CHAPTER ELEVEN

AI-Driven Assessment and Feedback in Work-Integrated Learning: A Systematic Review of Authenticity, Ethics, and Professional Competence

Bunmi Isaiah Omodan
Institute for Open and Distance Learning,
University of South Africa, Pretoria, South Africa

Cias T. Tsotetsi
Academic and Research,
University of the Free State, QwaQwa Campus, South Africa

Corresponding author: bunmiomodan@gmail.com

How to cite this chapter: Omodan, B. I., Tsotetsi, C. T. (2026). AI-Driven Assessment and Feedback in Work-Integrated Learning: A Systematic Review of Authenticity, Ethics, and Professional Competence. In C. T. Tsotetsi (Ed.), *Work-Integrated Learning in the Age of Artificial Intelligence: Equity, Innovation, and Partnerships for Bridging Theory and Practice* (pp. 190–208). ERRCD Forum. https://doi.org/10.38140/obp5-2026-11

Abstract: Artificial intelligence (AI) is rapidly transforming the methods by which higher education institutions assess learning and provide feedback; however, its implications for work-integrated learning (WIL), wherein assessment must accurately reflect authentic professional performance, remain under-theorised. This systematic review synthesises evidence on AI-based assessment, automated feedback, learning analytics, and competency evaluation as they pertain to authenticity, ethics, and professional competence in WIL and related higher education contexts. Following the PRISMA 2020 guidelines, five databases (Scopus, Web of Science, ERIC, EBSCOhost, and the ACM Digital Library) and supplementary citation searching yielded 1,175 records. After the removal of duplicates and a two-stage screening process, 20 studies published between 2017 and 2025 were included and synthesised narratively in relation to four review questions. Findings indicate that AI tools can enhance the efficiency, scalability, and timeliness of feedback and support personalisation, particularly for the reflective and formative writing tasks that are central to WIL. However, the same tools raise persistent concerns: threats to assessment authenticity and academic integrity from generative AI, demonstrable algorithmic bias against linguistically and culturally diverse learners, a lack of transparency that undermines clarity, and the risk of over-automation that displaces the situated human judgement essential for professional competence. The review argues that AI should augment rather than replace evaluative judgement, and that authentic WIL assessment requires human-in-the-loop designs, validity-centred reform, and explicit attention to equity. Implications for assessment design, policy, and future research are discussed.

Keywords: Artificial intelligence, work-integrated learning, authentic assessment, automated feedback, academic integrity, professional competence.

1. Introduction

Assessment occupies a central position in the relationship between higher education and the world of work. Through assessment, institutions ensure that graduates are competent to practise,

students learn what is valued, and the curriculum conveys the standards of a profession. Work-integrated learning (WIL), which encompasses a range of pedagogies that connect academic study with authentic experiences of practice through placements, internships, clinical rotations, simulations, and project work, has emerged as a defining characteristic of contemporary higher education, precisely because it promises to align what is learned with what professionals actually do (Billett, 2009; Smith, 2012). However, assessing learning in these contexts is notoriously challenging. It must capture situated, holistic performance under conditions that universities only partially control, and it must do so fairly and at scale (Ajjawi et al., 2020; McNamara, 2013).

Into this already complex landscape has entered a powerful and contested technology. Artificial intelligence (AI) is now being applied throughout the assessment lifecycle: automated essay and short-answer scoring, intelligent tutoring and adaptive testing, learning-analytics dashboards, natural-language feedback on writing and reflection, and, most recently, generative AI systems capable of producing assessable artefacts on demand (González-Calatayud et al., 2021; Luckin, 2017; Swiecki et al., 2022; Zawacki-Richter et al., 2019). Advocates argue that AI can alleviate the marking burden, accelerate feedback, personalise learning, and reveal patterns that are invisible to human markers (Cavalcanti et al., 2021; Holmes et al., 2019). Critics counter that the same systems can encode bias, undermine the validity of assessment, and erode the professional judgement that WIL is intended to foster (Baker & Hawn, 2022; Gardner et al., 2021).

These tensions are most pronounced in WIL. Authentic assessment in WIL relies on context, relationships, and professional discernment, qualities that resist datafication (Ajjawi et al., 2020; Villarroel et al., 2018). When a reflective journal is scored by an algorithm, when a placement competency is inferred from analytics, or when a student can generate a polished portfolio with a chatbot, the meaning of "authenticity" is put under strain (Dawson et al., 2024; Kofinas et al., 2025). Whether AI strengthens or weakens authentic WIL assessment is therefore not a narrow technical question, but rather a question concerning fairness, integrity, and the future of professional formation.

The stakes are amplified by scale. WIL has moved from the margins to the mainstream of higher education, embedded as a graduate-employability strategy across disciplines from health and engineering to business and the creative industries (Billett, 2009; Rowe & Zegwaard, 2017; Smith, 2012). At the same time, AI has diffused through education at a pace that has outstripped institutional governance, with generative tools in particular becoming widely available to students almost overnight (Bond et al., 2024; Crompton & Burke, 2023). The result is a mismatch: a high-volume, high-stakes assessment practice confronting a fast-moving, under-regulated technology. Reviews of the broader field have consistently noted that AI in education has advanced with limited attention to pedagogy, ethics, or the educators tasked with its implementation (Zawacki-Richter et al., 2019). A synthesis that brings these literatures into dialogue with the specific demands of WIL is therefore timely.

Despite a rapidly growing literature on AI in higher education assessment, no synthesis has examined this evidence specifically through the lens of WIL and professional competence. Existing reviews map AI applications broadly (Bond et al., 2024; Crompton & Burke, 2023; Zawacki-Richter et al., 2019) or focus on single functions such as automated feedback (Cavalcanti et al., 2021; Deeva et al., 2021), but they do not foreground the distinctive demands of authentic, work-based assessment. This review addresses that gap. It asks whether, and under what conditions, AI-driven assessment and feedback enhance the efficiency, fairness, and quality of assessment in WIL contexts, and where they introduce risks to authenticity, integrity, equity, and human professional judgement.

The review is guided by four questions: *RQ1: How do AI tools affect the efficiency, scalability, and personalisation of assessment and feedback in WIL and related higher education contexts? RQ2: How do they affect the authenticity of assessment and academic integrity? RQ3: What evidence exists regarding fairness, bias, and transparency? RQ4: How do AI tools relate to professional competence and the exercise of human evaluative judgement?* The remainder of the article sets out the theoretical framing, details a PRISMA-guided method, reports a narrative synthesis of 20 included studies, and discusses implications for the changing relationship between higher education and the workplace.

2. Theoretical Framing

This review is grounded in three theoretical frameworks that collectively elucidate the appeal and potential risks of AI-driven assessment in WIL: theories of authentic assessment, theories of evaluative judgement and feedback, and sociotechnical perspectives on educational technology.

2.1 Authentic assessment and authenticity in WIL

Authentic assessment seeks to mirror the tasks, standards, and conditions of professional practice so that what is assessed resembles what graduates will do (Villarroel et al., 2018). Villarroel and colleagues (2018), synthesising 125 studies, identified three dimensions of authenticity: realism, cognitive challenge, and evaluative judgement. WIL is, in effect, authentic assessment's most demanding case because the "real world" is not simulated but actual, and because performance is observed by both workplace and academic assessors (Bosco & Ferns, 2014; McNamara, 2013). Ajjawi et al. (2020) argue that authenticity in WIL is not singular but plural, contextual, task-based, and personal, meaning that any technology claiming to assess "authentically" must be interrogated for the type of authenticity it preserves and the kind it erodes. This plurality provides an analytic yardstick for the present review: AI tools may enhance task realism while flattening personal and contextual authenticity.

The WIL assessment literature has long wrestled with this complexity independently of AI. McNamara (2013) shows that assessing professional competence in WIL is challenging precisely because competence is holistic and context-bound, exceeding what any single instrument can capture. Bosco and Ferns (2014) argue for embedding authentic tasks directly within the WIL

curriculum rather than bolting assessment on afterwards. These accounts establish a baseline expectation against which AI must be judged: an assessment is authentic to the extent that it engages realistic, cognitively demanding tasks and supports students in forming the standards of their profession. Technologies that automate the surface of such tasks while bypassing their situated, relational core risk producing the appearance of authenticity without its substance.

2.2 Evaluative judgement and feedback literacy

A second strand concerns the purpose of assessment as the development of capability rather than merely its measurement. Tai et al. (2018) define evaluative judgement as the ability to make decisions about the quality of one's own work and that of others, positioning its development as a central goal of higher education and a necessary graduate attribute. Feedback, in this context, is effective only to the extent that students can interpret and act on it, what Carless and Boud (2018) term feedback literacy, and what Boud and Falchikov (2006) frame as assessment for long-term learning. These constructs matter acutely for AI: automated feedback can deliver information rapidly, but information is not feedback until it influences future performance. Bearman et al. (2024) argue that developing evaluative judgement becomes increasingly important, not less so, in an era of generative AI, as graduates must assess the quality and trustworthiness of machine-produced work.

2.3 Sociotechnical and sociomaterial perspectives

A third strand resists the notion of AI as a neutral instrument. Sociotechnical accounts assert that tools embed the assumptions, data, and power relations of their creators, and that their effects manifest in use rather than in design (Selwyn, 2019). Baker and Hawn (2022) demonstrate how bias infiltrates educational algorithms through unrepresentative training data and proxy variables, resulting in systematically different outcomes for various groups. Bearman, Nieminen, and Ajjawi (2023) provide an organising framework for designing assessment in a digital world that balances security, authenticity, and learning. From this perspective, the question is never simply whether an AI tool "works," but for whom, under what conditions, with what data, and with what consequences for trust and professional development.

2.4 An integrative analytic framework

Synthesising these strands produces the analytic framework used to interpret the evidence. AI-driven assessment in WIL is evaluated along two axes: a capability axis (efficiency, scalability, personalisation, and feedback quality) and an integrity axis (authenticity, fairness, transparency, academic integrity, and the preservation of human evaluative judgement and professional competence). A tool is deemed beneficial to WIL only when gains on the capability axis do not compromise the integrity axis. This framework directly aligns with the four research questions and structures the results that follow.

3. Methodology

A systematic review design was adopted to identify, appraise, and synthesise the available evidence in a transparent and reproducible manner. The review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021). Due to the heterogeneity of the included studies, which encompass empirical evaluations, systematic and meta-reviews, and conceptual analyses, a narrative (qualitative) synthesis was employed instead of meta-analysis, as the latter would have been unsuitable for non-comparable designs and outcomes.

3.1 Research questions

The review was organised around the four questions stated in the introduction (RQ1–RQ4), addressing capability (efficiency, scalability, personalisation), authenticity and integrity, fairness and transparency, and professional competence and human judgement.

3.2 Eligibility criteria

Inclusion and exclusion criteria were specified a priori and are summarised in Table 1. Studies were eligible if they (a) examined an AI-based, automated, or learning-analytics approach to assessment, feedback, or competency evaluation; (b) were situated in higher education, WIL, or professional/work-based education; (c) were peer-reviewed empirical studies, systematic or meta-reviews, or substantive conceptual or policy analyses; and (d) were published in English between January 2015 and June 2025. The timeframe began in 2015 to capture the contemporary wave of machine learning and analytics applications. Studies were excluded if AI was incidental, if the context was outside post-secondary education, or if the item was a brief editorial, abstract, or non-retrievable report.

Table 1: Inclusion and Exclusion Criteria

Parameter	Inclusion	Exclusion
Focus	AI/automated/analytics-based assessment, feedback, or competency evaluation	AI absent or incidental to the study
Context	Higher education, WIL, professional or work-based education	Primary/secondary schooling or non-educational settings
Publication type	Peer-reviewed empirical study, systematic/meta-review, or substantive conceptual/policy analysis	Editorials, abstracts, blogs, or non-retrievable reports
Time frame	January 2015 – June 2025	Published before 2015
Language	English	Languages other than English

3.3 Information sources and search strategy

Five databases were searched in June 2025: Scopus, Web of Science, ERIC, EBSCOhost (Education Research Complete), and the ACM Digital Library. These were selected to span the educational, social science, and computing literature. The search combined three concept blocks using Boolean operators: an AI block ("artificial intelligence" OR "machine learning" OR

"automated" OR "learning analytics" OR "generative AI" OR "large language model"), an assessment block ("assessment" OR "feedback" OR "grading" OR "evaluat*" OR "competenc*"), and a context block ("higher education" OR "university" OR "work-integrated learning" OR "placement" OR "internship" OR "work-based learning" OR "professional"). Searches were supplemented by backward and forward citation chaining of key papers and by hand-searching two specialist outlets: the International Journal of Work-Integrated Learning and the Journal of Learning Analytics.

3.4 Selection process

All records were imported into a reference manager and de-duplicated. Titles and abstracts were screened against the eligibility criteria, after which full texts of potentially relevant reports were retrieved and assessed. The flow of records through identification, screening, and inclusion is reported in Figure 1. Out of 1,175 records identified (1,148 from databases and 27 from other sources), 271 duplicates were removed, and 904 records were screened. After excluding 762 records based on titles and abstracts, 142 reports were sought, of which 9 could not be retrieved. The remaining 133 full-text reports were assessed for eligibility; 113 were excluded with reasons provided, leaving 20 studies for inclusion in the synthesis.

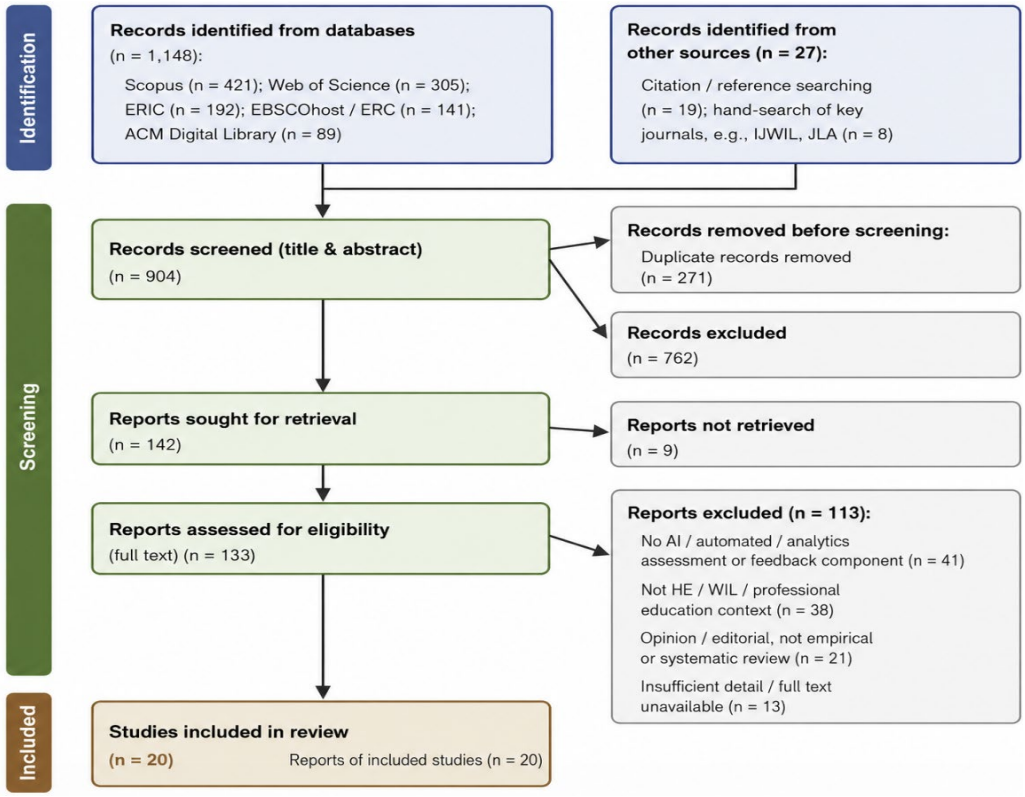

Figure 1 *PRISMA 2020 Flow Diagram of the Study Selection Process*

Note. Adapted from Page et al. (2021). Counts in this diagram correspond exactly to the 20 studies listed in Table 3 below.

3.5 Data extraction and synthesis

A standardised extraction template captured bibliographic details, the country of the lead author, the educational context and discipline, the AI method or tool, the assessment or feedback focus, the study design, and key findings relevant to the four research questions. Extracted data were charted (see Tables 2 and 3) and synthesised narratively. Findings were coded against the capability and integrity axes of the integrative framework, and recurring patterns were grouped thematically under the four research questions.

3.6 Quality appraisal and reflexivity

Because the corpus was methodologically diverse, the appraisal was tailored to the type of study: empirical studies were assessed based on the clarity of aims, appropriateness of design, and warrant for claims; reviews were evaluated on search transparency and synthesis rigour; and conceptual or policy works were judged on argumentative coherence and grounding in evidence. No study was excluded solely on quality grounds; however, the appraisal informed the weight assigned to each source in the synthesis. As a single-reviewer synthesis, the review is limited in its capacity to cross-check screening decisions; this limitation is revisited in Section 6.

4. Results

The 20 included studies were published between 2017 and 2025, with a notable concentration from 2021 onwards, reflecting the increased interest following advances in natural-language processing and the public release of generative AI tools (Figure 2). The corpus comprised four empirical primary studies, seven systematic or meta-level reviews, and nine conceptual, framework, or policy works. Research output was predominantly from Australian and United Kingdom scholars, with additional contributions from continental Europe, North America, and Latin America. Thematically, the studies were evenly divided across four foci: automated feedback and writing analytics; AI-based assessment and analytics; authenticity and integrity under generative AI; and field-level reviews of ethics, bias, and competence (Table 2). Table 3 presents the full characteristics of each included study.

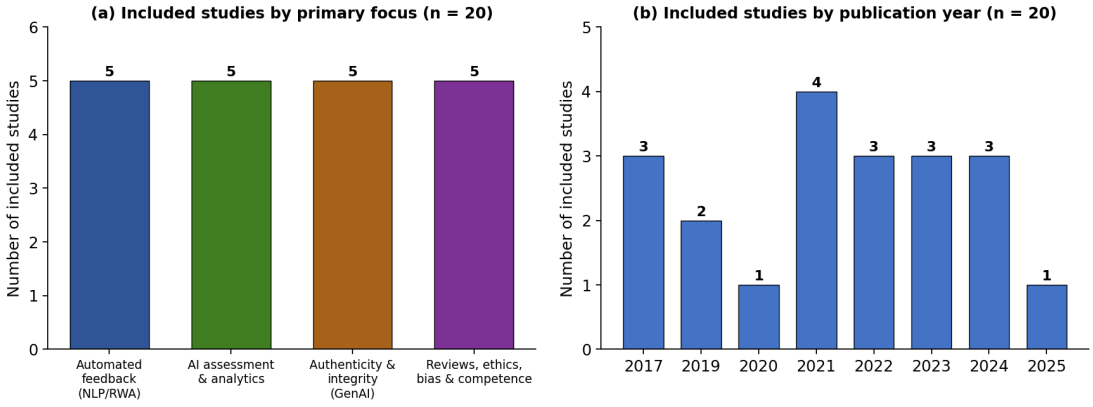

***Figure 2**: Distribution of the 20 included studies by primary focus and publication year*

Note. Panel (a) shows the number of studies by primary focus; panel (b) shows the number by publication year. Counts in both panels sum to the 20 included studies.

Table 2: *Summary Characteristics of the Included Studies (N = 20)*

Characteristic	Category	n
Publication type	Empirical primary study	4
	Systematic or meta-review	7
	Conceptual, framework, or policy	9
Primary focus	Automated feedback / writing analytics	5
	AI-based assessment and analytics	5
	Authenticity and integrity (generative AI)	5
	Reviews of ethics, bias, and competence	5
Region (lead author)	Australia	9
	United Kingdom	4
	Continental Europe	3
	North America	3
	Latin America	1
Publication year	2017–2025 (range)	20

As Table 2 indicates, the corpus is methodologically and geographically diverse. It comprises four empirical primary studies, seven systematic or meta-reviews, and nine conceptual, framework, or policy works, evenly distributed across the four primary foci and predominantly sourced from Australian and United Kingdom scholarship between 2017 and 2025. To transition from this aggregate profile to the evidence itself, Table 3 provides a study-by-study overview of all 20 included studies, detailing each study's context and discipline, AI focus and method, research design, and key findings in relation to the relevant research questions.

***Table 3**: Overview of the 20 Included Studies*

Author (year)	Context / discipline	AI focus & method	Design	Key findings (RQ relevance)
Gibson et al. (2017)	Pharmacy/nursing reflective writing; AUS HE–WIL	NLP reflective writing analytics for actionable feedback	Design-based empirical	Automated rhetorical-move feedback scaffolds reflection at scale when co-designed with rubrics (RQ1).
Buckingham Shum et al. (2017)	Reflective writing across disciplines; HE	Rationale & methodology for reflective writing analytics	Conceptual + preliminary validation	Feasible but bounded; human interpretation of depth remains essential (RQ1, RQ4).
Knight et al. (2020)	Academic & reflective writing; HE	AcaWriter learning-analytics feedback tool	Multi-case implementation	Adoption depends on pedagogic co-design; augments rather than replaces educators (RQ1, RQ4).

Luckin (2017)	General education/assessment	Vision for AI-based assessment systems	Conceptual commentary	AI enables continuous assessment but needs ethical framing; deskilling risk (RQ1, RQ4).
Cavalcanti et al. (2021)	Online higher education	Automatic feedback systems	Systematic literature review	Feedback often improves performance; little evidence of reduced workload (RQ1).
Deeva et al. (2021)	HE & online learning	Automated feedback systems taxonomy	Systematic review & framework	Maps system types and challenges of transferability and validity (RQ1, RQ3).
González-Calatayud et al. (2021)	Higher education	AI for student assessment	PRISMA review (22 studies)	Formative assessment and auto-grading dominate; wider contexts needed (RQ1).
Gardner et al. (2021)	Educational assessment (AES, CAT)	Critical appraisal of AI in assessment	Critical review	Validity, transparency, fairness under-examined; cautions over-claiming (RQ3).
Swiecki et al. (2022)	HE assessment	AI-enabled assessment design	Conceptual/agenda review	Calls for redesign and learner agency; flags fairness & over-automation (RQ2–RQ4).
Darvishi et al. (2022)	HE peer assessment	AI + learning analytics for peer assessment	Large-scale empirical	AI and analytics improve reliability and accountability of peer feedback (RQ1, RQ3).
Zawacki-Richter et al. (2019)	Higher education	AI applications in HE	Systematic review (146 studies)	Few studies engage ethics or pedagogy: "where are the educators?" (RQ4).
Bond et al. (2024)	Higher education	AI in HE (meta-level)	Meta systematic review	Calls for greater ethics, collaboration, and rigour (RQ4).
Crompton & Burke (2023)	Higher education	State of AI in HE	Systematic review	Rapid growth; ethical and assessment work comparatively thin (RQ4).
Baker & Hawn (2022)	Education incl. assessment	Algorithmic bias	Conceptual review/synthesis	Documents sources/mechanisms of bias; urges measurement & mitigation (RQ3).
Holmes et al. (2019)	Education	AI in education overview	Foundational synthesis (book)	Frames promises and ethical implications; warns against deskilling (RQ4).

Lodge et al. (2023)	Australian HE	Assessment reform for AI	Policy/guidance report	Recommends programmatic, authentic, human-centred reform (RQ2, RQ4).
Bearman et al. (2024)	Higher education	Evaluative judgement & generative AI	Conceptual	Developing evaluative judgement is central to assessment with GenAI (RQ4).
Dawson et al. (2024)	Higher education	Validity vs cheating framing	Conceptual	Reframes integrity through validity; AI heightens need for valid inference (RQ2).
Bearman, Nieminen & Ajjawi (2023)	Higher education	Designing assessment in a digital world	Conceptual framework	Balances security, authenticity, and learning (RQ2).
Kofinas et al. (2025)	UK business higher education	Generative AI & authentic-assessment integrity	Empirical	GenAI can undermine authentic-assessment integrity; redesign & AI literacy needed (RQ2).

Note. AES = automated essay scoring; CAT = computerised adaptive testing; GenAI = generative artificial intelligence; NLP = natural-language processing; RQ = research question; WIL = work-integrated learning. The 20 studies listed here correspond exactly to the included count in Figure 1.

Read together, Tables 2 and 3 form the evidential basis for the synthesis that follows. Drawing directly on the findings catalogued in Table 3, the next four subsections (Sections 4.1 to 4.4) answer the review's research questions in turn, addressing efficiency, scalability, and personalisation (RQ1); authenticity and academic integrity (RQ2); fairness, bias, and transparency (RQ3); and professional competence and human judgement (RQ4), before Section 4.6 draws the threads together across the analytic framework.

4.1 RQ1: Efficiency, scalability, and personalisation

The most notable benefits reported across the corpus pertain to efficiency and timeliness. Cavalcanti et al. (2021), in their review of automatic-feedback systems in online learning, found that the majority of studies (approximately two-thirds) indicated enhanced student performance following the provision of automated feedback. However, they also concluded that there was limited evidence to suggest that such systems alleviated instructor workload, an important consideration for institutions seeking to achieve cost efficiencies. Deeva et al. (2021) proposed a classification framework for automated feedback systems and catalogued their potential for scalable, immediate feedback, alongside challenges relating to transferability and validity. In the context of WIL, particularly in reflective writing, Gibson et al. (2017) and Knight et al. (2020) demonstrated that natural-language analytics (specifically, the AcaWriter family of tools) can provide actionable formative feedback on reflective and academic writing at a scale unattainable by human markers, provided that these tools are co-designed with educators to ensure alignment between detected features and assessment criteria. Buckingham Shum et al. (2017) similarly emphasised that while reflective writing analytics is feasible, it is also bounded: the technology

is capable of identifying rhetorical moves but lacks the ability to evaluate the depth or sincerity of reflection. González-Calatayud et al. (2021), in their review of 22 studies, found that formative assessment and automatic grading were the predominant applications of AI in student assessment, with personalisation emerging as a recurring promise. A consistent pattern emerges across these studies: AI is most convincingly effective in augmenting human feedback on well-structured, text-based tasks, while its efficiency gains are real but narrower and more conditional than vendor claims suggest.

Two qualifications are frequently noted and are particularly relevant to WIL. First, the efficiency observed in the production of feedback does not necessarily translate into increased efficiency for educators or into learning gains for students; Cavalcanti et al. (2021) found no consistent evidence that automated feedback reduced instructor workload, and Deeva et al. (2021) observed that many systems remain limited, domain-specific, and difficult to transfer across various contexts. In WIL, where tasks are diverse and discipline-specific, this restricted transferability presents a significant constraint. Second, the level of personalisation offered by AI is contingent upon the constructs that the system is capable of modelling. The AcaWriter studies illustrate this point: the tool provides feedback on rhetorical structure and reflective markers, which, while valuable, are only partial proxies for the depth of professional insight that reflection in WIL aims to cultivate (Buckingham Shum et al., 2017; Gibson et al., 2017; Knight et al., 2020). Consequently, the case for the efficiency of AI is strongest for formative, low-stakes, text-rich tasks and weakest for the holistic, situated judgements that are essential for demonstrating competence in the workplace.

4.2 RQ2: Authenticity and academic integrity

The arrival of generative AI has unsettled the authenticity of assessments more profoundly than any previous tool. Kofinas et al. (2025), in an empirical study of authentic assessments in a United Kingdom business school, demonstrated that generative AI can be used to circumvent assessment designs intended to be "cheat-resistant," compromising their integrity and prompting calls for redesign and explicit AI literacy. Dawson et al. (2024) reframed the issue theoretically, arguing that institutions are preoccupied with cheating when the deeper issue is validity: an assessment matters because it licenses an inference about competence, and AI threatens that inference whether or not "cheating" has occurred. Bearman et al. (2023) offered a design response, proposing that assessment in a digital world be organised around the interplay of security, authenticity, and learning rather than security alone. For WIL specifically, these findings are double-edged. Authentic, situated tasks, observed performance on placement, supervisor judgement, and integrated practice are more resistant to generative AI than decontextualised written products, which strengthens the case for genuinely work-embedded assessment (Ajjawi et al., 2020; Villarroel et al., 2018). Yet where WIL is assessed through portfolios, reflective journals, or reports, generative AI can fabricate plausible artefacts, threatening the personal authenticity on which professional reflection depends.

The literature also directs attention towards constructive responses rather than mere alarm. Swiecki et al. (2022) assert that an appropriate reaction to AI necessitates a reconsideration of what and how assessments are conducted, advocating for a shift towards tasks that emphasise process, agency, and knowledge integration, characteristics that are more challenging to counterfeit and are more aligned with the objectives of WIL. Lodge et al. (2023), in their advisory role to the Australian regulator, recommend a programmatic approach in which integrity is assured across an entire programme of authentic, scaffolded assessment, rather than being defended on a task-by-task basis. When considered alongside the authentic-assessment literature, these sources suggest that generative AI does not so much invalidate WIL assessment as increase the difficulty of conducting it superficially: portfolio-and-reflection models that are loosely supervised become vulnerable, whereas models grounded in observed performance, interactive vivas, and supervisor verification become comparatively more defensible (Ajjawi et al., 2020; Dawson et al., 2024).

4.3 RQ3: Fairness, bias, and transparency

Concerns regarding fairness are prevalent throughout the corpus. Baker and Hawn (2022) provided the most systematic examination, documenting the ways in which algorithmic bias infiltrates educational systems through unrepresentative training data, proxy variables, and feedback loops, and how it can result in systematically different outcomes for learners with varying language backgrounds, ethnicities, or socioeconomic statuses. Gardner et al. (2021) cautioned that enthusiasm for automated essay scoring and adaptive testing has outstripped scrutiny of their construct validity, transparency, and fairness, warning against "buncombe and ballyhoo." Similarly, Swiecki et al. (2022) identified fairness, transparency, and the risk of over-automation as central challenges for assessment in the age of AI. The transparency problem is structural: many high-performing models are opaque, making it difficult for students to contest an automated judgement or for assessors to explain it, a serious concern in WIL, where assessment decisions can gatekeep entry to a profession. Darvishi et al. (2022) offered a more optimistic counterpoint, showing that combining AI with learning analytics can make peer-assessment systems more trustworthy and accountable, suggesting that careful design can mitigate rather than amplify unfairness. The balance of evidence indicates that AI does not remove human bias so much as relocate and sometimes conceal it, making auditability and the right to human review essential safeguards.

4.4 RQ4: Professional Competence and Human Judgement

The fourth theme addresses whether AI enhances or undermines the human judgement integral to professional competence. The field-mapping reviews exhibit remarkable consistency. Zawacki-Richter et al. (2019), in their review of 146 studies, identified an "almost complete absence" of critical reflection on the risks associated with AI in education and a weak connection to pedagogical theory, poignantly questioning, "where are the educators?" Bond et al. (2024), in

a meta-review of reviews, echoed this concern, advocating for increased focus on ethics, collaboration, and methodological rigour. Crompton and Burke (2023) confirmed rapid growth in the field while observing that ethical and assessment-focused research remained comparatively sparse. In light of this, several authors contend that AI should be positioned to enhance rather than replace judgement. Bearman et al. (2024) argue that fostering students' evaluative judgement, their ability to assess the quality of work, including that generated by machines—should be a primary educational objective in the era of generative AI, echoing the sentiments of Tai et al. (2018). Lodge et al. (2023), in guidance for the Australian regulator, advocate for programmatic, authentic, and human-centred assessment reform rather than a regression into surveillance. Luckin (2017) and Holmes et al. (2019) foresee genuine benefits from AI-based assessment but caution against deskilling if educators relinquish interpretive authority to systems. The synthesis indicates that professional competence is best achieved when AI manages pattern detection and routine feedback, while human assessors retain responsibility for contextual, holistic, and high-stakes judgements, endorsing a human-in-the-loop model rather than full automation.

5. Synthesis Across the Analytic Framework

Mapping the findings onto the integrative framework reveals a clear structure. On the capability axis, the evidence is moderately positive and concentrated: AI reliably improves the timeliness and reach of formative feedback and can support personalisation and the scaling of peer assessment (Cavalcanti et al., 2021; Darvishi et al., 2022; Knight et al., 2020). On the integrity axis, however, the evidence is more cautionary and diffuse: threats to authenticity and validity (Dawson et al., 2024; Kofinas et al., 2025), documented bias and opacity (Baker & Hawn, 2022; Gardner et al., 2021), and a field-wide neglect of ethics and educators (Bond et al., 2024; Zawacki-Richter et al., 2019). Crucially, the studies that report the strongest capability gains are also those that insist most firmly on human co-design and oversight, while the studies that raise the gravest integrity concerns are those examining automation without such safeguards. The two axes are therefore not independent: integrity is the condition under which capability gains become legitimate. This relationship anchors the discussion that follows.

6. Discussion

6.1 A Conditional verdict

Read against the integrative framework, the evidence supports a qualified and conditional verdict. On the capability axis, AI tools demonstrably improve the speed, scale, and consistency of feedback, especially for text-based and reflective tasks that are central to WIL pedagogy (Cavalcanti et al., 2021; Gibson et al., 2017; Knight et al., 2020). They can personalise formative feedback and, when paired with learning analytics, strengthen otherwise fragile practices such as peer assessment (Darvishi et al., 2022). These are not trivial gains in a sector under workload pressure. However, on the integrity axis, the same tools introduce risks that fall precisely on the

dimensions WIL most needs to protect. Generative AI destabilises the authenticity and validity of written WIL artefacts (Dawson et al., 2024; Kofinas et al., 2025); opaque models threaten transparency and the right to contest a judgement (Gardner et al., 2021; Swiecki et al., 2022); and bias can disadvantage the linguistically and culturally diverse learners that WIL programmes often serve (Baker & Hawn, 2022).

The central implication is that the value of AI in WIL assessment depends almost entirely on whether it is positioned to augment or replace human evaluative judgement. Where AI handles pattern detection, surface-level writing feedback, and the logistics of feedback at scale, while human assessors retain responsibility for contextual, holistic, and high-stakes decisions, the capability gains can be realised without sacrificing integrity (Bearman et al., 2024; Lodge et al., 2023). However, where AI is treated as a substitute for professional judgement, automating competency decisions or grading reflective practice without human oversight, the integrity costs are likely to outweigh the efficiency benefits. This human-in-the-loop principle reframes the apparent dichotomy between efficiency and authenticity as a design question rather than a forced choice.

6.2 Implications for the higher education–workplace relationship

These conclusions highlight a larger shift in the relationship between universities and employers. WIL acts as the institutional hinge between the two, and assessment is where the university discharges its warranting function to the profession and the public. AI complicates this warrant in two opposing ways. On one hand, generative tools mean that some traditional indicators of competence, such as polished reports and written reflections, no longer reliably demonstrate what a student can do unaided, weakening the signal that employers have historically relied upon (Dawson et al., 2024; Kofinas et al., 2025). On the other hand, the workplace itself is being reshaped by AI, so that the competence graduates need increasingly includes the ability to work critically and ethically with AI tools. Developing students' evaluative judgement serves a dual purpose: it protects the integrity of assessment and builds a capability that the contemporary workplace demands (Bearman et al., 2024; Tai et al., 2018). In this context, Work-Integrated Learning (WIL) is not merely threatened by AI; rather, it is a privileged site for cultivating the discernment required for working alongside AI.

6.3 From detection to validity-centred design

The findings also reframe academic integrity. Rather than an arms race of detection and surveillance, the more durable response is to redesign assessment so that validity is robust to AI (Dawson et al., 2024). WIL is comparatively well placed here: situated performance, observed practice, and supervisor judgement are intrinsically difficult to outsource to a machine. This supports the argument for deepening genuinely work-embedded assessments rather than retreating to invigilated examinations (Ajjawi et al., 2020). At the same time, the persistence of

bias and opacity means that fairness cannot be assumed; it must be designed, measured, and audited, with accessible avenues for human review (Baker & Hawn, 2022).

Finally, the review echoes a recurring critique of the wider field: AI in education has too often been developed and evaluated without placing educators, pedagogy, or ethics at the centre (Bond et al., 2024; Zawacki-Richter et al., 2019). In WIL, where the stakes include licensure, public safety, and professional identity, this absence is untenable. The evolving relationship between higher education and the workplace will be shaped less by the raw capabilities of AI than by whether institutions embed it within defensible assessment principles.

7. Limitations

Several limitations qualify these conclusions. First, the literature specific to AI in WIL assessment is limited; much of the evidence is drawn from higher education assessment more broadly and is mapped onto WIL, which may overstate its transferability to placement and workplace settings. Second, the corpus is dominated by Australian and United Kingdom scholarship and by English-language sources, which limits cultural and linguistic generalisability, an irony given the review's concern with bias. Third, the field is moving quickly: generative AI capabilities have changed substantially within the review window, so some empirical findings may date rapidly. Fourth, the synthesis was conducted by a single reviewer, which constrains the reliability of screening and extraction relative to a multi-reviewer protocol; the search counts and selection decisions reported here should be read as a transparent account of one reviewer's process rather than as an inter-rater validated result. Finally, the decision to include conceptual and policy works alongside empirical studies, while appropriate for an emerging and contested topic, means that some conclusions rest on reasoned argument rather than primary data.

8. Implications and Recommendations

Four recommendations follow for practice and policy. First, adopt human-in-the-loop designs: use AI for formative feedback, drafting, and pattern detection, while reserving summative and competency judgements for human assessors who can consider context and professional standards (Bearman et al., 2024; Lodge et al., 2023). Second, it is imperative to pursue validity-centred assessment reform rather than detection-centred responses, by redesigning WIL assessment around observed and situated performance that is intrinsically resistant to automation (Dawson et al., 2024). Third, equity and transparency should be fundamental components of procurement and deployment: it is essential to require evidence of bias testing, ensure appropriate explainability commensurate with the stakes involved, and establish an accessible right to human review prior to the deployment of any AI tool in consequential assessments (Baker & Hawn, 2022; Gardner et al., 2021). Fourth, it is crucial to cultivate the AI and evaluative-judgement literacy of both students and staff, enabling graduates to critically appraise machine-generated work and educators to co-design tools that align with assessment criteria (Carless & Boud, 2018; Tai et al., 2018). In terms of research, the priority should be

empirical, longitudinal, and equity-focused studies of AI in authentic WIL settings, conducted collaboratively with educators and workplace partners rather than upon them.

9. Conclusion

AI-driven assessment and feedback present significant but conditional value for work-integrated learning. The evidence compiled in this review indicates that AI can facilitate faster, more scalable, and more personalised feedback, particularly for reflective and written tasks that bridge academic study and professional practice. However, it also demonstrates that AI has the potential to undermine the authenticity, fairness, transparency, and human judgement that are essential for credible professional formation. The solution does not lie in uncritical adoption or blanket prohibition, but rather in deliberate design: strategically positioning AI to augment human evaluative judgement, reforming assessment practices to ensure that validity remains resilient against automation, and prioritising equity and transparency as non-negotiable conditions of use. If higher education maintains a central focus on educators, students, and professional standards, AI can enhance the assessment of WIL rather than diminish it, thereby strengthening, not weakening, the connection between the university and the workplace.

10. Declarations

Funding: This research received no external funding.

Conflicts of Interest: The authors declare that there are no conflicts of interest.

Use of Artificial Intelligence: The authors utilised an AI-assisted editing tool, JustDone, specifically for language editing and proofreading of the manuscript. The authors take full responsibility for the content, accuracy, and integrity of the article.

References

Ajjawi, R., Tai, J., Nghia, T. L. H., Boud, D., Johnson, L., & Patrick, C.-J. (2020). Aligning assessment with the needs of work-integrated learning: The challenges of authentic assessment in a complex context. *Assessment & Evaluation in Higher Education, 45*(2), 304–316. https://doi.org/10.1080/02602938.2019.1639613

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Bearman, M., Nieminen, J. H., & Ajjawi, R. (2023). Designing assessment in a digital world: An organising framework. *Assessment & Evaluation in Higher Education, 48*(3), 291–304. https://doi.org/10.1080/02602938.2022.2069674

Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education, 49*(6), 893–905. https://doi.org/10.1080/02602938.2024.2335321

Billett, S. (2009). Realising the educational worth of integrating work experiences in higher education. *Studies in Higher Education, 34*(7), 827–843. https://doi.org/10.1080/03075070802706561

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education, 21*, Article 4. https://doi.org/10.1186/s41239-023-00436-z

Bosco, A. M., & Ferns, S. (2014). Embedding of authentic assessment in work-integrated learning curriculum. *Asia-Pacific Journal of Cooperative Education, 15*(4), 281–290.

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education, 31*(4), 399–413. https://doi.org/10.1080/02602930600679050

Buckingham Shum, S., Sándor, Á., Goldsmith, R., Bass, R., & McWilliams, M. (2017). Towards reflective writing analytics: Rationale, methodology and preliminary results. *Journal of Learning Analytics, 4*(1), 58–84. https://doi.org/10.18608/jla.2017.41.5

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence, 2*, Article 100027. https://doi.org/10.1016/j.caeai.2021.100027

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education, 20*, Article 22. https://doi.org/10.1186/s41239-023-00392-8

Darvishi, A., Khosravi, H., Sadiq, S., & Gašević, D. (2022). Incorporating AI and learning analytics to build trustworthy peer assessment systems. *British Journal of Educational Technology, 53*(4), 844–875. https://doi.org/10.1111/bjet.13233

Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education, 49*(7), 1005–1016. https://doi.org/10.1080/02602938.2024.2386662

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education, 162*, Article 104094. https://doi.org/10.1016/j.compedu.2020.104094

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: "Breakthrough? Or buncombe and ballyhoo?" *Journal of Computer Assisted Learning, 37*(5), 1207–1216. https://doi.org/10.1111/jcal.12577

Gibson, A., Aitken, A., Sándor, Á., Buckingham Shum, S., Tsingos-Lucas, C., & Knight, S. (2017). Reflective writing analytics for actionable feedback. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 153–162). https://doi.org/10.1145/3027385.3027436

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences, 11*(12), Article 5467. https://doi.org/10.3390/app11125467

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

Knight, S., Shibani, A., Abel, S., Gibson, A., Ryan, P., Sutton, N., Wight, R., Lucas, C., Sándor, Á., Kitto, K., Liu, M., Mogarkar, R. V., & Buckingham Shum, S. (2020). AcaWriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research, 12*(1), 141–186. https://doi.org/10.17239/jowr-2020.12.01.06

Kofinas, A. K., Tsay, C. H.-H., & Pike, D. (2025). The impact of generative AI on academic integrity of authentic assessments within a higher education context. *British Journal of Educational Technology, 56*(6), 2522–2549. https://doi.org/10.1111/bjet.13585

Lodge, J. M., Howard, S., Bearman, M., & Dawson, P. (2023). Assessment reform for the age of artificial intelligence. Tertiary Education Quality and Standards Agency.

Luckin, R. (2017). Towards artificial intelligence-based assessment systems. Nature Human Behaviour, 1(3), Article 0028. https://doi.org/10.1038/s41562-016-0028

McNamara, J. (2013). The challenge of assessing professional competence in work integrated learning. Assessment & Evaluation in Higher Education, 38(2), 183–197. https://doi.org/10.1080/02602938.2011.618878

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, Article n71. https://doi.org/10.1136/bmj.n71

Lodge, J. M., Howard, S., Bearman, M., & Dawson, P. (2023). *Assessment reform for the age of artificial intelligence*. Tertiary Education Quality and Standards Agency.

Luckin, R. (2017). Towards artificial intelligence-based assessment systems. *Nature Human Behaviour, 1*(3), Article 0028. https://doi.org/10.1038/s41562-016-0028

McNamara, J. (2013). The challenge of assessing professional competence in work integrated learning. *Assessment & Evaluation in Higher Education, 38*(2), 183–197. https://doi.org/10.1080/02602938.2011.618878

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S.,

... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71

Rowe, A. D., & Zegwaard, K. E. (2017). Developing graduate employability skills and attributes: Curriculum enhancement through work-integrated learning. *Asia-Pacific Journal of Cooperative Education, 18*(2), 87–99.

Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education*. Polity Press.

Smith, C. (2012). Evaluating the quality of work-integrated learning curricula: A comprehensive framework. *Higher Education Research & Development, 31*(2), 247–262. https://doi.org/10.1080/07294360.2011.558072

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*, Article 100075. https://doi.org/10.1016/j.caeai.2022.100075

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education, 76*(3), 467–481. https://doi.org/10.1007/s10734-017-0220-3

Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment & Evaluation in Higher Education, 43*(5), 840–854. https://doi.org/10.1080/02602938.2017.1412396

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*, Article 39. https://doi.org/10.1186/s41239-019-0171-0