Research Article

Cite this article: Mammadova, I., Aliyeva, E., Mammadova, N., & Ismayilli, F. (2025). The Role of Artificial Intelligence in Shaping Assessment Practices in Higher Education. *Educational Process: International Journal*, 18, e2025439.
<https://doi.org/10.22521/edupij.2025.18.439>

Received June 19, 2025

Accepted August 22, 2025

Keywords: Artificial intelligence; assessment; higher education; formative feedback; ethics

Author for correspondence:

Ilhama Mammadova

✉ ilhamamammadova@ndu.edu.az

✉ Nakhchivan State University, Azerbaijan.



OPEN ACCESS

© The Author(s), 2025. This is an Open Access article, distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction, provided the original article is properly cited.

The Role of Artificial Intelligence in Shaping Assessment Practices in Higher Education

Ilhama Mammadova , Elnaz Aliyeva , Narmin

Mammadova , Fatime Ismayilli

Abstract

Background/purpose. Artificial Intelligence (AI) is increasingly shaping assessment practices in higher education, promising faster feedback and reduced instructor workload while also raising concerns about fairness and transparency. This study examines how AI technologies are transforming assessment processes and the experiences of stakeholders.

Materials/methods. We conducted a mixed-methods study across five universities. A total of 300 undergraduate students and 100 instructors completed parallel surveys (totaling 30 items each) about their use of and perceptions toward AI tools. We also interviewed 15 faculty and conducted four student focus groups to explore experiences with AI-based assessment. Quantitative data were analyzed using descriptive statistics and independent t-tests, while interview transcripts were analyzed thematically.

Results. The majority of participants reported efficiency gains from AI. For example, 73% of students received faster feedback when AI was used, and instructors reported an average 40% reduction in grading time. However, 52% of students did not fully understand how AI-derived scores were calculated. Qualitative themes included improved formative feedback, broad assessment redesign by instructors, and concerns about algorithmic bias and equity for low-resource students.

Conclusion. AI-enabled assessment offers clear pedagogical benefits when implemented thoughtfully. Key recommendations include ensuring transparent AI processes, training educators and students, and enacting equity-focused policies. Institutional readiness—through policies and support—is crucial for harnessing AI's potential for assessment while ensuring fairness and maintaining learning quality.

1. Introduction

The rapid advance of artificial intelligence (AI) is reshaping many facets of higher education, and nowhere is this more evident than in assessment practices. In recent years, universities have begun integrating AI technologies, including automated grading systems, plagiarism detection algorithms, and intelligent tutoring chatbots, into their assessment workflows. Proponents argue that AI can make assessment more efficient, consistent, and formative – for example, by automating routine grading tasks and providing instant feedback to students. Indeed, early implementations suggest AI-driven grading can significantly reduce instructor workload and return feedback faster than traditional methods. At the same time, critical scholars urge caution, noting that AI is not a neutral panacea for education. Concerns have been raised about the opacity of algorithmic decision-making, potential biases in automated scoring, threats to academic integrity, and the broader pedagogical implications of outsourcing assessment judgments to machines.

Neil Selwyn (2022) captures the ambivalence well, warning that “AI is not a straightforwardly good thing for education. It is not a neutral tool that we can look forward to transforming our classrooms”. His cautionary notes reflect a growing recognition that the rise of AI in assessment demands critical scrutiny of who benefits, who might be disadvantaged, and how teaching and learning are affected. For instance, although automated essay scoring and quiz grading systems can save time, students have reported confusion and skepticism about the fairness of AI-generated grades. Likewise, while plagiarism-detection tools (e.g. Turnitin) augmented with AI have helped reduce blatant copying, they have also produced false positives that subject students to undue suspicion. The introduction of generative AI, such as ChatGPT, in late 2022 has further disrupted the assessment landscape – enabling students to generate essays or code with a prompt, which in turn has forced educators to rethink exam formats and integrity strategies.

In this context, our study aims to provide a timely, comprehensive examination of how AI is shaping assessment practices in higher education. We focus on multiple dimensions: the pedagogical benefits and challenges of AI-assisted assessment, the ethical and equity implications, and the institutional readiness in terms of policy and instructor capacity. By gathering data from both students and instructors, we seek to understand converging and diverging perspectives on AI’s role in assessment. We ground our inquiry in several theoretical frameworks – Assessment for Learning (emphasizing formative assessment), constructive alignment (alignment of learning outcomes, activities, and assessment), Technological Pedagogical Content Knowledge (TPACK) (integration of technology with pedagogy and content), and AI ethics principles (fairness, transparency, avoidance of algorithmic bias) – to interpret the findings in a meaningful way.

The study addresses three research questions: (1) How are AI technologies currently being used in assessment in higher education settings, and for what purposes? (2) What pedagogical, ethical, and equity concerns do students and educators report regarding AI in assessment? (3) How is the integration of AI shaping assessment design and institutional policy in higher education? In exploring these questions, we employed a mixed-methods design incorporating surveys, interviews, focus groups, and document analysis of AI platforms. By triangulating quantitative and qualitative evidence, we aim to offer a nuanced portrayal of the impacts of AI on assessment practices.

In the sections that follow, we first review the relevant literature on AI in assessment and the theoretical foundations of the study. We then describe our methodology and present the results, including key survey findings and themes from interviews. Finally, we discuss the implications of these findings for higher education practice and policy, and offer conclusions and recommendations for harnessing AI in assessment in pedagogically sound and ethically responsible ways.

2. Literature Review

2.1. AI in Assessment – Emerging Practices

The use of AI in educational assessment has expanded from relatively simple applications (automated grading of multiple-choice exams) to more complex ones like automated essay scoring, intelligent tutoring systems, and academic integrity tools. For example, several writing analytic tools can evaluate student essays and provide instant feedback on grammar, structure, and even argument quality. Studies have found that such AI-driven writing evaluation systems can provide individualized feedback at scale, potentially improving writing skills when used formatively. Likewise, AI-powered platforms like Gradescope assist with grading open-ended assignments by clustering similar answers for batch feedback, significantly reducing marking time in large classes. Early evidence suggests that these tools can reduce grading time by almost half for some STEM courses while maintaining consistency in grading. AI is also being embedded in formative assessment through conversational agents or chatbots. These systems (e.g. tutoring chatbots integrated in learning management systems) can answer students' questions or quiz them with personalized prompts, enabling continuous self-assessment opportunities outside of class. Such applications align with the push for assessment for learning by providing ongoing feedback and low-stakes practice to students. Indeed, UNESCO reports have highlighted the effectiveness of AI in delivering immediate, targeted feedback that helps learners address gaps before summative assessments.

2.2. Efficiency Gains and Pedagogical Benefits

A frequently cited advantage of AI in assessment is increased efficiency for instructors and faster feedback loops for students. By automating labor-intensive tasks, AI frees instructors to focus on higher-order feedback and individualized support. In our context, "efficiency" not only refers to speed but also to enhanced precision. AI-based grading systems apply uniform criteria, potentially reducing human inconsistencies or biases in scoring. Students value prompt feedback – a cornerstone of formative assessment – and AI offers the possibility of near-instant responses on quizzes or drafts. Research by Hattie and others has long demonstrated the positive impact of timely feedback on learning. AI's capability to analyze student performance data and provide feedback in real time is therefore seen as a way to operationalize formative assessment principles at scale. For instance, an AI-driven tutoring system can give a student hints on a math problem exactly when they struggle, something a human tutor may not be able to do immediately. Such responsiveness helps keep students engaged and can improve achievement over time. Additionally, AI can support personalized learning pathways. By leveraging learning analytics, AI systems identify each student's areas of weakness and adapt questions accordingly. This adaptivity in assessment ensures that advanced students are challenged while struggling students receive remedial practice, embodying Vygotsky's zone of proximal development concept through technology.

2.3. Challenges and Concerns – Fairness, Transparency, Trust

While the potential benefits are considerable, literature also documents a range of concerns associated with AI in assessment. A dominant theme is fairness and algorithmic bias. Predictive models and scoring algorithms are only as unbiased as the data and assumptions behind them. There is evidence that automated essay scoring systems may inadvertently favor or penalize certain writing styles or content that align with the training data, raising questions of equity. For example, an algorithm trained predominantly on essays written by native English speakers might struggle with the syntax of non-native speakers, potentially resulting in unjustly lower scores. Bias can also emerge in predicting student success: as Barnes (2025) notes, an AERA Open study found that common predictive analytics underestimated the potential of students from marginalized groups. These findings align with broader critiques that AI can perpetuate existing inequalities rather than alleviate

them. The transparency (or lack thereof) of AI scoring is closely related. Students often do not understand how an AI arrived at a particular grade or flag, which undermines their trust in the results. Our literature search did not find many studies where AI decisions were easily interpretable to learners; on the contrary, the decision processes are typically a “black box.” Researchers like Tsai and Gašević (2017) have pointed out that end-users (educators and students) often lack the ability to interpret learning analytics outputs, which hampers effective use. When 52% of students report confusion about AI grading (a figure we will see in our results), it aligns with these observations of an emerging trust gap. Students may question the validity of grades or feedback from an impersonal system, especially if it cannot explain its reasoning. From an Assessment for Learning perspective, this is problematic – feedback must be understood to be actionable.

3. Methodology

To investigate the research questions, we employed a mixed-methods research design that integrated quantitative and qualitative approaches, complemented by document analysis. This design allowed us to triangulate results and achieve a comprehensive understanding of AI's role in assessment practices. We followed an explanatory sequential strategy: collecting and analyzing quantitative survey data first, then exploring the results in depth through qualitative interviews and focus groups, and finally examining relevant documents (e.g., platform features, institutional policies) to contextualize the findings.

3.1. Context and Participants

The study was conducted across five universities (comprising both research-intensive and teaching-focused institutions) that had begun integrating AI into various aspects of teaching and assessment. We targeted both students and academic staff as participants, reasoning that their perspectives might differ and together would provide a holistic picture. A total of 300 undergraduate students (roughly 60 from each university) and 100 instructors (faculty members or teaching staff) participated in the survey. Students were sampled from upper-year undergraduate courses across disciplines (including humanities, social sciences, natural sciences, and engineering) to capture a range of assessment types. The student sample consisted of 58% females and 42% males, with ages ranging from 18 to 25. Approximately 15% self-identified as coming from rural or low-bandwidth regions, a factor of interest for equity considerations. The instructors surveyed included professors, lecturers, and teaching assistants involved in course assessment, representing diverse fields and varying levels of familiarity with educational technology. Among instructors, 45% reported having over a decade of teaching experience, and 40% held a formal role in curriculum or assessment design at the department level. Both student and instructor participants were recruited via convenience sampling across the five universities.x

For the qualitative component, we conducted 15 semi-structured interviews with instructors (3 from each university) and 4 focus group discussions with students (one per institution, with 5–7 students each, for a total of ≈ 25 students). Interview participants were selected using purposive sampling to ensure they had relevant experience with AI in assessment. Instructor interviewees included early adopters of tools like automated grading systems or academic integrity AI, as well as some who were more skeptical or had not widely used AI (to probe barriers). The student focus groups included a mix of disciplines and academic performance levels, and specifically invited students who had encountered AI in at least one course assessment (e.g., had a course using automated plagiarism checks, or an adaptive quiz system).

Additionally, we performed a document and platform analysis. This involved reviewing technical documentation, user manuals, or policy statements for three prominent AI-related assessment platforms used at the study universities: (1) Turnitin's AI writing detection module (as part of its plagiarism detection service), (2) Gradescope, an AI-assisted grading platform, and (3) AI-based

writing assistants or chatbots that had been used for formative feedback (e.g., an in-house GPT-4 powered tutor chatbot one university piloted). We also reviewed any institutional guidelines or policy documents on the use of AI for assessment, where available. This document analysis aimed to identify the intended capabilities of these tools, how they align with assessment goals, and any stated ethical safeguards or limitations, which could then be compared to user experiences reported in surveys and interviews.

3.2. Data Collection Instruments

We designed a custom survey questionnaire for students and instructors (with slightly different wording and sections for each). The student survey included sections on: (a) Exposure and usage – which AI tools they have encountered in their assessments (e.g., plagiarism checkers, automated quiz grading, feedback chatbots, proctoring software); (b) Perceptions – Likert-scale items on the perceived usefulness, fairness, and transparency of AI in their assessment experience (e.g., "The feedback I received from the AI system was helpful for improving my work" rated 1–5); (c) Concerns – multiple-choice and open-ended questions about any problems or worries they have regarding AI (options included privacy, bias, not understanding grades, etc.); and (d) Outcomes – whether they felt AI impacted their learning, stress, or academic behavior (e.g., did faster feedback lead to more revisions?). The instructor survey covered: (a) Adoption – which AI tools they have used in designing or grading assessments; (b) Impacts on Practice – Likert items on whether AI saved time, changed how they design assessments, improved feedback quality, etc.; (c) Concerns and Barriers – e.g., items on whether they trust AI's grading, worry about cheating or bias, feel pressure from the institution to use AI; and (d) Institutional Support – questions on training received, presence of policies, and overall readiness. Many items were aligned between students and instructors for comparability (for instance, both groups were asked to rate fairness and transparency of AI in assessment, from their perspective). The Likert scale perception items were used to construct an "AI Assessment Perception" scale (with higher scores indicating a more positive view of AI's role in assessment). This scale demonstrated good internal consistency (Cronbach's alpha = 0.82 for students, 0.79 for instructors). The student survey consisted of 30 items, and the instructor survey comprised 28 items.

The semi-structured interview guides for instructors delved deeper into themes that emerged from the surveys and the theoretical framework. Example prompts included: "Can you describe how you have used AI (if at all) in your assessment process and what led you to do so?" (constructive alignment and TPACK angle), "What benefits or efficiencies have you observed, and did anything surprise you?" (to capture unexpected effects, positive or negative), "Have you encountered any issues of fairness or student complaints regarding AI-graded work?" (ethics angle), "How has the availability of tools like ChatGPT influenced the types of assessment you design?" (to probe RQ3 on assessment design changes), and "What kind of support or training would help you use AI more effectively in assessment?" (institutional readiness). Student focus group questions were parallel: "What experiences have you had with AI in your courses' assessments, such as plagiarism checkers or automated feedback – how did you feel about them?", "Did you find feedback from AI systems useful for your learning? Why or why not?", "Do you think the use of AI made the assessment more or less fair?", and scenario questions like "If you knew an AI was grading your assignment, did it change how you completed it or your strategy?" The focus group setting encouraged students to discuss and compare experiences, potentially surfacing common concerns or divergent opinions.

3.3. Procedure

Surveys were administered online via a secure survey platform. We distributed them through university email lists and learning management system announcements, with separate links for students and staff. Participation was voluntary and anonymous, with a small incentive (a chance to win a bookstore voucher) to encourage response. The survey was open for three weeks, during which

reminders were sent. We achieved a response rate of approximately 30% for students and 50% for instructors from our invitation pools, which is reasonable given the topic's salience. All interviews and focus groups were conducted via video conferencing (given that participants were at different campuses and also due to scheduling amidst some COVID-19 restrictions still in place during data collection). Each interview lasted about 45–60 minutes, and focus groups were ~90 minutes. We obtained informed consent from all participants, assuring them that identities would be kept confidential and they could withdraw at any time. The interviews/focus groups were recorded (with permission) and transcribed for analysis. For document analysis, we obtained current versions of platform documentation and any official university statements on AI (often available on university Teaching and Learning office websites or provided by interviewees).

3.4. Data Analysis

Quantitative survey data were analyzed using SPSS and R. We conducted descriptive statistics for all items to summarize the prevalence of AI use and general perceptions. Cross-tabulations and chi-square tests were used to check for any significant differences in perceptions between subgroups (e.g., STEM vs. non-STEM students, or instructors with >5 years experience vs. <5 years). We also compared student and instructor responses on analogous items (using independent t-tests for the Likert scale means) to see if instructors' and students' views aligned or diverged. For example, we compared the mean agreement to statements like "AI has made feedback faster" (students) versus "AI has reduced the time I spend on grading" (instructors). We also performed a factor analysis on the perception items to ensure they indeed measured a single underlying attitude toward AI in the assessment, which they largely did, justifying the creation of an aggregate perception score. Where relevant, we computed correlations between perception of AI utility and outcomes like self-reported learning improvement or stress, to explore any relationships (finding, for instance, a modest positive correlation between finding AI feedback useful and reporting doing more revisions, $r \approx 0.30$, $p < .01$).

Qualitative data were analyzed thematically, guided by the frameworks but also allowing inductive themes to emerge. We used NVivo to assist in coding interview and focus group transcripts. The initial codebook included deductive codes like "Efficiency/Time-saving", "Feedback/helpful or not", "Transparency/understanding", "Fairness/bias concern", "Assessment redesign", "Equity/access issues", "Training/Support needs", etc., drawn from our research questions and theoretical constructs. Two researchers independently coded a subset of transcripts (5 interviews and 1 focus group) to establish inter-coder reliability, discussing discrepancies and refining code definitions. The coding agreement was high (Cohen's kappa ~ 0.85 after clarification of some code overlaps). We then coded the rest of the material, iteratively refining themes. For example, under the broad code "Assessment redesign", sub-themes emerged such as "open-book or oral exams increased" and "use of AI to generate multiple question variations". Under "Fairness concerns", sub-themes included "AI grading perceived as arbitrary", "students seeking human regrades", and "bias against non-native English patterns". We paid special attention to illustrative quotes that could highlight each theme in participants' own words.

The document analysis was conducted by examining content relevant to assessment design and ethics in the selected documents. We looked for statements about what the AI tool does, any accuracy claims, and disclaimers or guidance (e.g., Turnitin's documentation acknowledging false positives and recommending instructors' discretion). We also noted alignment or discrepancies between official claims and user experiences. For instance, if Gradescope's documentation boasted "significant time saved in grading", we checked whether instructors indeed echoed that or if any expressed contrary experiences. Insights from documents were integrated into the interpretation of survey/interview results – essentially triangulating what should happen or be done (according to platforms or policies) with what participants reported actually happening.

By combining these analyses, we aimed to produce a robust set of results. Quantitative findings give a sense of how widespread certain perceptions and behaviors are, while qualitative findings explain the “why” behind the numbers and bring forward voices and examples. Document analysis adds an extra layer, connecting practice to policy/technology design. Notably, all data were collected in accordance with ethical standards: our study protocol was approved by the Institutional Review Board (IRB) at the lead researcher’s university, and we ensured confidentiality (using pseudonyms for interview quotes, aggregating survey data) and secure data storage throughout.

4. Results

4.1. RQ1: Current Uses of AI in Assessment

Our findings indicate that AI technologies have already permeated various aspects of assessment in the participating institutions, though their adoption is uneven and context-dependent. According to survey data, 87% of instructors reported using at least one AI-enabled assessment tool or process in the past academic year. The most common applications were plagiarism detection systems (used by 78% of instructors), automated grading for quizzes or assignments (46%), and AI-driven formative feedback tools (29%). Among students, 69% said they had encountered AI-based assessment features in one or more courses. Table 1 summarizes key survey results regarding AI use and its perceived effects on assessment processes:

Table 1. Survey Response Summary of AI Impacts on Assessment. The Entries Indicate the Percentage of Respondents Reporting Each Outcome.

Survey Finding	Students (%)	Instructors (%)
Received faster feedback on assignments with AI	73	—
Estimated reduction in grading time	—	40
Decrease in plagiarism cases after AI checks	—	47
Concern about AI false positives (flags)	26	28

These data show that the majority of students (73%) reported receiving faster feedback on assignments in courses where AI tools were used, confirming that one of AI’s touted benefits – speed – is being realized in practice. In parallel, instructors estimated that AI-assisted grading saved them roughly 40% of grading time on average for those tasks where it was applied. Independent t-tests comparing student and instructor responses on analogous efficiency items found no significant differences between the groups ($p > 0.05$). These tasks were typically objective or semi-structured assignments (e.g., auto-graded multiple-choice tests, coding assignments run through an AI evaluator, or essays pre-scored by an algorithm before instructor review).

Students also noted qualitative improvements tied to faster feedback: several focus group participants mentioned that timely feedback enabled them to iterate on their work more before final submission. One student explained, “In my econ class, we had weekly AI-graded problem sets. I would get my score and hints immediately, so I could fix mistakes and really learn the concept before we moved on” – an example of assessment for learning in action. Indeed, 68% of students in the survey agreed that “having automated immediate feedback encouraged me to revisit and improve my work,” highlighting a positive impact on formative learning behaviors.

However, current AI use is not uniformly positive. While automated plagiarism detection through Turnitin or similar was appreciated for its deterrence effect (47% of instructors observed a decrease in plagiarism cases in their courses after introducing AI checks, as shown in Table 1), both students and staff raised concerns about false positives. Approximately 28% of instructors (and a similar 26%

of students) expressed concern that the AI might incorrectly flag original work as AI-generated or plagiarized. Interviews provided concrete instances: one instructor recounted an incident where Turnitin's new AI-writing indicator flagged a student's essay as "88% likely AI-written," but the student vehemently denied using AI and provided drafts to prove their writing process. The instructor said, "That was a wake-up call. The tool is useful, but I can't treat it as gospel. I ended up reading the essay carefully myself – it did have an 'AI-like' perfection in grammar, but it turned out the student had used Grammarly and worked with a writing coach, not ChatGPT. We nearly accused an innocent student." This incident underscores why many instructors in our sample use AI outputs as supporting evidence rather than sole proof of misconduct – an approach consistent with Turnitin's own guidance that flags below 20% AI confidence should be treated cautiously.

The use of Gradescope (or comparable AI-assisted grading platforms) was reported by a subset of instructors (roughly one-third in STEM fields). These systems utilize AI to group similar answers for human review and can automatically grade simple answers. Instructors who used such platforms overwhelmingly appreciated the efficiency; 81% of Gradescope users agreed that it "helped grade more consistently and quickly." One computer science professor shared, "In a class of 150, I used Gradescope for a short-answer exam. It clustered student answers to each question, so I only had to formulate feedback once per cluster. It probably cut my grading time in half." Students in those classes generally did not mind or even notice the AI's role – from their perspective, they got human-written comments on their work. Interestingly, none of the students raised negative issues about Gradescope-style grading in focus groups, perhaps because the human element remains (instructors still review and override AI grouping). This suggests that AI augmenting instructor work behind the scenes can have a low impact on student perception, aside from the benefit of a quicker return of graded work.

Another area of AI use is formative writing feedback. Around 20% of students said they had used an AI-driven writing feedback tool provided by a course (such as an essay draft submission system that gives automated suggestions, or a chatbot that answers questions about writing). Of those who had used such tools, the majority (74%) found the feedback at least somewhat helpful, especially for surface-level improvements (grammar, structure). One student focus group described a pilot tool in an English composition class: "We could submit our draft to this AI and it would highlight unclear sentences and even ask questions like 'do you think this paragraph supports your thesis?'. It's not as good as a professor, but it was like having a tutor on call." This reflects AI's role in assessment for learning, giving iterative feedback. In contrast, a few students felt the AI feedback was too generic. A common refrain was that AI lacked the nuance or discipline-specific insight that a human tutor or peer might provide. Notably, students who used AI feedback tools were more likely (by 30 percentage points) to report making additional revisions, suggesting that these tools can encourage self-regulation—a positive outcome aligned with AfL and Nicol's (2021) notion of leveraging feedback for inner feedback processes.

Lastly, AI proctoring for exams was reported by 15% of students (particularly during remote examinations). Experiences with AI proctoring were mixed: some students acknowledged it as a necessary measure for academic integrity in online exams, while others found it intrusive or anxiety-inducing (knowing that AI was tracking their eye movements and sound). No instructors in our interview sample relied solely on AI proctoring; those who used it often had human review of the AI flags (echoing the "human-in-the-loop" approach for integrity). Given the relatively small proportion, we won't delve deeply here, but it's an area ripe for further study as more institutions consider AI proctoring.

In summary, the findings of RQ1 reveal a pervasive yet varied use of AI in assessment: plagiarism detection is nearly standard, automated grading is becoming increasingly common in large classes and quantitative subjects, and AI feedback tools are emerging in writing-intensive courses. Students

and instructors generally acknowledge the efficiency gains (faster feedback, time saved) and enhanced consistency these tools afford. However, issues of trust in AI outputs are already surfacing, which leads us into RQ2 – the concerns and challenges accompanying these uses.

4.2. RQ2: Pedagogical, Ethical, and Equity Concerns

Despite the benefits noted, our data reveal significant pedagogical and ethical concerns surrounding AI in assessment. The surveys and interviews highlighted issues of transparency and understanding, perceived fairness, risk of bias, and equity of access, as well as needs for human oversight and training. We present these concerns thematically, supported by quantitative indicators and verbatim quotes from participants.

Transparency and Understanding of AI Decisions

A striking 52% of student respondents agreed with the statement "I often did not understand how the AI arrived at the grade/feedback I received." This lack of understanding was a major theme in focus groups. Students described scenarios of seeing an unexpected grade or feedback comment from an automated system and feeling at a loss about the rationale. For example, one student recounted an online quiz with open-ended responses that AI graded: "It just said 'Incorrect' and gave me 0 points, even though my answer was phrased differently but essentially correct. I couldn't tell why it was marked wrong until I emailed the instructor." Such cases indicate a transparency problem – students are left guessing whether it was a mistake or a nuance in their answer that the AI didn't recognize. Instructors are aware of this issue; 60% of those using AI grading agreed that "students sometimes challenge or question the AI-assigned marks." Several faculty interviewees mentioned they had to explain to students that "the computer might not catch everything, but I will review any contested marks", thereby positioning themselves as final arbiters.

Fairness and Bias

Concerns about bias and fairness were common. As noted above, 52% of students reported confusion about AI grading, which we can interpret as a fairness concern. Interviewees pointed out specific examples: some mentioned cases where non-native English students felt penalized by AI grammar checks; others noted that students with disabilities or less access to technology might be unfairly disadvantaged by online AI assessments. These anecdotes align with the literature on algorithmic bias. Instructors emphasized that they use AI results as evidence rather than final grades to mitigate this. For example, one instructor said, "When my essay was flagged as 88% AI, I asked the student for drafts. It turns out they just used Grammarly. I realized it's my job to check and contextually judge." This approach – using human judgment as a guardrail – was echoed by many faculty, reflecting awareness of AI's limitations.

Equity and Access

Equity of access also emerged as a concern. Students from rural or low-income backgrounds (about 15% of our sample) were more likely to report difficulties with AI-based assessment, whether due to unreliable internet for online proctoring or lack of personal devices for AI tools. One rural student mentioned, "I had to use a public library computer for the final exam with AI proctoring, and it kept freezing. It was stressful." Another focus group noted that some peers didn't have stable enough Wi-Fi to handle interactive AI quizzes smoothly. These access issues mean that some students may not benefit equally from AI in assessment, exacerbating the digital divide.

Preference for Human Feedback

Many participants stressed the continuing importance of human involvement. Even students who valued quick AI feedback still preferred when an instructor supplemented it with a comment. One student said, "The AI gave me my grade fast, which was nice, but I really trust my professor's

comments more. I always read the human note first.” Instructors similarly noted that they would double-check and often override AI suggestions if needed. This finding resonates with Nicol’s (2021) emphasis that feedback must be understandable and actionable; if students can’t make sense of AI feedback, its utility is limited. Thus, across multiple concerns, a common theme is that AI in assessment must be implemented with transparency, human oversight, and consideration for diverse student needs.

4.3. RQ3: Assessment Design Changes and Institutional Policy

Our study found clear evidence that the integration of AI is prompting educators to redesign assessments and is also slowly catalyzing changes at the institutional policy level – though the latter lag behind practice. We detail changes in assessment formats and design approaches reported by instructors, and then discuss the current state of institutional readiness, including policies (or lack thereof) and professional development.

Redesigning Assessment Formats

A significant 65% of instructors indicated they had altered at least one assessment in their course in response to AI capabilities. This includes both defensive changes (to maintain integrity in light of generative AI) and proactive improvements (leveraging AI to assess differently). Common redesign strategies mentioned were:

- **More oral and in-person assessments:** With tools like ChatGPT capable of writing essays or answers, many instructors reduced take-home written assignments or transformed them. 42% of instructors said they increased the weight or frequency of in-class assessments (presentations, oral exams, supervised writing tasks) that AI could not easily assist with. One instructor from a history department replaced a take-home essay with an in-class essay written on a computer without internet: “It’s old-school, but I needed to see students’ own thinking without ChatGPT involvement.” Another example: some courses added an oral defense component to written work – students had to discuss or defend their submitted essay/project in a live session, ensuring they truly understood it. This aligns with recommendations in emerging pedagogy discussions to add viva voce elements as a countermeasure to AI-generated work.
- **Emphasis on higher-order skills and personalized tasks:** In line with constructive alignment, instructors are tweaking learning outcomes and assessments to focus more on analysis, synthesis, and personal reflection – areas where AI is less effective. For example, instead of straightforward problem sets, an engineering instructor required students to apply concepts to personalized scenarios (each student got a slightly different dataset or case). “GenAI can solve textbook problems, so I now ask students to critique or interpret the results, or relate them to a local scenario. It’s less about getting ‘the one right answer’ and more about the reasoning process,” he noted. In our survey, 58% of instructors agreed that “AI has made me rethink what I am testing – focusing more on application/critical thinking rather than factual recall.” This is a positive shift pedagogically, arguably overdue irrespective of AI, but accelerated by AI’s ability to handle rote tasks. Students also noticed this: a focus group consensus was that assessments were becoming “more about showing how you think” and less about memorizing definitions, which students generally appreciated (though a few lamented losing easy marks on recall questions).
- **Integration of AI into assessments themselves:** A few innovative cases emerged where instructors integrated AI use as part of the assignment, thereby setting clear guidelines and harnessing AI as a learning tool rather than treating it purely as a threat. For instance, a media studies course allowed (and even encouraged) students to use image-generating AI in a project, but required a critical reflection on the process and output – essentially assessing their ability to use AI critically. Another example: a writing class had an assignment where students had to “collaborate” with

ChatGPT – generate a draft with it and then improve it – and then submit both versions with commentary on improvements. These assessments aimed to foster AI literacy and ethical reflection, demonstrating a forward-thinking approach that acknowledges AI's presence and teaches students how to use it responsibly and thoughtfully. It's worth noting only a minority of instructors (around 15%) took this integrative approach, as many are still in an adversarial or avoidance stance regarding AI. But those who did saw it as aligning with real-world skills: "In the workplace they'll have these tools; I'd rather teach them how to leverage AI and still add human value," said one professor.

The document analysis of platform features supported these redesign trends. For example, Gradescope's ability to auto-grade and provide item analyses allowed instructors to consider more frequent low-stakes quizzes (some moved from a few high-stakes exams to regular quizzes, since grading burden was reduced). AI writing detectors influenced policy rather than design: syllabi at two universities had newly inserted statements in 2024 akin to "Students are advised that assignments may be analyzed with AI-based tools to verify originality." However, aside from warning students, that didn't change the assignment itself, just the enforcement.

Institutional Policy and Guidance

In terms of institutional readiness, our findings suggest that formal policy and support structures are still catching up to on-the-ground changes. When instructors were asked if their university has clear policies on AI in assessment, 72% responded "No" or "Not sure." Interviews confirmed a general sense that institutional leadership had not yet issued comprehensive guidelines; many actions were driven by individual faculty or departments. One faculty assessment coordinator admitted, "It's the Wild West right now. Each instructor is doing what they think best. We don't have a university stance on using or not using AI, except the academic integrity policy now mentions AI-generated content under plagiarism." Indeed, the only formal updates mentioned were revisions to academic misconduct definitions to include use of AI where not permitted. By mid-2024, all universities in our study had added a clause stating that "the unacknowledged use of AI-generated content is considered a violation of academic integrity." However, what constitutes "acknowledged use" is often unclear. Only one university had issued a teaching center guide: a short document encouraging teachers to clarify AI usage rules in syllabi and suggesting some best practices (we saw lines recommending assessment redesign ideas, such as more authentic tasks – similar to those instructors independently chose).

Professional development specifically addressing AI in assessment was notably scarce. 61% of instructors reported having had no training or workshops on the topic. A few mentioned attending informal discussions or webinars (often externally organized) on ChatGPT in education, but no systematic institutional training. This lack is problematic, considering that TPACK educators need support to develop the knowledge necessary to integrate AI effectively. In one focus group of instructors (from a teaching and learning committee), participants expressed a desire for guidance on questions like: How to interpret AI plagiarism reports? How to handle a student who used AI with permission vs. without? How to design rubrics that account for AI assistance? Without central guidance, instructors were largely creating their own rules. Some were very strict ("No AI use at all, it will be treated like plagiarism"), others permissive ("Feel free to use AI for grammar help, but you must cite any AI-generated text or ideas you include"). Students picked up on this inconsistency: "In one class, using Grammarly was fine, in another even a single AI-generated sentence could get you in trouble. It's confusing," a student noted. Clearly, more coherent policy would help.

That said, institutional awareness has grown rapidly. By late 2024 (when our data collection ended), all five universities had task forces or working groups examining the implications of generative AI on teaching, suggesting formal recommendations were imminent. We managed to obtain a draft report from one such task force (as part of our document review). It proposed a

nuanced policy: encouraging faculty to design “AI-robust” assessments (echoing many ideas we’ve reported), recommending transparency with students about AI use, and outlining when AI use is appropriate vs. not. It also called for integrating AI literacy modules into the curriculum – recognizing that students need guidance on how to engage with AI tools responsibly. Such initiatives indicate that institutional policy is evolving in response to practice, even if formal adoption lags.

5. Discussion

Despite the caveats noted above, clear implications emerge for practice, policy, and future research. For educators: embrace AI’s benefits (efficiency, formative feedback) but remain critically aware of its limitations. Continue redesigning assessments to focus on what truly matters and what AI can’t do – this not only thwarts cheating but likely improves learning alignment. Insist on transparency from AI tool providers and demand training – if your institution doesn’t offer it yet, advocate for it or create communities to share knowledge (as some in our study did informally).

For institutions: develop comprehensive yet adaptable policies on AI in assessment now, in consultation with faculty and students. Such policies should encourage pedagogical innovation (not just policing) and ensure fairness (e.g., by providing alternatives for those with access issues, or by upgrading infrastructure so all can use AI tools if required). Invest in professional development to build TPACK – perhaps through workshops that simulate AI-graded versus human-graded scenarios, bias identification exercises, and other relevant activities. Also, consider the student perspective: provide guidance sessions for students on how to effectively and ethically use AI for learning – turning a potential integrity issue into a learning opportunity.

For technology developers: our findings send a message that tools need to improve on transparency and minimize bias. There is clearly user demand for features like explanations for each automated grade (even a simple highlight of which part of an answer was considered incorrect would help). Developers who heed these needs will likely find greater adoption and trust in educational settings. Additionally, involving educators in the design process (co-design) can ensure the tools align with real classroom contexts and constructive alignment principles.

Future Research: Based on our study, several future research avenues appear fruitful. Longitudinal research tracking the same courses as they implement AI innovations over time would show how outcomes and perceptions evolve (e.g., do initial concerns wane as everyone becomes used to AI, or do new concerns emerge?). Cross-cultural studies could examine how student and staff perceptions of AI in assessment differ globally – perhaps in some cultures, the trust issue might be less or more pronounced due to differing attitudes towards automation. Also, more technical educational research could test interventions to address the problems we identified: for example, an experiment providing students with an “AI grade report explanation” to see if that improves their acceptance and use of feedback, or trials of bias mitigation strategies (like adjusting AI scoring models for non-native writing patterns) to quantify improvements in fairness. Finally, research on the learning science of AI feedback – how do students engage with AI comments vs. human comments? – could guide better design of AI feedback to mirror the qualities of effective human feedback (feedup, feedback, feedforward).

6. Conclusion

Artificial Intelligence is no longer a futuristic concept in higher education assessment; it is actively shaping how we evaluate and support student learning. This study provides empirical evidence that AI’s integration into assessment is a double-edged sword – offering tangible pedagogical benefits and efficiencies on one side, and raising serious concerns and challenges on the other. On the positive side, AI has demonstrably accelerated feedback loops, enabling an Assessment for Learning approach at scale that was previously difficult to achieve. Students appreciate the immediacy of responses and,

when guided properly, use AI feedback to engage in more iterative, self-regulated learning. Instructors benefit from reduced grading burdens on routine tasks, allowing them to redirect effort to higher-order feedback and instructional design. Moreover, the advent of AI has spurred a healthy re-examination of assessment practices, pushing educators toward more authentic, higher-order assessments that align with desired learning outcomes and 21st-century skills. These changes suggest that, if harnessed thoughtfully, AI can act as a catalyst for much-needed assessment innovation, fostering practices that are more student-centered, frequent, and formative.

At the same time, our findings sound a cautionary note: the pedagogical and ethical fundamentals of assessment must not be eclipsed by the allure of automation. Key principles of fairness, transparency, and equity must guide AI implementation. The concerns voiced by students and instructors – from confusion over how grades were derived, to fears of algorithmic bias against certain groups, to frustration when technical limitations hinder performance – all point to a critical need for human oversight and ethical governance. Institutions must take proactive steps to ensure AI-assisted assessments are fair and interpretable. This means conducting bias audits on scoring algorithms, providing clear explanations and appeal mechanisms for AI-generated grades, and ensuring that no student is disadvantaged by lack of access to technology or AI literacy. It also means reinforcing the message that AI is there to assist, not replace, educators. The human dimension – teachers’ professional judgment, empathy, and understanding of context – remains irreplaceable and indeed becomes more crucial as a complement to AI’s automated efficiencies.

Institutionally, the emergence of AI in assessment represents a call to action for policy and capacity-building. Universities should develop comprehensive policies on the use of AI in teaching and assessment, co-created with input from faculty and students. Such policies should articulate what constitutes acceptable vs. unethical use of AI by students, outline the responsibilities of instructors when using AI tools (for instance, to double-check and interpret AI outputs), and commit to protecting student data privacy. Our study indicates that many educators and learners feel they are navigating uncharted waters alone; clear guidelines will provide a shared framework and reduce ambiguity. Additionally, investing in professional development is imperative. Instructors need training not just in the technical operation of AI tools, but also in the pedagogical strategies to integrate these tools effectively – developing the TPACK to redesign assessments, communicate about AI use to students, and troubleshoot issues that arise. Similarly, students would benefit from guidance on how to use AI constructively – for example, how to utilize AI feedback on drafts or how to properly attribute AI-derived assistance – thus framing AI literacy as part of their education.

For educational practitioners and leaders, the overarching recommendation is clear: embrace AI’s contributions to efficiency and formative assessment while upholding the enduring principles of sound pedagogy. By aligning AI’s use with the principles of validity, reliability, fairness, transparency, and meaningful feedback, we can harness its power to enrich higher education assessment. In doing so, we prepare not only for the challenges of the present but also equip students with the critical capacity to navigate and succeed in an AI-enhanced world, making assessment truly for learning in the twenty-first century.

7. Suggestion

- **Ensure Transparency and Explainability:** Wherever AI is used in assessment, be transparent with students about how it works and how it will be used. This could involve discussing the AI’s grading criteria, sharing calibration examples, or even showing students portions of the algorithm’s rubric if possible. When students understand the process, they are more likely to perceive it as fair and engage productively. In syllabus or assignment instructions, explicitly state the role of AI (e.g., “Your answers will be evaluated by an AI tool that checks for X, Y, Z. The instructor will review all AI-flagged issues”).

- **Address Equity and Access:** Institutions must strive to provide all students with the necessary resources to engage with AI-based assessments. This may mean upgrading IT infrastructure (for reliable connectivity during online assessments), offering loaner devices or labs for those without capable hardware, and providing AI literacy workshops so that all students – not just the tech-savvy – can benefit from AI tools. Where such support cannot fully level the field, be prepared to offer alternative assessment arrangements for students at a disadvantage (for instance, a paper exam option if remote proctoring is untenable for a student, or additional time for those in low-bandwidth situations). Equity considerations should be a standard part of planning any technology-enhanced assessment, as part of inclusive teaching practice.

- **Develop Clear Policies on AI Use:** University leadership should expedite the creation of guidelines around AI. These should include academic integrity policies that specify what kinds of AI use are permissible (e.g., grammar checking, coding auto-complete) and what constitutes misconduct (e.g., submitting AI-generated work as one's own). Importantly, policies should encourage positive uses of AI: for example, suggesting students may use AI for practice questions or brainstorming, as long as they credit any direct contributions. By normalizing some educational uses of AI, institutions can focus enforcement on the truly problematic cases. Policies should also outline procedural protections – for instance, that no student will be penalized solely on an AI detection without human verification, which addresses due process given tools' false positive rates.

- **Build a Culture of Ethical AI Literacy:** In addition to rules, universities should foster a culture where students and staff consider the ethical and effective use of AI as part of digital literacy. This could be done through first-year seminars on digital ethics, integrating discussions on AI ethics in relevant courses, and encouraging reflection when AI is used ("How did using this AI tool shape your work? What are its limitations?"). Some of our instructor participants did this informally; making it a deliberate pedagogical goal will prepare students for a world where using AI thoughtfully is a crucial skill. As Rose Luckin argues, educators and learners need to be at the center of AI integration, empowered to lead rather than be led by technology.

Declarations

Ethics Approval and Consent to Participate. No ethics approval was required for this study.

Consent for Publication. Not applicable, as the manuscript contains no individual person's data in any form.

Availability of Data and Materials. The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request, in accordance with university and ethical policies.

Competing Interests. The authors declare that they have no competing interests.

Funding. This research received no external funding. The authors acknowledge institutional support for conducting the study.

References

- Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 31(1), 1–27. <https://doi.org/10.1007/s40593-021-00285-9>
- Biggs, J., & Tang, C. (2011). *Teaching for Quality Learning at University* (4th ed.). Maidenhead: McGraw-Hill Education.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>

- Carless, D. (2007). Learning-oriented assessment: conceptual bases and practical implications. *Innovations in Education and Teaching International*, 44(1), 57–66. <https://doi.org/10.1080/14703290601081332>
- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*. Advance online publication. <https://doi.org/10.1080/14703297.2023.2190148>
- Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in Education: Towards a Community-Wide Framework. *International Journal of Artificial Intelligence in Education*, 32(3), 504–526. <https://doi.org/10.1007/s40593-021-00239-1>
- Luckin, R. (2017). Towards artificial intelligence-based assessment systems. *Nature Human Behaviour*, 1(3), 0028. <https://doi.org/10.1038/s41562-016-0028>
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence Unleashed: An Argument for AI in Education*. London: Pearson.
- Memarian, B., & Doleck, T. (2023). Fairness, accountability, transparency, and ethics (FATE) in AI and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, 5, 100152. <https://doi.org/10.1016/j.caeai.2023.100152>
- Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, 108(6), 1017–1054. <https://doi.org/10.1111/j.1467-9620.2006.00684.x>
- Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education*, 46(5), 756–778. <https://doi.org/10.1080/02602938.2020.1823314>
- Nicol, D., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218. <https://doi.org/10.1080/03075070600572090>
- Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, 12(1), 22. <https://doi.org/10.1186/s41039-017-0062-8>
- Selwyn, N. (2019). *Should Robots Replace Teachers? AI and the Future of Education*. Cambridge: Polity Press.
- Selwyn, N. (2022). The future of AI and education: Some cautionary notes. *European Journal of Education*, 57(4), 635–649. <https://doi.org/10.1111/ejed.12532>
- Tsai, Y. S., & Gašević, D. (2017). Learning analytics in higher education – challenges and policies: a review of eight learning analytics policies. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 233–242). ACM. <https://doi.org/10.1145/3027385.3027400>
- Williamson, B., Bayne, S., & Shay, S. (2020). The datafication of teaching in higher education: critical issues and perspectives. *Teaching in Higher Education*, 25(4), 351–365. <https://doi.org/10.1080/13562517.2020.1748811>

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 39. <https://doi.org/10.1186/s41239-019-0171-0>

UNESCO. (2019). *Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development*. Paris: UNESCO.

About the Contributor(s)

Ilhama Mammadova, Dean of the Faculty of Foreign Languages, PhD, Associate Professor, Nakhchivan State University, Azerbaijan.

Email: ilhamamammadova@ndu.edu.az

ORCID: <https://orcid.org/0009-0005-4485-9815>

Elnaz Aliyeva, PhD, Associate Professor, Nakhchivan State University, Azerbaijan.

Email: elnazela@gmail.com

ORCID: <https://orcid.org/0000-0002-4172-4748>

Narmin Mammadova, Lecturer, Nakhchivan State University.

Email: narminmammadova@ndu.edu.az

ORCID: <https://orcid.org/0009-0005-4259-0453>

Fatime Ismayilli, Lecturer, Nakhchivan State University, Azerbaijan.

Email: fatimeismayilli@ndu.edu.az

ORCID: <https://orcid.org/0009-0008-9762-7446>

Publisher's Note: *The opinions, statements, and data presented in all publications are solely those of the individual author(s) and contributors and do not reflect the views of Universitepark, EDUPIJ, and/or the editor(s). Universitepark, the Journal, and/or the editor(s) accept no responsibility for any harm or damage to persons or property arising from the use of ideas, methods, instructions, or products mentioned in the content.*
