IIAR JOURNALS

Innovating Knowledge. Advancing Research.

IIAR Journal of Education, Learning, and Human Development

Vol 1, No 1, 2026

Assessment Integrity in the Age of Artificial Intelligence: A Theoretical Framework of Validity, Authorship and Trust in Educational Measurement

OGAH Sunday Ogbo[1], NWOGBOJI Emmanuel Usulor[2], IGWE Emmanuel Sunday[3]

[1,3] Veristzon Global Research Lab; [2] International Institute of AI and Research

ABSTRACT

Generative artificial intelligence has broken the assumption behind most educational assessment: that submitted work is evidence of what its submitter knows and can do. Institutions have mostly treated this as cheating, but the literature lacks a measurement-grounded account of why some responses help and others harm, and neglects systems where national examinations govern access to further study. This theoretical paper reframes it as a validity problem, recasting authorship as attribution of evidence and trust as the downstream value of credentials, aiming to derive a framework and testable propositions for assessment redesign. The population is the peer-reviewed literature on validity theory, academic integrity, AI-generated work, AI detection and assessment redesign. The sample is 37 sources: 35 journal articles published 2023–2026 and two foundational works cited for the theories they originated, chosen by criterion-based (purposive) selection, verified against Crossref and screened for retractions. Analysis was a conceptual synthesis, mapping each reported threat onto the inferences linking observed performance to claims about a learner, then comparing responses and classifying tasks to derive propositions. Detection emerges as unreliable and inequitable, and prohibition without structural change as unenforceable; only redesign changes the evidence on which the inference rests, though at a cost in staff time and with fairness risks. The framework classifies tasks by the inference they support and their exposure to generative assistance, yielding four decisions: retain, relocate, reconstruct or reduce the stakes. Eight propositions state which task types keep their validity, which must move to supervised or oral conditions, which should be rebuilt to assess the use of AI itself, and how these choices affect the trust employers and receiving institutions place in credentials. The paper recommends that examination bodies such as WAEC, NECO and JAMB anchor high-stakes claims in supervised papers and audit coursework, that universities stop treating detector scores as evidence of misconduct, and that quality-assurance agencies require programme-level validity cases and fund tests of the propositions. These steps matter most in Nigeria and other African systems where, the paper assumes, high-stakes examinations carry exceptional weight and many candidates write English as a second language.

Keywords: assessment; academic integrity; validity; generative AI; AI detection

Assessment Integrity in the Age of Artificial Intelligence: A Theoretical Framework of Validity, Authorship and Trust in Educational Measurement

1. INTRODUCTION

Every assessment rests on a wager. A lecturer grading an essay or an examiner scoring a script judges something unseen: the knowledge, reasoning or skill of the person who produced the work. Work could always be copied, bought or ghost-written, but producing fluent, tailored work without the underlying capability used to take effort, money or accomplices. Generative artificial intelligence (GenAI) has removed most of that cost. In a blind test at a UK university, answers written entirely by AI were submitted to unsupervised take-home examinations in five psychology modules; 94% went undetected, and on average they outscored the students' own work (Scarfe et al., 2024).

Institutions have mostly framed this as a cheating crisis. Early commentary recommended new policies, staff training and detection methods (Cotton et al., 2024), and of the world's fifty highest-ranked universities examined by Moorhouse et al. (2023), just under half had public guidance, with academic integrity one of its three main areas. The framing is too narrow. Cheating is a behaviour; the deeper problem is that the evidence no longer supports the claims built on it. A take-home essay grade that once licensed the claim "this student can construct a disciplinary argument in writing" now licenses a weaker, more ambiguous one, whether or not any rule was broken. Several authors have begun to say so (Dawson et al., 2024; Corbin et al., 2025; Weidlich, 2026), but three gaps remain. First, there is no systematic account of why detection, prohibition and redesign differ in their effect on validity, as opposed to misconduct rates. Second, most redesign guidance classifies tasks by format (essay, exam, oral) or permitted AI level rather than by the inference the task should support; Weidlich (2026) starts from the claim but offers a reasoning procedure for course and programme teams, not a classification crossing inference type with graded AI exposure, so institutions still lack a ready way to tell which tasks yield trustworthy evidence. Third, the literature is concentrated in Australia, the United Kingdom, North America and East Asia, while systems where assessment carries the heaviest social weight are underrepresented. In Nigeria, results from the West African Examinations Council (WAEC), the National Examinations Council (NECO) and the Joint Admissions and Matriculation Board (JAMB) gate entry to tertiary education, as national examinations do elsewhere in Africa; the paper treats that weight as a premise rather than an empirical finding. Many candidates there are second-language (L2) writers of English, and detection tools have been shown to misclassify non-native English writing (Liang et al., 2023).

This paper responds to these gaps theoretically. Its significance is practical as well as conceptual, because institutions are already choosing between detection, prohibition and redesign without a measurement rationale. Its objectives are to conceptualise the disruption caused by GenAI in terms of validity theory, to compare the main institutional responses against that standard, to build a framework for classifying assessment tasks, and to draw out the consequences for credential trust in examination-centred African systems. Being conceptual, it advances testable propositions in place of tested hypotheses. Four research questions follow. RQ1: How should the disruption caused by GenAI be conceptualised in terms of validity theory rather than misconduct? RQ2: How do detection, prohibition and redesign compare when judged against the validity of the inferences they are meant to protect? RQ3: What framework can classify assessment tasks so that institutions can decide which retain validity, which must move to supervised or oral conditions, and which should be rebuilt to assess the use of AI itself? RQ4: What follows for trust in credentials, and specifically for examination bodies, universities and quality-assurance agencies in Nigeria and comparable African systems?

© IIARJournal. All rights reserved.

2. LITERATURE REVIEW

2.1 CONCEPTUAL REVIEW

Four concepts carry the argument. The first is *validity*, understood in the measurement tradition not as a property of a test but as the degree to which evidence and theory support the interpretation and use of its scores (Messick, 1995). A perfectly secure examination can therefore be invalid, and a task completed with AI help is not invalid by default. Two of Messick's threats are central. *Construct-irrelevant variance* arises when scores reflect something other than the intended construct, as when a polished essay reflects prompting skill rather than disciplinary reasoning. *Construct underrepresentation* arises when a task captures too little of the construct, as when a ban on AI pushes assessment toward short, recall-based questions that omit the extended reasoning a programme claims to teach.

The second concept is *authorship*. Traditional integrity policy treats authorship as binary: the student wrote it or someone else did. Eaton (2023) argues that this binary is dissolving in what she calls a postplagiarism era, in which AI is woven into ordinary reading, writing and communication and an arms race over detection may be futile. For measurement the useful question is less "who wrote this?" than "what share of the evidence was generated by the person about whom we are drawing a conclusion, and can we tell?" Authorship thus becomes attribution, and misattributed performance is a validity threat in its own right (Weidlich, 2026).

The third concept, introduced here as an analytic dimension, is *AI exposure*: the degree to which generative assistance can produce or substantially shape the evidence a task collects, given how it is administered. Exposure does not follow from format: an essay written in an invigilated room has low exposure, the same essay set for a week at home high exposure. Exposure also depends on current model capability, which varies across cognitive demands; Nguyen Thanh et al. (2023) found that GenAI tools performed very well on lower-order tasks in Bloom's taxonomy and least well on tasks requiring original creation.

The fourth concept is *trust*. Assessment results circulate as credentials, whose value depends on whether employers, other institutions and the public believe the claims they encode. Trust also operates between students and teachers. Luo (2025) found that one-way transparency eroded students' trust in AI-mediated assessment: students were asked to disclose their AI use, sometimes by submitting chat logs, while the criteria by which teachers judged AI-assisted work remained hidden, leaving students feeling unsafe to explore GenAI openly.

2.2 THEORETICAL REVIEW

The framework rests on the argument-based approach to validation. Kane (2013) proposes that every score interpretation rests on an interpretation/use argument, a chain of inferences from observed performance to a scored observation, from that observation to a universe of similar performances (generalisation), from that universe to the target domain of real-world capability (extrapolation), and finally to a decision. Validation means testing the weakest links. Messick's (1995) unified view adds that the consequences of score use, including adverse effects on particular groups, belong within validity.

Kane's framework shows *where* GenAI breaks the argument. Scoring is largely untouched: a marker can still apply a rubric to a script. The break comes earlier, where the observation is assumed to be the candidate's own performance, and later, in extrapolation, where task performance is assumed to indicate what the candidate can do elsewhere. Weidlich (2026) develops a similar analysis, separating misattribution of performance from construct underrepresentation, construct-irrelevant variance, what he terms conditional extrapolation, and unsupported score use. He proposes, for course and programme teams, a reasoning sequence from the intended claim and the AI conditions through threat diagnosis to proportionate redesign and local evidence, with a "Validity Reasoning Matrix" aligning concerns, threatened inferences, redesigns and evidence needs. This paper builds on his analysis by turning the AI conditions into a graded exposure dimension, crossing it with a typology of intended inferences to yield default decisions, and extending the analysis to the system level of credentials and national examinations.

Dawson et al. (2024) argue that validity matters more than cheating and that efforts against cheating can

© IIARJournal. All rights reserved.

reduce validity, as when an anti-cheating technology creates problems for inclusion. Corbin et al. (2025) distinguish discursive changes to assessment, which alter what students are told, from structural changes, which alter how an assessment works; discursive change looks like control but leaves evidential conditions untouched. Bearman et al. (2024) argue that students now urgently need evaluative judgement, in particular of generative AI outputs and of AI-related processes, which identifies a legitimate construct that only an AI-permitted task can measure.

The measurement community has mainly treated GenAI as a tool for examiners: Hao et al. (2024) weigh efficiency and customised evaluation against threats to validity, fairness and test security, and Kaldaras et al. (2024) propose extending the *Standards for Educational and Psychological Testing* to AI-based assessments, for which no validation guidelines yet exist. Candidate-side AI has had far less theoretical attention, though it matters most where results are used for selection.

2.3 EMPIRICAL REVIEW

Benchmarking studies agree on direction but differ in rigour. Nikolic et al. (2023) entered prompts from ten engineering subjects at seven Australian universities into ChatGPT in early 2023; it passed some subjects outright, often with little prompt adjustment. Nguyen Thanh et al. (2023), working with economics assessments in Vietnam, reported strong performance at lower Bloom's levels and weaker performance on theory-based argument, coherence and referencing. Scarfe et al. (2024) offer the most ecologically valid evidence, since unaware markers graded their AI submissions in the ordinary way. All three used models since superseded, so their exposure estimates are likely conservative.

The detection literature is larger and more contested. Weber-Wulff et al. (2023) tested fourteen tools, including Turnitin, and concluded that none was accurate or reliable; the tools leaned toward classifying text as human-written, and paraphrasing or translation degraded performance further. Elkhatat et al. (2023) found five detectors more accurate on GPT-3.5 than on GPT-4 output and inconsistent on human-written controls. Walters (2023) offers a counterweight: of sixteen detectors tested on first-year composition essays, three commercial tools, including Turnitin, performed well on unmodified text, though most others could not separate GPT-4 essays from student writing. Chaka (2024), reviewing seventeen studies across twelve countries, found detector performance inconsistent and recommended pairing tools with human judgement. The apparent disagreement resolves once adversarial behaviour is considered: Perkins, Roe, et al. (2024) applied simple evasion techniques to 805 samples, reduced the accuracy of six major detectors by 17.4%, and concluded that detectors should not be used to decide integrity cases. Laboratory accuracy is therefore an upper bound on accuracy in a system where some students try to evade detection.

Human judgement fares no better. Fleckenstein et al. (2024) found that neither novice (N = 89) nor experienced (N = 200) teachers could reliably identify ChatGPT essays, and both groups were overconfident. Across two UK universities, Kofinas et al. (2025) found that markers generally could not distinguish GenAI-assisted from unassisted work and that their suspicions generated both false positives and false negatives, even on assessments designed to be authentic; they recommend performative, process-based assessment.

The most consequential finding for African systems concerns bias. Liang et al. (2023) ran seven detectors on 91 TOEFL essays by non-native English speakers and 88 essays by US eighth-graders. The detectors classified the American essays accurately but labelled the TOEFL essays as AI-generated at an average false-positive rate of 61.3%; 97.8% were flagged by at least one detector. The essays came from a single Chinese forum, but the mechanism the authors identified, low text perplexity in writing that uses a narrower range of expressions, is not specific to one group, which suggests (though the study did not test it) that the bias may extend to other L2 writers. Gorichanaz (2023) documented the student's side: in 49 Reddit posts, students accused of using ChatGPT, many of whom said the accusation was false, described adopting legalistic defences, and shifting trust between students and technology emerged as a theme.

Evidence on redesign is thinner and mostly single-site. Davey et al. (2025) replaced a written final examination with one-to-one interactive oral assessments in a bioscience course and analysed 722

© IIARJournal. All rights reserved.

students across cohorts from 2009 to 2023. Performance improved after the change, with no significant differences by gender, international status or language background, and the format resisted misconduct, including inappropriate GenAI use. At Durban University of Technology, Nyamadzawo (2025) reported that structured oral examinations promoted equity for students from disadvantaged backgrounds and worked as a translanguaging tool, but were time-consuming and stressful and needed staff and infrastructure. Rae et al. (2026) caution that oral formats can disadvantage women and lose validity unless the interaction is carefully designed. Process evidence is a second strand. Mehta et al. (2026) found that keystroke-dynamics models detected LLM-assisted writing better than text-only detectors or human evaluators, though less well under adversarial conditions. Yang et al. (2026) used writing traces to identify distinct trajectories of GenAI use, from AI-critical to AI-dependent, associated with different cognitive processes. This suggests that process data can show not only misuse but *how* a student used AI, which is itself assessable.

On policy, graded permission scales such as the AI Assessment Scale (AIAS; Perkins, Furze, et al., 2024) have spread quickly, but Corbin et al. (2025) argue that they are mostly discursive. Curtis (2025) rejects the alternative "two-lane" model of fully secured and fully open tasks, arguing that a middle lane of limited, specified use is pedagogically essential. In February 2025 only 46% of South African public universities had publicly documented institution-wide GenAI guidelines, and these stressed student disclosure while treating staff accountability and inclusiveness as advisory (Jacobs-Basadien & Pather, 2026). Members of the task team behind the University of KwaZulu-Natal's guidelines argue that the case shows a resource-constrained institution can build coherent governance (Thaldar et al., 2025). A survey of 899 students at 17 Nigerian universities found GenAI engagement driven by ease of use and alignment with study goals but limited by the need for technical support (Essien et al., 2024). A Nigerian-led survey of 1,217 respondents in 76 countries reported widespread use for information retrieval and paraphrasing, with concern about dishonesty that varied with cultural orientation (Yusuf et al., 2024). Across Ghana, Rwanda, Namibia, Kenya and Zambia, Aryee et al. (2026) found technical change outpacing governance, with equity barriers falling hardest on women and on rural and part-time students. Outside education, Bone et al. (2025) analysed about 11 million UK vacancies; mentions of degree requirements for AI roles fell by 15% as employers shifted toward hiring on skills. The authors link this to skill shortages, not doubts about assessment, and it is the only employer-side study here, but it shows that employers can substitute other evidence for credentials.

The empirical record has clear weaknesses. Detection studies use small, purpose-built corpora and soon-superseded models. Redesign studies are single-site, usually evaluated by their designers, and seldom report effect sizes. Experience data are mostly self-reported, and African studies document adoption and policy rather than assessment outcomes.

3. MATERIALS AND METHODS

3.1 RESEARCH DESIGN

The study is a conceptual, theory-building synthesis: it collects no new data but integrates existing theory and evidence into a framework and testable propositions.

3.2 POPULATION AND SAMPLE

The population is the peer-reviewed literature on validity theory in educational measurement, academic integrity, AI-generated work, AI detection and assessment redesign. The sample is 37 sources: 35 peer-reviewed journal articles published 2023–2026, together with two foundational works, Messick (1995) and Kane (2013), cited for the theories they originated. Five of the 35 concern African settings directly and two more are African-led. Sampling was criterion-based (purposive), selecting sources for their bearing on the research questions.

3.3 INCLUSION AND EXCLUSION CRITERIA

Inclusion required a peer-reviewed journal article published 2023–2026 with a working DOI, a Crossref record of type journal-article, and no retraction, correction or expression of concern. Preprints, conference papers, books, reports, commentaries and web sources were excluded, as were articles from outlets whose peer review could not be established (several African-focused items among them) and articles whose findings could not be confirmed from a retrievable abstract. The two

© IIARJournal. All rights reserved.

foundational works are the sole exception to the date criterion.

3.4 DATA COLLECTION PROCEDURE AND TOOLS

Literature was retrieved through OpenAlex, Crossref and targeted searches of assessment, educational-technology and academic-integrity journals, run in October 2026 for publications from 2023 onward. Search terms combined generative AI and ChatGPT with validity, detection, false positives, non-native writers, oral and interactive assessment, process evidence, keystroke logging, two-lane assessment, policy, credentials and country names (Nigeria, Ghana, Kenya, South Africa, Rwanda, sub-Saharan Africa). Metadata came from Crossref and abstracts from OpenAlex, and publisher pages filled Crossref gaps.

3.5 DATA ANALYSIS

The synthesis had three steps. First, each empirical finding was coded by the Messick (1995) threat involved and by the link it threatens in an interpretation/use argument (Kane, 2013), extended here with a preceding attribution link: attribution of the observation, scoring, generalisation, extrapolation or decision. Second, each institutional response was judged by whether it acts on the threatened inference, its cost to fairness and its enforceability. Third, tasks were cross-classified by intended inference and AI exposure, and propositions written for each response comparison, grid cell or system-level boundary where the evidence suggested a predictable outcome.

3.6 BASES OF JUDGEMENT

Judgement rested on consistency with validity theory, convergence across independent studies, and explicit weighting of study quality, so that a finding from a blind field test (Scarfe et al., 2024) counted for more than one from opinion data. Propositions were retained only if they could be stated in falsifiable form.

3.7 RELIABILITY AND VALIDITY

Every reference was checked twice against Crossref and screened for retraction, correction and concern notices, with OpenAlex retraction flags as a further check. Every citation was independently re-verified against its source abstract (and full text for key figures), and claims outrunning their source were narrowed. A separate review checked that counts, claims and hedges agree across sections, and the tests that could disconfirm each proposition are set out in §4.2. The framework still inherits the weaknesses of its evidence (end of §2.3); its propositions are claims to be tested, not findings.

3.8 VARIABLE MEASUREMENT

No variables were measured directly. The two framework dimensions were defined from the literature and operationalised as categories: intended inference as four claim types, and AI exposure as three levels set by administration conditions (§4.1). Each proposition names measurable indicators, such as score gaps or false-allegation rates.

3.9 AVAILABILITY OF MATERIALS

The full reference list with DOIs is in §7. The verification log and coding notes are available from the corresponding author on request, without restriction.

3.10 AI USAGE DISCLOSURE

Generative AI tools assisted literature retrieval, drafting and formatting under the authors' direction. The authors verified all sources against publisher and Crossref records and take full responsibility for the content.

3.11 OTHER DISCLOSURES

The study involved no human participants or personal data, so no ethical approval was required. It received no external funding, and the authors declare no conflict of interest.

4. RESULTS AND DISCUSSION

4.1 RESULTS

GenAI leaves scoring largely intact but breaks the attribution of the observation to the candidate and weakens extrapolation (§2.2), so the three responses are judged by whether they repair those links. Each acts on a different part of the argument (Table 1). Detection tries to repair attribution after the fact, inferring authorship from the product, and its errors are not only frequent but unevenly distributed. False positives fall on writers whose prose has low perplexity, which in the one controlled comparison meant non-native writers (Liang et al., 2023), while false negatives favour students who know how to

© IIARJournal. All rights reserved.

evade the tools (Perkins, Roe, et al., 2024). A detection-based integrity regime therefore adds construct-irrelevant variance correlated with language background, a consequential validity failure in Messick's (1995) sense. Prohibition tries to secure attribution in advance by rule. In an unsupervised task nothing makes the rule true, so prohibition is discursive (Corbin et al., 2025), and the paper expects its main observable effect to be less disclosure rather than less use, a conjecture consistent with Luo's (2025) finding on one-way transparency (§2.1). Redesign alone changes the evidence itself, either restoring conditions under which attribution holds or changing the claim so that AI use belongs to the construct.

The three are not cleanly separable. Process evidence such as keystroke logs is redesign that collects new evidence, but used punitively it becomes detection by other means, with the same risks of surveillance and adversarial evasion (Mehta et al., 2026). A prohibition becomes structural once attached to supervised conditions. Oral assessment is redesign, yet it can introduce its own construct-irrelevant variance, linked to gender and anxiety (Rae et al., 2026) and, plausibly, to language. What matters is not a measure's label but which inference it secures, and for whom.

Table. 1. Detection, prohibition and redesign compared against the validity argument

Response	Inference it acts on	Main evidence	Fairness consequence	Enforceability	Residual validity risk
Detection (software or marker judgement)	Attribution, inferred after the fact from the product	Unreliable and easy to evade (Weber-Wulff et al., 2023; Perkins, Roe, et al., 2024); markers no better (Fleckenstein et al., 2024; Kofinas et al., 2025)	False positives concentrated among non-native writers (Liang et al., 2023)	Only probabilistic; contestable in appeals	Misattribution plus biased misconduct decisions
Prohibition (rules, permission scales)	Attribution, assumed by rule	Rules students can ignore; "enforcement illusion" (Corbin et al., 2025)	Rewards non-disclosure; one-way transparency (Luo, 2025)	Low for unsupervised tasks	Unknown share of AI-generated evidence
Redesign: relocation (supervised, oral)	Attribution, restored by conditions	Oral formats resist misuse (Davey et al., 2025)	Equity gains possible (Nyamadzawo, 2025); gender and anxiety risks (Rae et al., 2026)	High, at a cost in staff time	Narrower sampling of the construct
Redesign: reconstruction (assess AI use)	Construct and extrapolation, redefined	Evaluative judgement as a target (Bearman et al., 2024); process traces (Yang et al., 2026)	Depends on equal access to tools	High, because use is expected	Weak rubrics for quality of AI use

The framework crosses two dimensions. The first is the *intended inference*, the claim the result is meant to support. Four types cover most educational uses: unaided capability ("can solve this without help", as in foundational numeracy, licensure knowledge or language proficiency); reasoning and understanding ("can explain, justify and adapt"); authentic production ("can produce work of professional quality under realistic conditions"); and accountable AI use ("can direct, evaluate, verify and take responsibility for AI-assisted work"). The second, *AI exposure*, is graded low (supervised, device-controlled conditions), moderate (staged or partly observed work with process evidence or a follow-up conversation) or high (unsupervised work judged on the product alone). The cells yield four decisions

© IIARJournal. All rights reserved.

(Table 2): retain, relocate, reconstruct or reduce the stakes.

Table. 2. Task classification by intended inference and AI exposure

Intended inference	Low exposure (supervised)	Moderate exposure (staged, process evidence, follow-up dialogue)	High exposure (unsupervised, product only)
Unaided capability	**Retain**: invigilated papers, practicals, computer-based tests in controlled centres	Retain if the dialogue checks the capability itself	**Relocate** to supervised conditions, or reduce stakes
Reasoning and understanding	**Retain**: structured orals, in-class problem solving	**Retain or relocate**: brief viva attached to written work	**Relocate**: the product says little about whose reasoning it shows
Authentic production	Often artificial; limited extrapolation	**Reconstruct**: staged artefacts with process evidence and defence	**Reconstruct**, or treat as formative
Accountable AI use	Rarely meaningful	**Reconstruct**: AI use required, documented and evaluated	**Reconstruct**: the AI-permitted task is valid only if the quality of use is what gets scored

The framework locates the problem in claims, not formats. A take-home essay is not invalid in itself; it is invalid as evidence of unaided reasoning, and potentially valid as evidence of accountable AI use if assessed that way. Curtis's (2025) objection to two-lane models fits here: the moderate-exposure column is the middle lane, where most reconstruction happens. In these terms, validity is judged cell by cell, authorship enters as the attribution link that exposure threatens, and trust is the downstream stake taken up in P7.

Eight propositions follow from the response comparison and the framework.

P1 (divergence). For tasks meant to support claims about unaided capability or reasoning, high-exposure scores will diverge increasingly from the same students' supervised scores as GenAI access grows. AI submissions already outscore real students in unsupervised examinations (Scarfe et al., 2024), so inflated, decoupled scores are the expected signature.

P2 (detection inequity). Where detection outputs inform misconduct decisions, the rate of false allegations will be higher for L2 writers of English than for first-language writers, and higher in systems where most candidates are L2 writers. The perplexity mechanism reported by Liang et al. (2023) implies the first part; the system-level part is an untested extension.

P3 (prohibition and disclosure). In unsupervised tasks, prohibition without structural change will not lower actual GenAI use relative to permitted conditions, but will lower disclosure. This rests on Corbin et al.'s (2025) analysis of discursive change and is consistent with Luo (2025).

P4 (retention). Supervised tasks targeting unaided capability or reasoning will show stable relations with external criteria before and after GenAI became widespread. This follows from the exposure dimension rather than direct evidence; Kofinas et al. (2025) likewise recommend performative, process-based tasks over asynchronous written output.

P5 (structured relocation). Short, structured oral components attached to written work will raise agreement between assessed and independently demonstrated capability without adverse differences by language background or gender, provided they are blueprinted and the interaction standardised. Unstructured orals will not. The mixed evidence makes this testable (Davey et al., 2025; Nyamadzawo, 2025; Rae et al., 2026).

P6 (reconstruction). Where professional practice is saturated with AI, tasks that score the direction, verification and evaluation of AI output will predict later performance better than product-only grading of the same tasks. It follows from Bearman et al.'s (2024) account of evaluative judgement and from the different quality of AI use visible in process traces (Yang et al., 2026).

P7 (credential discounting). The larger the share of a qualification's weight carried by high-exposure,

© IIARJournal. All rights reserved.

unreconstructed tasks, the more employers and receiving institutions will discount it and substitute their own testing. The shift away from degree requirements in AI-intensive hiring is consistent with this, although the single UK study attributes it to skill shortages rather than doubts about assessment (Bone et al., 2025).

P8 (localisation in examination-centred systems). In systems dominated by terminal national examinations, validity risk will concentrate in coursework, continuous assessment, project and research components and in breaches of examination security, rather than in the supervised paper itself. These should be the first targets of relocation or reconstruction. It follows from classifying invigilated papers as low exposure; no peer-reviewed study yet tests it (see §4.2).

4.2 DISCUSSION

On RQ1, the analysis supports treating GenAI as a break in the attribution and extrapolation links of Kane's (2013) interpretation/use argument, not as a rise in misconduct. This agrees with Dawson et al. (2024) and Weidlich (2026) and applies Messick's (1995) consequential view by treating biased misconduct findings as a validity failure. For RQ2, detection and prohibition both act on attribution without changing the evidence, which fits the convergent detection and marker studies in §2.3 (Weber-Wulff et al., 2023; Perkins, Roe, et al., 2024; Kofinas et al., 2025). The analysis qualifies rather than rejects the redesign literature: relocation secures attribution (Davey et al., 2025) but narrows sampling and costs staff time, and its fairness depends on structure (Rae et al., 2026).

Turning to RQ3, the framework in Table 2 adds two things at task level to Weidlich's (2026) procedure: an exposure dimension, which explains why the same format can be secure in one setting and compromised in another, and a decision rule linking each inference to an action. Unlike permission scales such as the AIAS (Perkins, Furze, et al., 2024), it classifies tasks by what the result is meant to show rather than by permitted AI use, so the permitted level follows from the claim and, because each level is tied to administration conditions, the change is structural rather than discursive (Corbin et al., 2025). Its reconstruction cells adopt, rather than test, Bearman et al.'s (2024) case for evaluative judgement of AI outputs and processes.

As for RQ4, African systems sharpen every element of the analysis. Detection is least defensible where most tempting: in Nigeria, Ghana and Kenya many students are L2 writers of English, the group for which P2 predicts more false allegations, although the bias Liang et al. (2023) found has not been tested with these writers. Imported detectors therefore risk turning measurement error into misconduct findings, at greatest cost where appeals are weak. On the framework's logic (P8), supervised papers such as WAEC's West African Senior School Certificate Examination (WASSCE), the NECO Senior School Certificate Examination (SSCE) and JAMB's Unified Tertiary Matriculation Examination (UTME) are low-exposure by design, so their main AI-related threat would come through breaches of examination security, which a phone with GenAI access could make faster to exploit, a possibility no peer-reviewed study has examined. Policy readiness is uneven in South Africa (Jacobs-Basadien & Pather, 2026), and Ghana's innovation has proceeded in a policy vacuum, Rwanda's under a strong central vision constrained by institutional rigidity (Aryee et al., 2026). Resources are also scarce: large classes make one-to-one orals costly, although Nyamadzawo's (2025) reflective study in Durban suggests that structured orals can work as a translanguaging and equity tool, and reconstruction assumes equal access to AI tools, where Essien et al. (2024) and Aryee et al. (2026) both document infrastructural and equity barriers. Finally, wherever public confidence in certificates is fragile, P7 implies that unreformed coursework will add an AI discount to it; the paper expects that discount to fall on graduates rather than institutions.

None of the eight propositions is accepted or rejected here; all are offered for empirical testing. P1 and P4 can be tested by tracking the same students' supervised and unsupervised scores, and their relations with external criteria, as GenAI spreads: a gap that does not widen would count against P1, and weakening criterion relations for supervised tasks against P4. P2 calls for audits of detector-informed allegations by language background, P3 for comparisons of disclosed and process-traced use under prohibited and permitted conditions, P5 and P6

© IIARJournal. All rights reserved.

for multi-site trials of structured orals and AI-use rubrics with subgroup analyses, P7 for studies of employer screening and admissions practice, and P8 for audits of where integrity breaches occur across components. Future research should also repeat detection and capability studies with current models and with African L2 writers, report effect sizes for redesign, and develop rubrics for the quality of AI use, the weakest point in Table 1.

5. CONCLUSION AND RECOMMENDATION

GenAI has not made assessment impossible, but it has made much assessment evidence ambiguous. Treating it as cheating leads toward detection, which has proved unreliable and has been shown to misclassify non-native English writing, and toward prohibition, which cannot be enforced in unsupervised work. Treating it as a validity problem asks which claim each task supports, and under what conditions, and leads to four actions: retain, relocate, reconstruct or reduce the stakes.

Examination bodies (WAEC, NECO, JAMB and counterparts). Keep supervised papers as the anchor of high-stakes claims. Audit coursework and school-based components with the inference-by-exposure grid and reduce the weight of high-exposure ones. Strengthen device control and item security in computer-based centres, and publish the claim each component supports.

Universities. Do not use detector scores as evidence of misconduct. Map every module's assessments onto Table 2. Attach brief, blueprinted vivas to projects and theses. Build at least one assessment per programme in which accountable AI use is the construct. Publish marking criteria for AI-assisted work, so that transparency runs in both directions.

Quality-assurance agencies (Nigeria's National Universities Commission and its counterparts). Require institutions to set out an assessment-validity case for each programme in place of AI-detection procedures, fund staff development for oral and process-based assessment, and commission multi-site studies to test P1–P8 in African settings.

6. CONTRIBUTION TO EXISTING KNOWLEDGE

Theoretically, the paper reframes AI and assessment integrity in validity terms, compares detection, prohibition and redesign as validity strategies, and builds on Weidlich's (2026) argument-based treatment of candidate-side GenAI use by making AI exposure a graded dimension, distinct from format, crossed with the intended inference. Practically, it gives institutions a decision grid to apply module by module. For research, it states eight falsifiable propositions on assessment redesign and credential trust and identifies coursework components in examination-centred African systems as a neglected priority.

7. REFERENCES

Aryee, J. N. A., Davies, P., Torsah, G. A., Apaw, M. M., Mwando, S. M., Kwisanga, C., Boateng, C. D., Jobunga, E., Mutale, D., & Amekudzi, L. K. (2026). Rethinking AI integration in African higher education: A mixed-methods investigation of capacity, equity, and governance across sub-Saharan Africa. *Frontiers in Education*, *11*, Article 1968304. https://doi.org/10.3389/feduc.2026.1968304

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

Bone, M., González Ehlinger, E., & Stephany, F. (2025). Skills or degree? The rise of skill-based hiring for AI and green jobs. *Technological Forecasting and Social Change*, *214*, Article 124042. https://doi.org/10.1016/j.techfore.2025.124042

Chaka, C. (2024). Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review. *Journal of Applied Learning & Teaching*, *7*(1). https://doi.org/10.37074/jalt.2024.7.1.14

Corbin, T., Dawson, P., & Liu, D. (2025). Talk is cheap: Why structural assessment changes are needed for a time of GenAI. *Assessment &*

© IIARJournal. All rights reserved.

Evaluation in Higher Education, *50*(7), 1087–1097. https://doi.org/10.1080/02602938.2025.2503964

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Curtis, G. J. (2025). The two-lane road to hell is paved with good intentions: Why an all-or-none approach to generative AI, integrity, and assessment is insupportable. *Higher Education Research & Development*, *44*(8), 2151–2158. https://doi.org/10.1080/07294360.2025.2476516

Davey, S. K., Birbeck, D., Nallaya, S., Sallows, G., & Della Vedova, C. B. (2025). Utilising one-on-one interactive oral assessments as the major final assessment within a bioscience course. *Assessment & Evaluation in Higher Education*, *50*(7), 1154–1171. https://doi.org/10.1080/02602938.2025.2502577

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

Eaton, S. E. (2023). Postplagiarism: Transdisciplinary ethics and integrity in the age of artificial intelligence and neurotechnology. *International Journal for Educational Integrity*, *19*(1), Article 23. https://doi.org/10.1007/s40979-023-00144-1

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, *19*(1), Article 17. https://doi.org/10.1007/s40979-023-00140-5

Essien, A., Salami, A., Ajala, O., Adebisi, B., Shodiya, A., & Essien, G. (2024). Exploring socio-cultural influences on generative AI engagement in Nigerian higher education: An activity theory analysis. *Smart Learning Environments*, *11*(1), Article 63. https://doi.org/10.1186/s40561-024-00352-3

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S. D., Köller, O., & Möller, J. (2024). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers and Education: Artificial Intelligence*, *6*, Article 100209. https://doi.org/10.1016/j.caeai.2024.100209

Gorichanaz, T. (2023). Accused: How students respond to allegations of using ChatGPT on assessments. *Learning: Research and Practice*, *9*(2), 183–196. https://doi.org/10.1080/23735082.2023.2254787

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice*, *43*(2), 16–29. https://doi.org/10.1111/emip.12602

Jacobs-Basadien, M., & Pather, S. (2026). Assessing AI policy and guideline progress within the South African public higher education system: Insights from documented policies. *Universal Access in the Information Society*, *25*(2), Article 71. https://doi.org/10.1007/s10209-026-01340-9

Kaldaras, L., Akaeze, H. O., & Reckase, M. D. (2024). Developing valid assessments in the era of generative artificial intelligence. *Frontiers in Education*, *9*, Article 1399377. https://doi.org/10.3389/feduc.2024.1399377

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

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

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, *4*(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

© *IIARJournal. All rights reserved.*

Luo, J. (2025). How does GenAI affect trust in teacher-student relationships? Insights from students' assessment experiences. *Teaching in Higher Education*, *30*(4), 991–1006. https://doi.org/10.1080/13562517.2024.2341005

Mehta, A., Kumar, R., Singla, A., Bisht, K., Singla, Y. K., & Shah, R. R. (2026). Detecting LLM-assisted academic dishonesty using keystroke dynamics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *8*(5), 754–768. https://doi.org/10.1109/tbiom.2026.3683979

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), 741–749. https://doi.org/10.1037/0003-066x.50.9.741

Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open*, *5*, Article 100151. https://doi.org/10.1016/j.caeo.2023.100151

Nguyen Thanh, B., Vo, D. T. H., Nguyen Nhat, M., Pham, T. T. T., Thai Trung, H., & Ha Xuan, S. (2023). Race with the machines: Assessing the capability of generative AI in solving authentic assessments. *Australasian Journal of Educational Technology*, *39*(5), 59–81. https://doi.org/10.14742/ajet.8902

Nikolic, S., Daniel, S., Haque, R., Belkina, M., Hassan, G. M., Grundy, S., Lyden, S., Neal, P., & Sandison, C. (2023). ChatGPT versus engineering education assessment: A multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. *European Journal of Engineering Education*, *48*(4), 559–614. https://doi.org/10.1080/03043797.2023.2213169

Nyamadzawo, R. T. (2025). Oral versus written assessments: A reflective study on promoting equality in evaluations in higher education. *Educational Research for Social Change*, *14*(2), 14–33. https://doi.org/10.17159/2221-4070/v14n2a2

Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice*, *21*(6). https://doi.org/10.53761/q3azde36

Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., & Khuat, H. Q. (2024). Simple techniques to bypass GenAI text detectors: Implications for inclusive education. *International Journal of Educational Technology in Higher Education*, *21*(1), Article 53. https://doi.org/10.1186/s41239-024-00487-w

Rae, M., Tai, J., & Dawson, P. (2026). Giving voice to women students: Designing oral assessments for inclusion and validity. *Assessment & Evaluation in Higher Education*, *51*(6), 1083–1096. https://doi.org/10.1080/02602938.2025.2580622

Scarfe, P., Watcham, K., Clarke, A., & Roesch, E. (2024). A real-world test of artificial intelligence infiltration of a university examinations system: A "Turing Test" case study. *PLOS ONE*, *19*(6), Article e0305354. https://doi.org/10.1371/journal.pone.0305354

Thaldar, D., Botes, M., Badru, A., Chenia, H., Duma, S., Dlamini, S. B., Amin, N., Hugo, W., Govender, R., Bruce-Brand, J., Vosloo, A., Koorbanally, N. A., & Chuturgoon, A. (2025). Generative AI governance in higher education: A case study from Africa. *Frontiers in Political Science*, *7*, Article 1666661. https://doi.org/10.3389/fpos.2025.1666661

Walters, W. H. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science*, *7*(1), Article 20220158. https://doi.org/10.1515/opis-2022-0158

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for*

© IIARJournal. All rights reserved.

Educational Integrity, *19*(1), Article 26. https://doi.org/10.1007/s40979-023-00146-z

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

Yang, K., Cheng, Y., Zhao, L., Raković, M., Swiecki, Z., Gašević, D., & Chen, G. (2026). Ink and algorithm: Exploring temporal dynamics in generative AI-assisted writing. *British Journal of Educational Technology*, *57*(2), 508–534. https://doi.org/10.1111/bjet.70031

Yusuf, A., Pervin, N., & Román-González, M. (2024). Generative AI and the future of higher education: A threat to academic integrity or reformation? Evidence from multicultural perspectives. *International Journal of Educational Technology in Higher Education*, *21*(1), Article 21. https://doi.org/10.1186/s41239-024-00453-6

© IIARJournal. All rights reserved.