INTERNATIONAL JOURNAL OF STATISTICAL ANALYSIS

ISSN: 2690-2265
ACCEPTED: 06.10.2026
PUBLISHED: 07.10.2026

REDESIGNING ASSESSMENT IN HIGHER EDUCATION IN THE DIGITAL ERA: THE “IZLI” MODEL OF PROCESS-TRACE-BASED ASSESSMENT

Usmanova Shoira Djumabekovna

Associate Professor of the Department of "Educational Technologies" Branch (Regional) Center for Retraining and Advanced Training of Pedagogical Staff under the National University of Uzbekistan

Annotation

This article argues that product-based assessment in higher education is losing validity under generative artificial intelligence (AI) and proposes IZLI, a model based on process traces. The model combines four components – process trace, accountability declaration, evaluative-judgement capacity and interactive defence – with an authorship congruence coefficient that measures agreement between product and defence scores.

Keywords

higher education; assessment; generative artificial intelligence; assessment validity; academic integrity; process-based assessment; evaluative judgement; IZLI model.

Annotasiya: Maqolada generativ sun'iy intellekt (SI) sharoitida oliy ta'limda yakuniy mahsulotga tayanuvchi an'anaviy baholashning validligi pasayayotgani asoslanib, jarayon izlariga tayanuvchi «IZLI» modeli taklif etiladi. Model to'rt komponentni – jarayon izi, zimmadorlik deklaratsiyasi, baholovchi mulohaza layoqati va interaktiv himoyani – hamda mahsulot va himoya natijalarining mosligini o'lchovchi mualliflik muvofiqligi koeffitsientini birlashtiradi.

Kalit so'zlar: oliy ta'lim; baholash; generativ sun'iy intellekt; baholash validligi; akademik halollik; jarayonli baholash; baholovchi mulohaza; IZLI modeli.

Аннотация: В статье обосновывается снижение валидности традиционного оценивания в высшем образовании, опирающегося на итоговый

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 06.10.2026
PUBLISHED: 07.10.2026

продукт, в условиях генеративного искусственного интеллекта (ИИ), и предлагается модель «IZLI», основанная на следах учебного процесса. Модель объединяет четыре компонента – след процесса, декларацию ответственности, способность к оценочному суждению и интерактивную защиту – а также коэффициент авторского соответствия, измеряющий согласованность результатов по продукту и защите.

Ключевые слова: высшее образование; оценивание; генеративный искусственный интеллект; валидность оценивания; академическая честность; процессуальное оценивание; оценочное суждение; модель IZLI.

INTRODUCTION

Since the end of 2022, widely available generative artificial intelligence (AI) tools have called into question the core premise of assessment practice in higher education – that a product submitted by a student (an essay, an abstract, program code, a report) is a direct expression of that student's mastery of the subject.

Today, large language models can generate, within seconds, text that is coherently structured, academic in style and often satisfactory in content. This situation has produced a polarised response in the field of assessment: on the one hand, the path of prohibition and detection (detectors, strict supervision); on the other, the path of integrating AI into the learning process.

The weakness of the first approach is that the technical means of reliably establishing the authorship of a text are limited, while errors can seriously damage a student's academic reputation. The second approach, in turn, requires a rethinking of the logic of assessment itself. In our view, the essence of the problem is not "cheating" but a question of **validity**: as Messick noted, what is evaluated in assessment is not the test or task itself but the interpretation and use of the scores derived from it. If it is unclear who or what created the product, the interpretation of the score is weakened.

This issue is also pertinent to higher education in Uzbekistan. Decree of the President of the Republic of Uzbekistan No. PF-6079 of 5 October 2020 on the

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 06.10.2026
PUBLISHED: 07.10.2026

"Digital Uzbekistan–2030" Strategy[1] and Resolution of the President No. 358 of 14 October 2024 on approving the "Strategy for the Development of Artificial Intelligence Technologies until 2030" and measures for its effective implementation [2][2] establish the digitalisation of the education system as a priority task. At the same time, the use of AI tools in higher education institutions (HEIs) is expanding rapidly, while assessment procedures have not kept pace with this change. From a scholarly perspective, the development of a holistic model of assessment under AI conditions that encompasses process, evidence and defence has not been sufficiently studied.

The aim of this study is to develop a practically applicable model for preserving the validity of assessment in higher education under conditions of generative AI.

Literature Review

Evolution of Assessment Paradigms

Modern assessment theory distinguishes between assessment of learning, assessment for learning and assessment as learning. Black and Wiliam (1998) synthesised the evidence that formative assessment has a significant positive effect on learning outcomes; Sadler (1989), in turn, showed that the effectiveness of formative assessment depends on students' understanding of quality criteria. Nicol and Macfarlane-Dick (2006) interpreted feedback from the perspective of self-regulated learning, while Hattie and Timperley (2007) demonstrated that the power of feedback depends on its content. Through the idea of constructive alignment, Biggs (1996) calls for coherence among learning outcomes, teaching activities and assessment. Wiggins (1990) and Sambell et al. (2013) developed authentic assessment, emphasising that tasks should be close to real professional contexts.

However, even authentic tasks are not inherently protected against generative AI: a case analysis or project report grounded in a real situation can also be prepared to a high standard with the help of AI. For this reason, in assessment design, not only

[1]On approving the "Digital Uzbekistan–2030" Strategy and measures for its effective implementation. Decree of the President of the Republic of Uzbekistan PF-6079, 5 October 2020.
[2]On approving the "Strategy for the Development of Artificial Intelligence Technologies until 2030" and measures for its effective implementation. Resolution of the President of the Republic of Uzbekistan PQ-358, 14 October 2024.

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 06.10.2026
PUBLISHED: 07.10.2026

the nature of the task but also the **type of evidence** is becoming decisive.

Generative AI and Assessment

Zawacki-Richter et al. (2019), in systematising research on AI in higher education, identified assessment and monitoring as one of the principal areas of application. Once generative models appeared, research focused on two issues: the threat to academic integrity (Cotton et al., 2024) and the pedagogical opportunities and risks (Kasneci et al., 2023). Swiecki et al. (2022) stressed the need to reconsider assessment design in the AI era, and Lodge et al. (2023) emphasised that assessment reform must take human–AI collaboration into account. Dawson (2021) proposes viewing assessment security in the digital environment as a multi-layered defence.

The AI Assessment Scale (AIAS) proposed by Perkins et al. (2024) helps to specify, step by step, the level of AI use permitted in an assignment. Such scales ensure transparency; however, they typically do not measure the relationship between the declared level of use and **actual learning**. It is precisely this gap that is at the centre of the present article.

Evaluative Judgement and Self-Regulation

Tai et al. (2018) defined evaluative judgement as the capacity to understand the quality of work and to make decisions based on that understanding; Boud et al. (2018) regard it as one of the central goals of higher education. In the AI era, this capacity acquires new significance: the student must now be not merely a producer of text but an expert who evaluates the quality, errors and reliability of AI output. Carless and Boud (2018) developed the concept of feedback literacy, and Zimmerman (2002) the model of self-regulated learning; both provide a theoretical basis for assessing process traces.

Methodology

The study is design-based and conceptual-methodological in nature and was carried out in three stages.

In the first stage, a thematic synthesis of sources on assessment, AI and validity

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 06.10.2026
PUBLISHED: 07.10.2026

was conducted, and the validity inferences that are being weakened were identified (RQ1).

In the second stage, design principles were formulated on the basis of the synthesis, from which the model components and measurement rules were derived (RQ2).

In the third stage, a validity argument was constructed following the logic of Kane (2013), and a protocol for empirical validation was developed (RQ3).

An important note: the article does not report the results of empirical experiments. The numerical examples are illustrative in nature, intended to show how the model works, and should not be interpreted as real data. The proposed weights and threshold values are initial values that require calibration.

RESULTS: THE "IZLI" ASSESSMENT MODEL

Design Principles

Five principles were formulated on the basis of the synthesis:

- **P1. Priority of validity.** Assessment design is directed towards protecting the interpretation of scores before penalising cheating.
- **P2. Transparency and collaboration with AI.** The use of AI is not concealed; it is declared at graded levels.
- **P3. Multi-source evidence.** Conclusions rest not on a single artefact but on the triad of product, process and defence.
- **P4. Proportionality.** Additional evidence is kept to a reasonable volume for students and instructors (cognitive load).
- **P5. Equity and human oversight.** An algorithmic signal is never the final decision; all students have access to the same tools.

Structure of the Model

The name of the model derives from the Uzbek words *iz* (process trace) and *izli* (having a trace; traceable), and is also formed from the initial letters of its four components: **I**z (Trace), **Z**immadorlik (Accountability), **L**ayoqat (Capacity) and **I**nteraktiv himoya (Interactive defence). The product itself is also assessed; however,

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 06.10.2026
PUBLISHED: 07.10.2026

it is no longer the sole evidence but the "core" that is complemented by the four components (Table 1).

Table 1. Components of the IZLI model and initial weights

Component	Essence	Evidence (artefact)	Proposed weight
P – Product (core)	Quality of the content and form of the assignment	Final work	30%
I – Trace	Observable traces of the learning process	Version history, drafts, prompt log, short diary	15%
Z – Accountability	Open and precise declaration of AI use	Record of the Z-level (0–4), list of sources	10%
L – Capacity	Critical evaluation of AI output, identifying errors and correcting them with justification	"AI error map", annotated corrections	20%
D – Interactive defence	Short (5–7-minute) oral or online question-and-answer session on the work	Defence rubric, record of the session	25%

The overall index is calculated as follows:

$$IZLI = 0.30 \cdot S_P + 0.15 \cdot S_I + 0.10 \cdot S_Z + 0.20 \cdot S_L + 0.25 \cdot S_D \quad (1)$$

where S is the score on the corresponding component on a 0–100 scale. The weights shift depending on the Z-level (Table 2): the deeper the involvement of AI, the greater the share of the L and D components, because it is precisely these that show the student's personal contribution.

Accountability: Five Levels of AI Use

Table 2. Levels of AI use, mandatory evidence and weight shifts (to be calibrated)

https://creativecommons.org/licenses/by/4.0/

INTERNATIONAL JOURNAL OF STATISTICAL ANALYSIS
ISSN: 2690-2265
ACCEPTED: 06.10.2026
PUBLISHED: 07.10.2026

Level	Permitted use	Mandatory trace and evidence	Weight shifts (proposed)
Z-0	AI is not used (supervised environment)	Supervision record	Standard (L and I may be reduced; D ≥ 30%)
Z-1	AI for idea generation and planning	Draft plan, 2–3 sample prompts	Standard
Z-2	AI for language and style editing	Versions before and after editing	Standard
Z-3	Co-creation with AI	Full prompt log, versions, statement of own contribution	L = 25%, D = 30%, P = 20%
Z-4	AI-centred task (AI output is the object of analysis)	AI output, critical analysis, corrected version	L = 35%, D = 30%, P = 10%

The level is set in advance by the instructor for each assignment and is linked to the learning outcomes. For example, Z-0 or Z-1 is logically sound for an assignment that tests language skills, whereas Z-3 or Z-4 is appropriate for an assignment that tests competence in working with AI. For us, this approach is a continuation of Biggs's (1996) constructive alignment in the AI era.

Capacity Component: Rubric

The L component examines the student's evaluative judgement. The student is given 3–5 judgements or citations produced by AI and must verify them, identify the errors and correct them with justification. Table 3 presents a three-criterion rubric for L.

Table 3. Rubric for the L (Capacity) component

Criterion	High (80–100)	Medium (50–79)	Low (0–49)
Identifying errors	All major factual and logical errors in the AI output have been identified	Some of the errors have been identified	No errors have been identified, or a non-existent error has been reported

https://creativecommons.org/licenses/by/4.0/

INTERNATIONAL JOURNAL OF STATISTICAL ANALYSIS
ISSN: 2690-2265
ACCEPTED: 06.10.2026
PUBLISHED: 07.10.2026

Criterion	High (80–100)	Medium (50–79)	Low (0–49)
Verifying sources	All claims have been confirmed against independent sources	Claims have been partially confirmed	No verification has been carried out
Justified correction	Corrections are based on evidence and criteria	Corrections are present, but the justification is weak	No corrections, or the corrections are unfounded

Authorship Congruence Coefficient

The central novelty of the model is a coefficient that measures the difference between the product (P) and defence (D) scores. It is not suspicious for a student to perform better in the defence than in the product (the student may have reconsidered the work); therefore, the coefficient is defined as **one-sided**:

$$\Delta^{+} = \max(0; S_P - S_D) / 100 \qquad (2)$$

Interpretation rule (initial thresholds): $\Delta^{+} \leq 0.15$ – congruence is sufficient; $0.15 < \Delta^{+} \leq 0.30$ – 2–3 additional clarifying questions; $\Delta^{+} > 0.30$ – the final score is temporarily withheld and a second interview is held with the student. Importantly, the coefficient is **never an automatic basis for a penalty**: it is merely a signal that triggers additional human verification. So that factors such as defence anxiety, language barriers or health conditions can be taken into account, the student is offered an alternative format (written or online).

Illustrative Scenario

Let us consider the model using the example of a scientific mini-review assignment (Z-3 level) in the course “Information Technologies”. The student submits a 1,200-word review (P), provides a prompt log and version history (I), declares the use of AI (Z), completes an “error map” by verifying the authenticity of five citations generated by AI (L), and, in a 6-minute defence, explains the key

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 06.10.2026
PUBLISHED: 07.10.2026

decisions made in the work (D). The figures below are given solely for illustrative calculation (Table 4).

Table 4. Calculation of the IZLI index and Δ^+ (illustrative example; standard weights for the Z-3 level)

Student (hypothetical)	**S_P**	**S_I**	**S_Z**	**S_L**	**S_D**	**IZLI index**	**Δ^+ and decision**
A	88	70	90	65	52	71.9	0.36 – withhold; second interview
B	80	78	85	82	79	80.4	0.01 – congruent

Student A's product is of high quality (88), but in the defence (52) the student was unable to explain the logic of the work and showed a low result on the error map. Because $\Delta^+ = (88 - 52)/100 = 0.36 > 0.30$, the score is withheld. This does not mean that Student A cheated: the situation is clarified in a second interview. By contrast, Student B's product and defence are congruent, and the scores are supported by process traces. Calculation: for A, $0.30 \cdot 88 + 0.15 \cdot 70 + 0.10 \cdot 90 + 0.20 \cdot 65 + 0.25 \cdot 52 = 71.9$; for B, $0.30 \cdot 80 + 0.15 \cdot 78 + 0.10 \cdot 85 + 0.20 \cdot 82 + 0.25 \cdot 79 = 80.35$.

Validity Argument

Following the logic of Kane (2013), the validity argument for the model is constructed through four inferences (Table 5).

Table 5. Form of the validity argument for IZLI

Inference	**Basis (claim)**	**Method of evidence collection**
Scoring	Rubrics ensure a clear link between the score and the observed performance	Expert evaluation; inter-rater agreement (ICC)
Generalisation	Scores are stable across tasks, raters and time	Repeated measurement; Cronbach's α; G-study
Extrapolation	Scores reflect real academic and professional competence	Correlation with results in other courses and with portfolios

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 06.10.2026
PUBLISHED: 07.10.2026

Implication	The consequences of decisions are fair and beneficial for learning	Student and instructor surveys; equity analysis

Analytical Comparison with Alternative Approaches

Table 6 compares the IZLI model with three widely used approaches. The ratings are the author's analytical judgement and are not based on empirical measurement.

Table 6. Analytical comparison of approaches

Approach	Protection of validity	Collaboration with AI	Scalability	Fairness and security
AI detectors	Low	None	High	Low (risk of false signals)
Declaration-only scale	Medium	High	High	Medium
Traditional supervised examination	High	None	Medium	Medium
IZLI	High	High	Medium (high through selective defence)	High (human oversight)

Discussion

The proposed model changes the object of assessment: the question "What did the student submit?" gives way to "Is there evidence of what the student has learned and how?" This is a direct application of the validity logic of Messick (1995) and Kane (2013) and, at the same time, a development, in the AI era, of the idea of evaluative judgement put forward by Tai et al. (2018): the student uses AI, but the capacity to evaluate its output and to take responsibility for it is also assessed.

The practical risks of the model and ways of mitigating them are as follows:

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 06.10.2026
PUBLISHED: 07.10.2026

• **Workload.** A long oral defence for all students may not be feasible. Solution: a 5–7-minute micro-defence for each student; in large groups, random written questions in the learning management system (LMS) and an oral interview only in cases where Δ^+ signals a problem.

• **Equity.** If access to AI tools is unequal, Z-3 and Z-4 assignments may be unfair. Solution: institutional licences or the inclusion of free alternative tools in the list of permitted tools.

• **Privacy.** Prompt logs may contain personal data. Solution: student consent, a limited retention period and access restricted to the instructor.

• **Defence anxiety and language barriers.** Solution: the one-sided Δ^+, alternative formats, and orienting the defence rubric towards assessing content rather than personal expression.

• **Falsification of traces.** A prompt log can be "fabricated" after the fact. Solution: triangulating the traces from three sources rather than one (LMS version history, log, defence).

For HEIs in Uzbekistan, the following sequence is proposed as a practical path: (1) a one-semester pilot in one course and one type of assignment; (2) incorporating the Z-levels, the Δ^+ rule and the appeals procedure into internal regulations (a statute or guidelines); (3) teaching the L and D rubrics in professional development courses for instructors; (4) calibrating the weights on the basis of the pilot results.

Limitations and Future Research

The main limitation of the article is the absence of empirical validation: the weights and threshold values rest on theoretical reasoning and expert judgement. In addition, the model has been developed primarily for text-based and analytical assignments; adaptation is required for laboratory, creative or applied clinical disciplines. The following protocol is proposed for future research:

• **Delphi study (two rounds).** 15–20 experts in assessment and educational technology; agreement on the weights and thresholds to be assessed using Kendall's W coefficient.

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 06.10.2026
PUBLISHED: 07.10.2026

- **Quasi-experiment.** At least two groups (IZLI and traditional assessment), each with approximately 60 students; measures: the correlation between S_P and S_D, the stability of repeated measurements, and the time spent by instructors.
- **Psychometric analysis.** Criteria of ICC(2,k) ≥ 0.75 between raters and Cronbach's α ≥ 0.70 for internal consistency; testing the structure of the components by confirmatory factor analysis.
- **Qualitative study.** Semi-structured interviews with students and instructors on perceptions of fairness, anxiety and workload.

The expected **hypotheses** are as follows: H1 – under IZLI, the correlation between S_P and S_D in Z-3/Z-4 assignments will be higher than in the control group; H2 – the selective-defence strategy will keep instructor workload at an acceptable level; H3 – students will perceive the assessment procedure as sufficiently fair. If the hypotheses are not confirmed, the model will be redesigned.

CONCLUSION

Generative AI poses a direct threat to the validity of product-based assessment; however, the answer lies not in prohibition but in enriching the evidence for assessment. The proposed "IZLI" model protects the interpretation of scores by integrating the process trace, accountability, the capacity for evaluative judgement and interactive defence with the product, while the one-sided authorship congruence coefficient serves as a fair signal that triggers human oversight. The scholarly value of the model lies in its being grounded in the form of a validity argument and in its having testable hypotheses. The next step is pilot testing and psychometric validation. The results will contribute to shaping an assessment policy in HEIs that is suited to the AI era, fair and scientifically grounded.

REFERENCES

1. Decree of the President of the Republic of Uzbekistan No. PF-6079 of 5 October 2020. On approving the "Digital Uzbekistan–2030" Strategy and measures for its effective implementation.

https://creativecommons.org/licenses/by/4.0/

ISSN: 2690-2265
ACCEPTED: 06.10.2026
PUBLISHED: 07.10.2026

2. Resolution of the President of the Republic of Uzbekistan No. 358 of 14 October 2024. On approving the “Strategy for the Development of Artificial Intelligence Technologies until 2030” and measures for its effective implementation.

3. Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347–364.

4. Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

5. Boud, D., Ajjawi, R., Dawson, P., & Tai, J. (Eds.). (2018). *Developing evaluative judgement in higher education.* Routledge.

6. Boud, D., & Falchikov, N. (Eds.). (2007). *Rethinking assessment in higher education.* Routledge.

7. Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325.

8. Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 61*(2), 228–239.

https://creativecommons.org/licenses/by/4.0/