REVIEW OF CRIME, PEACE AND SOCIETY
ISSN (O): 3079-6105, ISSN (P): 3079-6091
Vol. 3, No. 8, 2026, 161-174

AI in Evaluating Teacher Education Portfolios: A Validity-Centred Framework

Dr. Abida Nasreen[1], Dr.Anjum Naz[2], Dr. Asma Sharif *[3]

Abstract

Professional portfolios act as high-stakes gatekeeping instruments in teacher education, yet their assessment is vulnerable to rater effects, thin feedback, limited scalability and cultural and gender bias. Artificial intelligence (AI) is increasingly proposed as a remedy. This conceptual article argues that AI can enhance portfolio evaluation only when its introduction is governed by a validated framework that places construct validity, fairness, transparency and accountability ahead of efficiency. Using a critical, argument-based analysis of the assessment, validity and AI-in-education literature, the article (a) diagnoses the problems that create demand for AI, (b) analyses the potential and the validity, ethical and practical challenges of AI-supported evaluation, and (c) reviews validated design responses, including structural redesign, triangulation, assessor AI literacy, ethical governance, humanAI collaboration, comparative judgement and aberrant-response detection. It then proposes a six-indicator evaluation framework (validity, fairness, transparency, accountability, scalability and ethical use) that synthesises Messick's unified validity theory, Kane's argument-based validation, Weidlich's Validity Reasoning Matrix and Quality Assurance Agency (QAA) principles. The article concludes that the central question is not whether AI will feature in portfolio assessment but whether it will be used with the critical scrutiny that professional certification demands, and it sets out implications for practice, policy, leadership and research.

Article History
Received 22 September 2026
Accepted 14 October 2026
Published 20 October 2026
Corresponding Author*
OPEN ACCESS
Keywords
Artificial intelligence; Professional portfolio; Teacher education; Assessment validity; Comparative judgement; Fairness; assessment Governance

1. Introduction

The professional portfolio is a mandatory component of many teacher training programmes and operates as a high-stakes gatekeeping instrument. It provides evidence that a prospective teacher has met the professional standards required for certification or employment, and typically comprises lesson plans, reflective commentaries, classroom artefacts, observations of teaching and records of professional development. Because it is intended to capture situated professional competence rather than declarative knowledge alone, the portfolio is among the most complex forms of assessment in teacher education. Assessors must make defensible judgements about practice from diverse and often incomparable evidence, at scale, within tight time limits, and under conditions long recognised as vulnerable to rater effects.

Artificial intelligence (AI) has been proposed as a way of addressing these difficulties. Automated scoring, natural language processing, learning analytics and multimodal data analysis appear to offer greater efficiency, consistency and scalability than conventional

[1] Associate Professor, Institute of Education and Research, University of the Punjab, Lahore.
Email: zoraiztahir55@gmail.com
[2] 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.
[3] Visiting Lecturer, Institute of Education and Research, University of the Punjab, Lahore.

human marking. Yet integrating AI into portfolio evaluation is not a straightforward technical upgrade. It raises basic questions about what the portfolio is meant to measure, whether AI can validly capture that construct, and whether the consequences of algorithmic judgement can be defended to candidates, programmes and the profession.

The issue has become pressing for two reasons. First, cohort sizes in many programmes have grown while assessment resources have remained static, which strains both consistency and the quality of feedback. In the programme context that prompted this article, for example, candidate numbers rose by roughly 40% over five years while allocated assessment hours did not change. Second, generative AI makes it easy for candidates to produce polished reflective writing that may not reflect their own professional reasoning, which threatens the inference from submitted evidence to demonstrated competence. The question is therefore no longer whether AI will feature in portfolio assessment, but how it can be used responsibly.

1.1 Aim and research questions

The article's central argument is that AI can enhance the evaluation of teacher education portfolios, but only if its introduction is governed by a validated framework that prioritises construct validity, fairness, transparency and accountability. Three questions structure the analysis:

1. **RQ1.** What problems in current portfolio evaluation create the case for AI?
2. **RQ2.** What potential and what validity, ethical and practical challenges does AI introduce?
3. **RQ3.** Which validated design responses, and which evaluation framework, can support responsible integration?

1.2 Contribution and structure

The article contributes (a) a validity-centred analysis of AI in a specific high-stakes setting, teacher certification portfolios, which is less examined than AI in examinations or essay marking; and (b) a six-indicator framework that translates abstract validity theory into questions that programme teams can use formatively and summatively. Section 2 outlines the conceptual approach. Sections 3 to 5 address RQ1 and RQ2. Section 6 reviews design responses and Section 7 presents the framework (RQ3). Section 8 discusses implications, Section 9 notes limitations and future research, and the article closes with a brief conclusion.

2. Conceptual approach

This is a conceptual, argument-based article rather than an empirical study. It does not report new data on AI scoring performance. Its method is a critical analysis of peer-reviewed and sector literature on assessment validity, AI in assessment, comparative judgement and academic integrity, supplemented by regulator and quality-assurance guidance (Ofqual, 2020; QAA, 2026). Sources were selected for their relevance to high-stakes, performance-based

assessment and for their capacity to inform design choices, not through a systematic search, so the review should be read as purposive rather than exhaustive.

Four validity lenses structure the analysis:

- **Messick (1995)** treats validity as a property of the interpretations and uses of scores, requiring attention to construct underrepresentation, construct-irrelevant variance and the social consequences of assessment.
- **Kane (2013)** operationalises validation as an interpretation and use argument, a chain of inferences (scoring, generalisation, explanation, extrapolation) each of which must be open to scrutiny.
- **Weidlich (2025)** offers a Validity Reasoning Matrix for the generative-AI era, linking concerns to the inference at risk, a proportionate redesign and the local evidence needed.
- **QAA (2026)** supplies a sector-level perspective on assessment validity, academic integrity, and staff and student support.

The lenses are applied in two ways: to evaluate the claims made for and to derive the framework proposed in Section 7.

3. Problems in evaluating teacher professional portfolios

A portfolio is not a single test but a body of evidence assembled over time, so its assessment depends on human judgement at every stage: selecting evidence, interpreting its significance and deciding whether it demonstrates the required standards. Four well-documented problems follow.

3.1 Subjectivity and rater bias

Research on portfolio assessment consistently identifies severity, leniency and halo effects as sources of score variation. A severe assessor may downgrade an otherwise adequate portfolio, while a lenient one may pass thin evidence. Halo effects arise when a positive impression in one area, such as an engaging reflective style, unduly shapes judgements in others, such as subject knowledge or classroom management. Because portfolios consist largely of self-documentation, assessors often cannot ask clarifying questions, which increases interpretive uncertainty. Two candidates of comparable competence may therefore receive different outcomes depending on who assesses them, which undermines the credibility of certification and can have serious consequences for careers.

3.2 Limited feedback quality

Portfolio assessment is resource-intensive, and programmes under pressure tend to prioritise summative decisions over formative feedback. Candidates may receive a grade or pass/fail outcome with little explanation of how their practice could improve. This is a particular loss in teacher education, where the portfolio is meant both to certify competence and to support

professional learning. Assessment literacy, for assessors and candidates alike, suffers when the process is opaque and hurried (Bennett, 2018); candidates who cannot understand why evidence was judged inadequate are unlikely to improve.

3.3 Scalability

Large cohorts demand substantial assessor time, and scaling human assessment without loss of quality is difficult. Reliance on part-time or less experienced assessors can exacerbate inconsistency, particularly where they do not share core staff's understanding of the professional standards. The problem is more than administrative: it bears on the reliability and credibility of the certification process.

3.4 Cultural and gender bias

Portfolio evidence is not culturally neutral. Conventions of reflective writing, expectations about self-presentation and assumptions about what counts as "professional" behaviour may disadvantage candidates from particular cultural backgrounds or genders. Assessors may unconsciously reward reflective styles that align with dominant norms, or interpret confidence differently by gender. Such biases are hard to remove through rubric design alone because they operate at the level of interpretation. As Weidlich (2025) notes, threats to validity are often embedded in the conditions under which evidence is produced, not only in the scoring instrument.

Together these problems make a plausible case for AI: if it could score more consistently, provide richer feedback and do so at scale, it might address several weaknesses at once. Whether available technologies can deliver those benefits without creating new and possibly more serious problems is the subject of the next two sections.

4. The potential of AI in portfolio evaluation

The potential is real and varied. Automation could reduce the administrative burden of scoring, freeing assessors for higher-order judgement. Scalability could make detailed feedback feasible for every candidate rather than a subset. Data-driven insights could reveal cohort-level patterns, such as recurring weaknesses in reflective writing or gaps in evidence, that programme leaders could use to improve training. Personalised feedback could point individual candidates to specific areas for development.

These possibilities are not merely speculative. Automated scoring is already used in high-stakes language assessment, and Gao, Gales and Xu (2024) describe classifiers that detect aberrant responses in automated spoken-English marking, such as incomprehensible or nonsensical answers, and so improve score accuracy. Analogous techniques could be applied to portfolio evidence: natural language processing could analyse reflective writing, computer vision could examine classroom artefacts, and learning analytics could track engagement across a programme. Zainuddin, Shujahat, Haruna and Chu (2020) show that gamified e-

quizzes can raise engagement and provide immediate feedback in formative assessment; similar principles could inform timely, motivating AI-supported portfolio feedback.

Bennett (2018) argues that assessment is being reshaped by technological and social change and is likely to become more personalised, more embedded in learning and more dependent on complex evidence. In principle, AI could make portfolio assessment more responsive to the richness of professional practice, for instance through multimodal learning analytics that combine video of teaching with written reflection. AI could also support comparative judgement by organising and presenting evidence to human judges, reducing the cognitive load of sorting large numbers of portfolios.

5. Challenges of AI-supported evaluation

The challenges are substantial and go to the heart of what portfolio assessment is for.

5.1 Validity threats

Messick's (1995) unified conception of validity directs attention to construct underrepresentation and construct-irrelevant variance. AI systems may underrepresent the construct of professional competence if they focus on features that are easy to quantify, such as word frequency, syntactic complexity or sentiment, rather than the pedagogical reasoning and situated judgement the portfolio is meant to evidence. They may also introduce construct-irrelevant variance if scores are influenced by factors unrelated to competence, such as the quality of a candidate's camera or the dialect features of their written English.

Weidlich (2025) distinguishes several problems that are often collapsed in debate: construct underrepresentation, construct-irrelevant variance, attribution of performance, conditional extrapolation and harmful use. Under generative-AI conditions, the central concern is not simply whether AI-generated text can be detected, but whether the evidence produced can support the intended interpretation. If a candidate uses AI to draft reflective commentary, the portfolio may no longer evidence that candidate's own reflective capacity, and the inference from performance to competence is threatened. Asking which inference is at risk and responding proportionately is more sophisticated than banning AI or relying on detection software.

5.2 Construct substitution

When AI is introduced, there is a temptation to redefine the construct in terms of what the AI can measure. A portfolio intended to capture professional judgement may become a test of writing style or digital literacy. The assessment may become more reliable without becoming more valid: a system trained to identify "high-quality reflection" may reward linguistic fluency rather than depth of pedagogical insight, systematically disadvantaging less confident writers and those writing in a second language.

5.3 Bias

AI systems are trained on data that may encode cultural and gender bias. If historical portfolio scores reflect bias, a model trained on them may reproduce and amplify it. There is also evidence that AI-text detectors misclassify non-native English writing as AI-generated, a specific risk for programmes that recruit internationally, where candidates could be penalised for linguistic features beyond their control. Ofqual (2020) identified unequal access to reliable technology as a major barrier to online assessment, and similar equity concerns apply here.

5.4 Transparency and accountability

Many AI systems are opaque: candidates and assessors cannot see how a score was produced, which makes judgements hard to challenge and undermines legitimacy. If a candidate fails because an algorithm classified their portfolio as inadequate, what recourse do they have? Lack of explainability is a threat to procedural fairness, not a technical inconvenience. Kane (2013) argues that validation requires a clear interpretation and use argument in which each inference is open to scrutiny, and proprietary systems often make such scrutiny impossible.

5.5 Ethical concerns

Portfolio evidence may include sensitive personal data, classroom recordings and information about pupils, raising questions of data protection, surveillance and consent. Corbin, Sharpe and Dawson (2026) show that wearable AI, such as smart glasses that can display generated text and process speech without any external signal, erodes the conditions on which physical exclusion of AI from assessment has depended. Their concept of "dual transparency" captures how AI can become invisible to both wearer and observer. Where portfolio evidence is produced away from direct supervision, this is a serious challenge. The authors also warn against a "regime of bodily adjudication" that may fall hardest on students with disabilities, health conditions and religious dress practices. A just system must not sacrifice equity for technological control.

Corbin, Dawson, Nicola-Richmond and Partridge (2025) offer a framework for acceptable uses of AI in assessment, which is useful for deciding where to draw the line. If assessors use AI to organise evidence or draft feedback, this may be acceptable; if candidates use it to generate reflective commentary, validity may be undermined. The difficulty is that the same technology serves both purposes and detection is unreliable.

5.6 Unreliable AI grading and detection

AI detectors produce false positives and false negatives, and automated scoring can be gamed. Using AI to grade portfolios may create an illusion of objectivity while introducing new forms of error. Cotton, Cotton and Shipway (2023) argue that academic integrity in the age of generative AI requires rethinking assessment design rather than relying on detection. The same logic applies here: detection-based approaches are reactive and adversarial, whereas design-based approaches are proactive and educational.

5.7 Practical and infrastructural barriers

Ofqual's (2020) review found that IT provision, security and staffing are significant obstacles to online assessment. Programmes vary widely in digital capacity, and a system that works in a well-resourced institution may be unworkable elsewhere. In the author's own institution, for example, the learning management system is not fully integrated with assessment tools and data protection protocols were not designed for AI processing. Scalability is therefore a matter of institutional readiness as well as processing power.

6. Validated ways forward

The challenges are significant but not insurmountable. A range of design and governance responses can mitigate the risks while capturing some benefits of AI.

6.1 Structural redesign

Rather than bolting AI onto existing portfolio tasks, programmes should redesign the portfolio to generate evidence that is more resistant to AI-related threats: a longitudinal record of practice in place of a one-off submission, multimodal evidence (video, audio, pupil work, observer notes), and a requirement that candidates justify professional decisions orally or in dialogue. Perkins (2026) makes a case for verbal examinations as a resilient form of assessment in the age of AI, drawing on Wiggins's theory of authentic assessment. Verbal assessment lets assessors probe reasoning and distinguish genuine understanding from surface performance. A short viva voce attached to the portfolio would not remove the need for AI governance, but would reduce reliance on written artefacts that can be easily generated.

6.2 Triangulation

No single source of evidence should determine a high-stakes judgement. Consistent with Messick's (1995) emphasis on multiple lines of evidence, AI-generated scores should be combined with human assessment and weighed against other evidence. If an AI system flags a portfolio as weak, a human assessor should review it independently, and divergences between the two judgements should be investigated. Triangulation also means drawing on varied sources: written reflection, observation, pupil outcomes and professional dialogue. It is more labour-intensive than a single AI score, but more defensible.

6.3 AI literacy for evaluators

Assessors need to understand what AI systems can and cannot do, how they produce outputs and where their limits lie. This is a matter of professional judgement as much as technical training: an assessor who understands the logic of automated scoring is better placed to decide when to rely on it and when to override it. QAA (2026) stresses supporting staff to use generative AI responsibly and to understand its implications for assessment validity. In practice this implies continuing professional development with hands-on experience of AI tools, critical discussion of bias and fairness, and case studies of AI-related misconduct. Assessment literacy, as Bennett (2018) argues, is less a fixed body of knowledge than a capacity to judge well in changing conditions.

6.4 Ethical governance

Programmes need clear policies on data protection, consent, transparency and the right to challenge AI-assisted decisions, supported by external oversight, regular bias audits and appeal mechanisms. QAA's (2026) work offers a starting point for sector guidance. Governance should also heed Corbin et al.'s (2026) warning that enforcing physical exclusion of AI may fall hardest on students with disabilities, health conditions and religious dress practices. In practice, any AI system used in portfolio assessment should be subject to a data protection impact assessment, and candidates should be told in advance how their data will be used.

6.5 HumanAI collaboration

The most defensible model is not full automation but a division of labour in which AI handles routine tasks (flagging anomalies, drafting feedback, organising evidence) while humans retain responsibility for high-stakes judgement. This preserves professional accountability while capturing efficiency gains, and is consistent with Swiecki et al.'s (2022) argument that assessment in the age of AI should centre on what humans do best: judging complex performances, giving nuanced feedback and making contextual decisions.

6.6 Evidence-centred design

Evidence-centred design provides a principled way to integrate these measures. The guiding question is not "What can AI do?" but "What evidence do we need to support the intended inference?" Once evidence requirements are specified, AI can be selected and configured accordingly. This aligns with Kane's (2013) argument-based validation: programmes specify what counts as evidence of professional competence, how it will be elicited and what interpretations are warranted, and introduce AI only where it can be shown to support those inferences.

6.7 Comparative judgement

Comparative judgement offers a validated alternative to rubric-based scoring that may suit portfolio evidence. Wheadon, Barmby, Christodoulou and Henderson (2020) report a large-scale study of primary writing in England in which teachers compared pairs of scripts rather than applying rubrics; the approach, which used anchoring to stop teachers inflating their own pupils' scores, showed promise as a fair and robust large-scale method. AI could organise and present evidence while human judges make holistic comparisons, combining technological scalability with human validity. It is not a panacea: it requires trained judges, careful sampling of evidence and enough comparisons for reliable scores, and may suit some portfolio components better than others.

6.8 Aberrant-response detection

Gao et al. (2024) show that classifiers can detect incomprehensible or nonsensical responses in automated spoken assessment and improve marking accuracy. Similar techniques could flag portfolio evidence that is inconsistent, incomplete or potentially fabricated, directing human attention where it is most needed. This is a more defensible use of AI than automated scoring because it supports human judgement rather than replacing it. It is, however, a quality-assurance tool and not a solution to AI-generated text.

6.9 Authentic assessment

Swiecki et al. (2022) argue that traditional assessment is poorly suited to the age of AI and call for rethinking what is assessed and how. Portfolios are already more authentic than many examinations, but can be made more so by embedding tasks in real professional contexts and requiring candidates to show reasoning in action. Shaw and Child's (2017) critique of PISA's collaborative problem-solving tasks shows that technology-mediated assessment of complex constructs is possible but demands careful attention to task design and construct validity: technology does not make assessment authentic by itself. In portfolio assessment, authenticity can be strengthened by requiring evidence from candidates' own classrooms, accompanied by pupil work and observer notes, and by assessing how they adapt practice in response to feedback.

6.10 Summary of responses

Response	Main threat addressed	Principal limitation
Structural redesign (longitudinal, multimodal, viva)	Attribution of performance under AI conditions	Assessor time; access and accommodation needs
Triangulation	Over-reliance on any single score or source	More labour-intensive than one score
AI literacy for evaluators	Misplaced trust in or rejection of AI outputs	Requires sustained CPD investment
Ethical governance	Opacity, data-protection and equity risks	Needs external oversight and audit capacity
HumanAI collaboration	Loss of professional accountability	Division of labour must be explicit and monitored
Comparative judgement	Rubric-related rater effects	Trained judges and many comparisons needed
Aberrant-response detection	Incomplete or fabricated evidence	A quality-assurance tool, not an AI-text detector

7. A framework for evaluating AI in teacher education portfolios

The responses in Section 6 need a coherent structure. The framework proposed here synthesises four sources.

7.1 Theoretical foundations

Messick's validity theory supplies the foundation. Validity is a property of the interpretations and uses of scores, not of a test, and validation must consider evidential and consequential aspects (Messick, 1995). For AI-based portfolio evaluation, it is not enough to show that scores are consistent; one must ask whether they support fair and defensible certification decisions, whose effects on careers and on the quality of education are long-lasting.

Kane's argument-based approach operationalises this concern. Validation specifies the interpretation and use argument, the chain of inferences from performance to score to interpretation to use, and tests the assumptions behind each link. Kane (2013) identifies scoring, generalisation, explanation and extrapolation inferences, each of which AI can threaten and each of which requires local evidence:

Inference	What it requires for AI-supported portfolio scoring	Example evidence
Scoring	Scores rest on relevant features of the portfolio	System is sensitive to pedagogical content, not only linguistic features
Generalisation	Scores are stable across tasks and occasions	Agreement across portfolio components and assessment occasions
Explanation	Scores reflect professional competence, not irrelevant variance	Scores not confounded by cultural background, gender or first language
Extrapolation	Scores relate to performance in other contexts	Later relationship with classroom practice or employer judgements

Weidlich's Validity Reasoning Matrix provides a practical diagnostic. It links concerns to threatened inferences, redesign options and evidence needs, and proposes a sequence that begins with specifying the claim and the AI conditions, proceeds through threat diagnosis and proportionate design response, and ends with local evidence (Weidlich, 2025). It prompts teams to ask: What claim are we making about this candidate? Which inference is at risk? What redesign would address the risk? What evidence would show it has worked? The matrix is a heuristic for professional judgement, not a checklist.

QAA principles add a sector-level perspective on standards and accountability. QAA (2026) emphasises assessment validity, academic integrity and support for staff and students in using AI responsibly, and its work reinforces the need for transparency, consistency and equity, as well as institutional governance and staff development.

7.2 The six indicators

Synthesising these sources yields six indicators, each framed as a question that a programme team can answer with evidence.

Indicator	Key question

Validity	Does the AI-supported assessment measure the intended construct of professional competence, and are the inferences from evidence to judgement defensible?
Fairness	Are all candidates, regardless of cultural background, gender, disability or access to technology, given an equal opportunity to demonstrate competence?
Transparency	Can candidates and assessors understand how AI-assisted judgements are produced, and are the criteria for judgement clear?
Accountability	Are there clear lines of responsibility for AI-assisted decisions, and are there mechanisms for challenge and appeal?
Scalability	Can the assessment be delivered reliably and sustainably across programmes and cohorts?
Ethical use	Are data protection, consent, privacy and the avoidance of harm built into the design and governance of the system?

The indicators interact. A highly scalable but opaque system may fail the transparency and accountability tests, and a transparent but biased system fails on fairness. The framework therefore supports holistic judgement rather than a tick-box exercise. It can be used formatively while designing an AI-supported assessment and summatively to evaluate an existing one.

8. Discussion and implications

The analysis supports a qualified answer to the central question. The current problems in portfolio evaluation (subjectivity, limited feedback, scalability constraints and bias) are real, and AI offers plausible remedies. The challenges are equally real: construct underrepresentation, construct substitution, algorithmic bias, opacity, ethical risk and infrastructural barriers. The way forward is neither to reject AI nor to embrace it uncritically, but to integrate it within a defensible assessment design in which humans remain accountable for high-stakes judgement.

Two points deserve emphasis. First, the strongest responses are about design rather than detection. Redesigning portfolios to yield richer, more AI-resistant evidence, triangulating sources and adding dialogic components addresses the validity threat at its source, whereas detection and bodily exclusion are reactive, error-prone and inequitable. Second, the most defensible uses of AI are those that support human judgement, such as organising evidence, flagging anomalies, drafting feedback and assisting comparative judgement, not those that replace it. In both cases the six indicators give programme teams a way to test whether a proposed use of AI preserves the intended construct and treats candidates fairly.

8.1 Implications for practice

1. **Pilot before embedding.** Trial AI tools on a small scale before using them in high-stakes decisions, and evaluate them against the six indicators.

2. **Invest in assessors.** Train assessors so that human judgement remains the anchor of the system.
3. **Set clear policies.** Develop explicit policies on data protection, consent and the right of appeal.
4. **Diversify evidence.** Use multimodal evidence and verbal-defence components to reduce reliance on any single source.
5. **Monitor for bias.** Conduct regular audits for differential outcomes across candidate groups.
6. **Communicate openly.** Explain to candidates and staff why AI is used and what safeguards apply.

8.2 Implications for leadership

Leadership for change should be distributed. Programme leaders, assessors, IT and data-protection staff, and candidates all have a role in shaping the system, and institutional readiness (integration of learning management and assessment tools, data-protection protocols, staff capacity) should be assessed before adoption.

8.3 Implications for policy

Policy-makers should develop sector-level guidance on AI in teacher certification, fund research on the validity and fairness of AI-supported portfolio assessment, and require transparency from vendors of AI assessment tools. They should resist mandating AI for efficiency reasons without adequate evidence of validity and fairness. QAA (2026) provides a foundation, but it needs to be translated into guidance specific to teacher education.

9. Limitations and future research

The article has three main limitations. It is conceptual and presents no empirical test of AI scoring or feedback on teacher portfolios, so the framework remains to be validated in practice. Its sources were purposively selected rather than systematically searched, and several recent sources are preprints or very recent publications whose findings may be refined. Finally, the illustrative institutional details come from a single postgraduate teacher training context and may not generalise.

Future research should therefore:

- conduct longitudinal studies of whether AI-assisted portfolio evaluation predicts later teacher effectiveness;
- compare AI configurations to establish whether they produce different patterns of bias, including for candidates writing in a second language;
- examine candidates' and assessors' perceptions of fairness and trust;
- assess cost-effectiveness across institutional contexts; and
- use participatory designs that involve candidates and teachers, so that AI design is informed by those most affected.

10. Conclusion

Teacher certification decisions shape who enters the profession and what kind of education pupils receive. This article has argued that AI can make portfolio-based decisions more consistent and better informed, but only where it is used with a clear understanding of what the portfolio is for and what it can and cannot evidence. By combining Messick's consequential view of validity, Kane's inference-by-inference scrutiny, Weidlich's reasoning about which inference is at risk and QAA's sector principles, the proposed six-indicator framework offers programme teams a structure for that understanding. In a rapidly changing assessment landscape, the greatest risk is not that AI will be used, but that it will be used without the critical scrutiny that professional judgement demands.

References

Bennett, R. E. (2018). Educational assessment: What to watch in a rapidly changing world. *Educational Measurement: Issues and Practice, 37*(4), 715.

Corbin, T., Dawson, P., Nicola-Richmond, K., & Partridge, H. (2025). 'Where's the line? It's an absurd line': Towards a framework for acceptable uses of AI in assessment. *Assessment & Evaluation in Higher Education, 50*(5), 705717.

Corbin, T., Sharpe, S., & Dawson, P. (2026). On AI glasses and wearable AI in assessment. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2026.2661367

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, 61*(2), 228239. https://doi.org/10.1080/14703297.2023.2190148

Gao, S., Gales, M., & Xu, J. (2024). Detecting aberrant responses in automated L2 spoken English assessment. In C. A. Chapelle, G. H. Beckett, & J. Ranalli (Eds.), *Exploring artificial intelligence in applied linguistics* (pp. 96117). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.07

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741749.

Ofqual. (2020). *Online and on-screen assessment in high-stakes, sessional qualifications: A review of the barriers to greater adoption and how these may be overcome*. Ofqual.

Perkins, G. (2026). Resilient assessment in the age of AI: Authentic design and the case for verbal examinations in business education. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2026.2644516

QAA. (2023). *Reconsidering assessment for the ChatGPT era: QAA advice on developing sustainable assessment strategies*. Quality Assurance Agency for Higher Education.

Shaw, S. D., & Child, S. F. J. (2017). Utilising technology in the assessment of collaboration: A critique of PISA's collaborative problem-solving tasks. *Research Matters: A Cambridge Assessment Publication, 24*, 17-22.

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

Weidlich, J. (2026). Which inference is at risk? Assessment validity reasoning and generative AI. *Preprint*. https://osf.io/preprints/socarxiv/f6hz7_v2

Wheadon, C., Barmby, P., Christodoulou, D., & Henderson, B. (2020). A comparative judgement approach to the large-scale assessment of primary writing in England. *Assessment in Education: Principles, Policy & Practice, 27*(1), 46-64. https://doi.org/10.1080/0969594X.2019.1700212

Zainuddin, Z., Shujahat, M., Haruna, H., & Chu, S. K. W. (2020). The role of gamified e-quizzes on student learning and engagement: An interactive gamification solution for a formative assessment system. *Computers & Education, 145*, 103729. https://doi.org/10.1016/j.compedu.2019.103729