DOI: https://doi.org/10.63056/jllsa.2.10.2026.419

Journal of Language, Literature & Social Affairs

journal homepage: https://scholarclub.org/index.php/jllsa

Who Owns Validity? AI Assessment, Institutional Boundaries, and the Future of Quality Assurance

Dr. Anjum Naz *[a] , Dr. Abida Nasreen[b]

a. PhD, FAHE, FSET, ATS, MSEd, PGCert, PGCE, CharTeach, Qualified Assessor OCR, PPD (Cambridge), Consultant, Commonwealth Scholarship Commission; Head of Teaching & Learning (CSA), AL, Arden University, London, UK; Birmingham Newman University, United Kingdom.
b. Associate Professor, Institute of Education and Research, University of the Punjab, Lahore, Pakistan.

ARTICLE INFO

Received:
September 11, 2026
Revision Received:
October 04, 2026
Accepted:
October 15, 2026
Available Online:
October 26, 2026

Keywords:
Artificial intelligence; assessment validity; higher education; quality assurance; cross-institutional partnerships; academic integrity.

ABSTRACT

Generative artificial intelligence (GenAI) challenges the relationship between assessed student work and the capabilities that universities claim to certify. This conceptual paper examines assessment validity as a shared governance problem, rather than primarily a matter of misconduct detection or technology procurement. It synthesises recent evidence from the UK Quality Assurance Agency, research on automated essay marking, and scholarship on validity-centred and process-informed assessment. The paper proposes a Cross-Institutional AI Assessment Compact comprising co-designed validity indicators, reciprocal audit mechanisms and student-centred accountability. It argues that institutional partnerships can provide a practical setting in which to test assessment redesign, align expectations and protect disciplinary autonomy while developing comparable standards of assurance. The proposed compact remains a framework for pilot evaluation, not an empirically validated intervention.

1. Introduction

Universities award qualifications on the assumption that assessment provides defensible evidence of students' learning and achievement. Generative artificial intelligence (GenAI) complicates that assumption. An essay can now be fluent, well structured and persuasive without its author independently demonstrating every capability apparently represented by the text. Conversely, thoughtful and legitimate use of AI may be an authentic dimension of professional competence. The fundamental question is therefore not simply whether AI was used, but what an assessed performance permits educators to infer about the learner.

This problem is especially important in cross-institutional arrangements, including collaborative provision, validated programmes and delivery partnerships. Different institutions may share responsibility for programme design, teaching, marking, moderation and awards, while maintaining distinct local practices and student support systems. Inconsistency in AI expectations can expose students studying towards comparable awards to materially different assessment conditions. The

* Correspondence to: PhD, FAHE, FSET, ATS, MSEd, PGCert, PGCE, CharTeach, Qualified Assessor OCR, PPD (Cambridge), Consultant, Commonwealth Scholarship Commission; Head of Teaching & Learning (CSA), AL, Arden University, London, UK; Birmingham Newman University, United Kingdom.
E-*mail address:* anaz@staff.newman.ac.uk (A.Naz)

3105-4919 & 3105-4927/© 2026 The Authors. Published by Scholar Club. This is an open access article.

Quality Assurance Agency (QAA, 2026a) identifies variation both between providers and within departments, programmes and modules. Its subsequent work emphasises the need for sector-owned approaches to protecting the reliability of awards (QAA, 2026b).

This paper advances three propositions: assessment integrity should be understood through validity reasoning; process and performance evidence should be triangulated rather than reduced to a single submitted product; and shared governance is required where learning, delivery and certification traverse institutional boundaries. Its contribution is a proposed Cross-Institutional AI Assessment Compact designed for practical piloting.

2. Approach and Conceptual Foundation

This is a conceptual and policy-oriented paper based on a focused synthesis of publicly available research and sector guidance. It does not report interviews, survey results or findings from an implemented institutional pilot. The synthesis centres on QAA's analysis of institutional variability, University of Cambridge evidence about AI marking, Weidlich's framework for validity reasoning, Swiecki and colleagues' account of assessment in the age of AI, UNESCO's human-centred principles and labour-market evidence from the World Economic Forum.

Assessment validity concerns the justification of interpretations and decisions made from assessment evidence. A polished assignment is an observation, not direct access to a student's knowledge, judgement or independence. Weidlich (2026) distinguishes threats such as construct underrepresentation, construct-irrelevant variance, uncertain attribution and unsupported extrapolation. If a task claims to measure independent critical analysis but rewards AI-produced synthesis without requiring student explanation, its evidence-to-claim relationship may be weakened. By contrast, if the intended outcome is the responsible evaluation of AI-assisted outputs, prohibiting those tools may compromise authenticity.

Validity is therefore conditional on the task's purpose, the permitted tools, the evidence available and the intended use of results. Rather than classify all AI involvement as either innovation or cheating, programme teams should specify the capability being certified, identify the likely threat to inference and gather appropriate supporting evidence.

3. Assessment Validity and the Limitations of Automated Judgement

AI has potential to support formative feedback, rubric development and workload management, but the reliability of algorithmic marking remains uneven. A University of Cambridge-led evaluation of more than 750 psychology essays across three UK universities reported that the systems examined matched human-awarded degree classification bands approximately 35-65% of the time (University of Cambridge, 2026; Talmi et al., 2026). The study also described sensitivity to surface presentation and difficulties judging exceptional and weak work. These findings apply to the systems, subject domain and conditions examined; they should not be treated as a universal accuracy rate for all AI marking.

The implication is not that human marking is infallible. Human assessment also requires moderation, calibration and scrutiny of bias. Rather, replacing expert judgement with an opaque system may add a further source of inconsistency without resolving disputes over what counts as quality. An algorithm can consistently apply an invalid proxy just as a human marker can inconsistently apply a sound standard.

An assessment policy must accordingly distinguish AI used by students to generate evidence from AI used by institutions to interpret evidence. The former raises attribution, authenticity and construct questions; the latter raises reliability, transparency, accountability and rights of appeal. Both should be governed through explicit evidence requirements and meaningful human oversight (Soni, A., Kumar, M., Kumar, S., Nawaz, I., Mohan, J. M., & Janardhan, R. 2025).

4. Institutional Boundaries as a Quality Assurance Risk

QAA's report, *The Perfect Storm: AI, Assessment and a Sector Under Pressure*, finds that variability in student use, staff confidence, policy and local implementation can produce inconsistent learning experiences (QAA, 2026a). A single university may permit AI for planning in one module, require detailed disclosure in another and prohibit related activities elsewhere. Differences can be justified by disciplinary aims; unexplained contradictions, however, undermine students' ability to understand and meet expectations.

Partnership delivery magnifies the issue. Awarding bodies may own regulations while teaching partners control routine assessment communication. External examiners may see final scripts but not the conditions under which they were produced. Central policy teams may mandate disclosure without checking whether students can realistically distinguish editing assistance from substantive generation. Fragmented governance produces a risk of responsibility without control, or control without accountability.

The answer should not be a universal list of forbidden tools. Programmes differ in what they legitimately seek to assess. Instead, institutions need shared principles for explaining AI conditions, mapping outcomes to evidence, recording exceptions, moderating consequential decisions and reviewing student impact. Such principles make local flexibility more defensible, not less.

5. From Product-Based to Evidence-Rich Assessment

Conventional assessments often privilege a final artefact: the essay, report, presentation or examination script. Swiecki et al. (2022) caution that isolated products offer limited visibility into learning processes and explore the opportunities and risks of richer, technology-supported assessment evidence. In AI-inclusive settings, an evidence-rich approach might combine a submitted product with a concise design rationale, annotated sources, selected drafts, reflective commentary or a short oral discussion.

For example, a management student could submit an AI-assisted strategic recommendation while also identifying the prompts or information sources that materially shaped it, explaining which suggestions were rejected and defending the final judgement in a brief viva. The aim is not to collect exhaustive surveillance logs. It is to obtain proportionate evidence that connects the student to the assessed learning outcomes. In some contexts, supervised application tasks or demonstrations may provide stronger evidence; in others, authentic collaboration with tools is itself the target capability.

Process evidence also has limitations. Reflections can be scripted, draft histories can be manufactured, and oral assessment can disadvantage some students if accessibility is neglected. Asking every learner for multiple extra submissions may increase workload. Evidence should therefore be selected according to stakes, accessibility, feasibility and the nature of the inference being made. High-stakes certification may justify stronger triangulation than low-stakes formative activity.

6. A Cross-Institutional AI Assessment Compact

The proposed Cross-Institutional AI Assessment Compact is a structured agreement among awarding institutions, delivery partners, programme teams and student representatives. It is intended as a mechanism for shared assurance rather than a new external regulator. Three mutually reinforcing commitments define its operation.

6.1 Co-Designed Validity Indicators

First, partners would co-design validity indicators. Each programme would state what its major assessments claim to measure, the conditions of permissible AI use, the evidence that links individual students to those claims and the implications for access and fairness.

Indicators might include clear task-level AI guidance, outcome–evidence alignment, student understanding of rules, moderation consistency, and documented review of alternative assessment formats. Shared definitions would enable comparison while preserving subject-specific criteria.

6.2 Reciprocal Audit Mechanisms

Second, the compact would establish reciprocal audit mechanisms. Partners could conduct annual sample-based reviews of assessment briefs, anonymised student work, moderation records and student-facing guidance.

Review teams would test whether assessment claims are sufficiently supported and whether exceptions are handled consistently. Calibration workshops could compare judgements on common exemplars and record the rationale for differences. Findings should feed into programme enhancement rather than default to punitive compliance reporting.

6.3 Mutual Accountability for Student Experience

Third, the compact would establish mutual accountability for the student experience. Students would receive coherent information about acceptable AI use, disclosure, feedback and appeals regardless of delivery site.

A joint oversight group would include academic leads, quality professionals, accessibility specialists and students. It would publish concise annual improvement actions and assign named owners, deadlines and routes for escalating unresolved disagreements. Commercial tool providers should never become the final arbiters of academic standards.

These elements are interdependent. Indicators without review are aspirational; audits without student participation risk becoming procedural exercises; and consultation without assigned responsibility rarely changes practice. The compact attempts to connect all three.

7. Implementation, Evaluation and Safeguards

A feasible pilot could begin with two or three partner programmes over one academic year. During the first term, a joint team would map learning outcomes, assessment formats, AI permissions and current moderation practices. It would identify a small number of consequential assessments for redesign. During the second term, teams would introduce proportionate process evidence, clearer AI conditions and shared marker calibration. At year-end, partners would review results and decide whether to revise or extend the model.

Evaluation should go beyond counting misconduct referrals. Suggested measures include the proportion of assessments with explicit validity rationales; inter-marker agreement on sampled work; staff and student comprehension of AI guidance; appeal and complaint patterns; perceived fairness across delivery sites; time required for assessment and moderation; and accessibility-related adjustments.

Evidence could be gathered through document audits, small-scale surveys, focus groups and anonymised moderation records. Baselines and differences between programmes would need to be documented before claims of improvement could be made.

Risks warrant active mitigation. Process capture can create privacy and data-protection concerns, particularly where third-party tools retain prompts or student material. Unequal access to paid AI services can introduce construct-irrelevant advantage. Students with disabilities or different linguistic backgrounds may be disproportionately affected by rigid oral or supervised formats.

Staff workload and academic freedom also require consideration. An effective compact would adopt data minimisation, transparent retention rules, accessible alternative evidence routes, documented human review and periodic equality-impact checks.

Any pilot should explicitly test whether the proposed controls improve confidence in what grades mean, rather than merely adding bureaucracy. Negative or mixed findings would be informative and should shape revision.

8. Discussion: Who Ultimately Owns Validity?

Validity cannot be delegated exclusively to a central quality office, an individual marker, an AI vendor or a partner institution. Awarding bodies retain formal responsibility for academic standards, but defensible evidence is produced through distributed decisions about curriculum, teaching, assessment design, marking and student support.

Ownership should therefore be understood as layered: academic teams justify their assessment claims; programme leaders ensure coherence; quality systems test those claims; and institutional partnerships address gaps at organisational interfaces.

UNESCO's 2026 ministerial statement stresses that AI adoption must protect human relationships, judgement and public accountability (UNESCO, 2026). This reinforces the need for human-governed processes rather than automatic reliance on either detection or algorithmic grades.

The World Economic Forum's *Future of Jobs Report 2025* likewise identifies analytical thinking, adaptability and creative thinking as prominent capabilities valued by employers (WEF, 2025). Higher education should assess such capabilities in credible contexts, while avoiding the unsubstantiated assumption that employer demand renders degrees obsolete.

The compact's principal value is thus institutional learning: partner organisations can make the reasoning behind assessment decisions visible, examine variation and accumulate evidence about which practices work.

Shared assurance does not imply identical assessments. It implies comparable transparency about what each assessment certifies and why that interpretation is justified.

9. Conclusion

GenAI has exposed a longstanding vulnerability in higher education: the distance between an assessment product and the claims institutions make from it. The response should move beyond detecting unauthorised tool use towards designing, moderating and reviewing credible evidence of learning.

Recent QAA analysis demonstrates that inconsistent practice is already a governance concern, while emerging validity scholarship offers a practical language for diagnosing specific threats.

This paper proposes a Cross-Institutional AI Assessment Compact founded on co-designed validity indicators, reciprocal audits and shared responsibility for student experience. As a conceptual proposal, its effectiveness remains to be tested. Nevertheless, it offers a route for awarding institutions and partners to reconcile local disciplinary judgement with collective assurance.

The future credibility of qualifications will depend less on proving that AI was absent and more on demonstrating clearly what students know, can do and can responsibly justify.

10.References

QAA (2026a) *The perfect storm: AI, assessment and a sector under pressure.* Quality Assurance Agency for Higher Education. Available at: https://www.qaa.ac.uk/news-events/policy-and-leadership/state-of-the-nation

QAA (2026b) *Assuring the standard of UK awards in the age of GenAI.* 16 September. Available at: https://www.qaa.ac.uk/news-events/news/assuring-the-standard-of-uk-awards-in-the-age-of-genai

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J.M., Milligan, S., Selwyn, N. and Gašević, D. (2022) 'Assessment in the age of artificial intelligence', *Computers and Education: Artificial Intelligence*, 3, 100075. https://doi.org/10.1016/j.caeai.2022.100075

Talmi, D., Benn, Y., Tibon, R., Corsi, G. and Marcoci, A. (2026) *AI in university assessment: Evaluating the opportunities and risks of automated marking.* University of Cambridge. Available at: https://www.ai.cam.ac.uk/reports/ai-in-university-assessment-evaluating-the-opportunities-and-risks-of-automated-marking/

UNESCO (2026) *Education ministers call for education to remain a common good in the age of AI.* 9 September. Available at: https://www.unesco.org/en/articles/education-ministers-call-education-remain-common-good-age-ai-unescos-digital-learning-week

University of Cambridge (2026) *AI not yet good enough to mark university essays, rewarding "style over substance".* 22 May. Available at: https://www.cam.ac.uk/stories/ai-university-essay-grading

Weidlich, J. (2026) 'Which inference is at risk? Assessment validity reasoning and generative AI', *Assessment & Evaluation in Higher Education*, published online 20 September. https://doi.org/10.1080/02602938.2026.2734795

World Economic Forum (WEF) (2025) *The future of jobs report 2025.* Available at: https://www.weforum.org/publications/the-future-of-jobs-report-2025/

Soni, A., Kumar, M., Kumar, S., Nawaz, I., Mohan, J. M., & Janardhan, R. (2025). The Role of Artificial Intelligence in Transforming Human Resource Management: Opportunities and Challenges in Ethical and Social Issues of Digitalization. *Advances in Consumer Research*, *2*, 1084-1097.