IIJP 2026 | Volume 2, Issue 1, pp. 1360-1372 | ISSN: 2636-4484

RETHINKING ASSESSMENT IN SCIENCE EDUCATION IN THE AGE OF GENERATIVE AI

Gabriel Emudiaga Esakpaide
Department of Science Education, Delta State University, Abraka, Nigeria
mudnero@gmail.com/+2348035494590

ARTICLE INFO
Article No.: 0233
Accepted Date: 12/03/2026
Published Date: 29/03/2026
Type: Research

ABSTRACT
Generative AI has disrupted educational assessment by threatening the inferential basis on which validity depends. When submitted work may represent either authentic reasoning or algorithmic output, assessment cannot support valid inferences about competence. Detection and prohibition approaches fail technically, conceptually, and behaviourally: students are driven by performance expectations, not policy. This paper proposes a four-principle framework for restoring assessment validity in AI-rich environments, grounded in validity theory (Kane, 2006, 2013; Messick, 1989) and contextualised within Nigerian science education. The framework integrates established traditions of authentic assessment, epistemic documentation, and equitable design, recontextualised under generative AI conditions in a resource-constrained setting where existing frameworks assume technological access that most Nigerian schools do not possess. The framework makes reasoning visible and evaluable through task design and proportional documentation, removing substitution incentives rather than attempting identification. Its validation rests on theoretical coherence and contextual appropriateness, pending empirical testing. Critically, the framework measures reasoning articulation (students' capacity to explain their thinking) as a proxy for reasoning execution (actual cognitive work), a shift that requires explicit acknowledgement. For Nigerian science education, it provides concrete operationalisation of assessment principles aligned with curriculum policy while directly addressing the AI validity crisis.

Keywords: Generative AI, science education, assessment validity, academic integrity, AI literacy, Nigerian education

Introduction

Generative AI has undermined a foundational assumption of educational assessment: that observed student performance reflects what the student knows and can do. Advanced language models (including ChatGPT, Claude, and Gemini) now produce academically credible responses technically indistinguishable from authentic student work (Garzon, Baldiris & Fabregat, 2025; Holmes, Persson, & Chounta, 2022; Luckin, 2023). Stribling, Clifton and Maclean (2024) demonstrated that GPT-4 performs at doctoral-level competency on graduate science examinations, while West, Shebab and Williamson (2023) documented laboratory reports sufficiently coherent to pass as authentic student submissions. When submitted work could equally represent genuine reasoning or algorithmic generation, assessment can no longer support the inferences it claims to make about student competence. The question is not how to detect AI use more effectively, but how to redesign assessment so that valid inferences remain possible regardless of which tools students employ.

The urgency of this question is acute in Nigerian educational institutions. A survey of 308 education students at Kwara State University found that 50.3% predominantly use AI tools for academic purposes, with 64.3% demonstrating high reliance (Adesokan, Salman & Raheem, 2025). Simultaneously, 90% of Nigerian academics reported the absence of institutional policies guiding AI use in academic writing, while 95% rejected AI as a co-author (Nwali & Udumukwu, 2025). A concurrent survey of 289 students at a northern Nigerian state university found that 77.9% received no guidance from lecturers or institutions on AI use, confirming that the institutional gap is national in scope (Yakubu, David & Abubakar, 2025). This disjunction between widespread adoption and absent guidance creates both practical and ethical urgency for frameworks that address, rather than prohibit, AI integration.

Science education occupies a distinctive position in this landscape. Unlike disciplines centred on factual recall, science prioritises reasoning, investigation, and evidence-based explanation (Kind & Osborne, 2017; Osborne & Dillon, 2008). Assessment is designed to capture authentic scientific thinking, not memorised content (Holmes, Bailik & Fadel, 2019). The National Research Council (2012) grounded scientific literacy in a knowledge-in-use paradigm integrating disciplinary core ideas, scientific practices, and crosscutting concepts, none of which is captured by selected-response formats measuring one-dimensional retrieval (Pellegrino, DiBello & Goldman, 2016). Generative AI poses a specific threat because it can simulate scientific reasoning without any underlying cognitive work.

The four principles proposed here are not individually novel. Authentic assessment (Wiggins, 1998), process portfolios (Darling-Hammond, Ancess & Falk, 2013), reflective documentation (Schon, 1983), and equitable design (Universal Design for Learning, 2023) each have established literatures. The framework's contribution is integrative: it reconceptualises these traditions as validity-restoring mechanisms under generative AI conditions in a resource-constrained national context where prior frameworks either assume technological access unavailable in most Nigerian schools, or address AI as an integrity problem rather than a validity problem. This distinction drives all four principles.

Methodology

This paper presents a conceptual framework grounded in narrative literature review and contextual analysis, intended for empirical validation rather than immediate implementation. Sources were identified through Google Scholar, ERIC, and Scopus using search terms including generative AI assessment validity, AI detection higher education, authentic assessment science, and Nigerian science education. Initial searches returned approximately 340 sources across all terms; 61 were selected for citation based on direct relevance to the framework's theoretical claims, empirical grounding in Nigerian or comparable resource-constrained contexts, and recency (prioritising peer-reviewed studies published between 2010 and 2025). Sources were excluded where they addressed AI in education without bearing on

assessment validity, or where empirical claims were insufficiently reported to support the argument. Several 2025 citations are preprints or manuscripts under review; where this is the case, claims drawn from them are treated as indicative rather than definitive.

Three safeguards address the confirmation bias inherent in purposive review. First, sources were actively sought that challenge the framework's central claims, including literature on pre-AI resistance to authentic assessment (Bennett, 2011; Hargreaves, 2005; Supovitz, 2009), the limitations of oral assessment at scale (Pearce & Chiavaroli, 2020), and the persistence of undeclared AI use under disclosure mandates (Pérez-Pérez et al., 2026). Second, the framework is treated as a normative proposal requiring empirical testing; claims are stated as principled positions rather than proven outcomes. Third, explicit falsification criteria are stated: the framework would require revision if authentic assessment tasks do not demonstrably improve validity of inferences about scientific reasoning in Nigerian classrooms, if documentation requirements prove more gameable than anticipated, or if professional development investment proves unsustainable at institutional scale.

Nigerian Educational Context

The structural realities of Nigerian science education shape every aspect of the framework's design. The Federal Ministry of Education (2015) mandates inquiry-based learning and authentic assessment, yet national examinations remain dominated by selected-response items measuring content recall rather than investigative reasoning (Eze & Obi, 2019; Ojimba, 2023). A systematic review of 40 studies on AI in Nigerian education found that 65% of rural schools lack consistent electricity supply and most secondary schools experience poor internet connectivity (Ibitoye et al., 2025). The World Bank (2023) reports that internet penetration remains between 35 and 40% with pronounced urban-rural disparities, while UNESCO's (2023) survey reveals that fewer than 10% of schools and universities have formal guidance on AI use in academic contexts. Within Nigerian tertiary institutions, high student-staff ratios, limited learning resources, and uneven digital capacity constrain learner-centred pedagogies while complicating timely diagnostic assessment (Usman et al., 2025). These constraints rule out detection-based and technology-dependent responses a priori, and make contextually grounded redesign the only viable approach.

Assessment Validity and the Generative AI Problem

Assessment validity depends on whether interpretations of performance support defensible claims about competence (Messick, 1989). Kane's (2006, 2013) argument-based framework specifies that valid inferences require coherent inferential chains from observable performance to intended constructs. Generative AI disrupts this chain through two mechanisms.

First, AI detection reliability varies substantially by task type: detection tools achieve F1 scores of 0.90-0.95 on simple classification tasks but only 0.61-0.82 on tasks requiring novel application (Wang, Xia & Ye, 2025; Yan et al., 2024). These figures measure the accuracy of detection systems, not AI task performance directly; the implication is that tasks demanding novel reasoning are both harder for AI to execute convincingly and harder for detection systems to flag reliably, creating a design opportunity at their intersection. Second, synthetic fluency (the production of coherent, credible-sounding text that masks shallow or absent understanding) creates a validity threat when assessment evaluates submitted outputs in isolation. Cotton, Cotton and Shipway (2023) demonstrated that ChatGPT generates academically acceptable work across disciplines without genuine comprehension. Perkins and Roe (2024), examining inductive thematic analysis conducted by ChatGPT, found that stochastic variability across model versions produced different results from identical data, pointing to a deeper indeterminacy: when the same analytical work could be performed by either human or machine, assessment cannot distinguish between them even in principle. In Nigerian contexts, Ya'u and Mohammed (2025) found that 75% of students used AI primarily

for surface-level tasks, with only a moderate correlation (r = 0.45) between AI use and performance, suggesting that performance metrics have begun to lose capacity to reflect actual learning when assessment design permits fluent substitution without reasoning documentation.

Synthetic fluency becomes pedagogically irrelevant, however, when assessment shifts to evaluating reasoning processes: a fluent submission paired with documented evidence of authentic reasoning demonstrates competence regardless of tool use. The validity problem that synthetic fluency creates is therefore a problem of assessment design, not an inherent feature of generative AI. Valid assessment must replace unverifiable authenticity with epistemic transparency as its governing criterion (Lincoln & Guba, 1985): assessment cannot verify authorship, but it can systematically document how students arrived at their claims regardless of which tools they used.

The technical case against detection-based responses is substantial. ChatGPT bypasses plagiarism detection systems (Gill et al., 2023); detection tools show substantial bidirectional classification errors (Erol et al., 2025); and LLM-generated responses frequently evade detection while receiving grades indistinguishable from authentic work (Scarfe, Watchirn & Gibbs, 2024). Most critically, GPT-4o achieved only 67.7% accuracy distinguishing its own generated text from human text (Abbas, 2025). Peterson (2025) demonstrated that probabilistic generation means detection signatures shift as model datasets update continuously. In Nigerian contexts, traditional detection tools suffer from high false-positive rates for non-native English writers, and 80% of academics expressed concern about data privacy when AI detection systems process unpublished manuscripts without clear safeguards (Nwali & Udumukwu, 2025). Conceptually, even perfect detection would not restore assessment validity. Salaudeen et al. (2025) distinguish measurement (the assignment of values to observable properties) from validity (which refers to whether evidence and theory support interpretations of those values for their proposed uses). When a task permits multiple causal pathways to the same observed performance (authentic reasoning or algorithmic generation), the inferential chain becomes indeterminate. Flawless identification of AI use would not restore the lost inferential capacity. For construct-targeted claims asserting scientific understanding, assessment must restore observational clarity through tasks where authentic reasoning and algorithmic generation produce sufficiently different performances that inference becomes again possible.

Deterrence-based approaches fail not only technically and conceptually, but behaviourally. Structural equation modelling among Nigerian university students found that perceived risk of academic consequences exerted no significant effect on intention to use generative AI, while performance expectancy was the strongest predictor (Yakubu et al., 2025). Attitude toward generative AI had zero effect on behavioural intention. Students are driven by performance expectations, not values or guidelines. This finding directly challenges the assumption underlying values-based guidance frameworks (European Network for Academic Integrity, 2024; University College Cork, 2025): if attitude and consequence awareness have no empirical effect on behaviour, then transparency expectations and ethics instruction alone cannot resolve the validity crisis. The persistent optimism in institutional guidance documents about values-based compliance also encounters the structural critiques raised by Hargreaves (2005) and Supovitz (2009): assessment reform fails not because frameworks are theoretically unsound, but because institutional incentive structures reward compliance behaviours rather than genuine reasoning. Bennett's (2011) critique that formative assessment gains are frequently shallow and unsustained applies equally here: documentation requirements that are appended rather than embedded in assessable criteria will be gamed rather than engaged. The framework responds to these critiques by designing documentation as an intrinsic component of the assessment task, not as a supplementary policy obligation, and by making the performance advantage of authentic reasoning explicit rather than relying on students' values alignment.

A Four-Principle Framework for Assessment in the Age of Generative AI

The proposed framework addresses assessment's inferential foundation directly, making reasoning visible and evaluable through task design and documentation regardless of which tools students use. It differs from detection approaches (Erol et al., 2025), institutional policy frameworks (Banta & Palomba, 2014), and existing AI literacy development programmes in three operationally distinct ways.

Recent AI literacy frameworks, such as the ED-AI Lit framework (Holmes, Tuomi & Kamarainen, 2022) and AI4K12 initiative (Touretzky et al., 2019), do embed contextual understanding, asking students to recognize how AI systems function across domains and what ethical implications arise from their use. However, this framework differs operationally. First, AI literacy curricula target understanding of AI mechanisms and sociocultural implications, while this framework targets demonstrated reasoning competence in disciplinary domains, assessed through epistemic transparency rather than AI comprehension. Second, students may develop sophisticated AI literacy yet still attempt to substitute algorithmic output when assessment tasks permit; conversely, students may use AI strategically while remaining poorly informed about its mechanisms. These are separate competencies. Third, Yakubu et al. (2025) found that attitude toward generative AI and perceived understanding of its capabilities had zero effect on behavioural intention to use it; only performance expectancy mattered. This framework sidesteps attitude-change approaches by making the performance advantage accrue to demonstrated reasoning, not to AI use. A student who reasons thoroughly (with or without AI consultation) receives higher assessment credit than a student who substitutes algorithmic output, regardless of their AI literacy level.

This framework complements rather than competes with critical AI education scholarship (Selwyn, 2019, 2022; Macgilchrist, 2021), which argues that technical solutions alone are insufficient to address AI's social and political embeddedness in education. Both this framework and critical perspectives reject techno-utopian assumptions. This framework's contribution is operationally specific: it proposes mechanisms for assessment design that restore valid inference about scientific reasoning when AI tools are available, functioning independently of students' AI literacy level and equally in contexts with or without explicit AI literacy instruction.

Principle 1: Expanded Validity Constructs

Assessment validity under generative AI requires reconceptualising what constitutes evidence of competence. Product-based validity evidence is insufficient when generative AI produces indistinguishably authentic-seeming artefacts. Salaudeen et al. (2025) distinguish criterion-aligned evidence (where the object of claim and measured object are identical), criterion-adjacent evidence (where a proxy is measured to infer a different criterion), and construct-targeted evidence (where proxies measure abstract constructs). For construct-targeted claims asserting scientific understanding, validity restoration requires establishing explicit nomological networks, mapping competence constructs to observable indicators that generative AI cannot plausibly replicate: troubleshooting unexpected experimental results within local material contexts, communicating findings to non-specialist audiences in contextually appropriate language, and recognising when standard solutions require adaptation to local epistemic conditions.

Expanded constructs must shift the burden of proof from what students submit to how they reason. Epistemic transparency (documented evidence of how students know what they claim to know) becomes the primary validity mechanism. This represents an intentional and significant shift in measurement: the framework measures reasoning articulation (students' capacity to explain and justify their thinking processes) as a proxy for reasoning execution (whether students actually engaged in genuine problem-solving). This is defensible but

important to make explicit. Traditional assessment asks: "Did the student produce a correct answer?" under the assumption that production itself demonstrates reasoning. Under generative AI, this assumption breaks. The framework instead asks: "Can the student articulate a coherent reasoning process, and does that articulation demonstrate engagement with evidence and authentic cognitive work?" The validity inference shifts from "correct product = reasoning occurred" to "articulated reasoning process + documented alternative consideration + explicit constraint acknowledgment = authentic intellectual engagement occurred." Importantly, articulation is necessary but not sufficient proof of execution. Teachers must exercise professional judgment about whether documented reasoning is authentic or fabricated. However, documented reasoning processes create observational clarity that final products alone cannot provide.

A second expansion recognises productive human-AI collaboration as a legitimate scientific competency, assessed through demonstrated metacognitive control rather than through the AI literacy goal of general AI awareness. Adesokan et al. (2025) examined 150 Nigerian secondary students and found that those receiving explicit instruction in tiered AI involvement could provide detailed, coherent accounts distinguishing AI-generated responses from their own synthesis. However, the study did not disaggregate outcomes by achievement level or socioeconomic status, leaving open whether metacognitive awareness equally benefits lower-achieving students and whether equity gaps narrow or widen. Usman et al. (2025) tested a tiered framework in two Nigerian tertiary institutions, reporting increased student attainment and higher pass rates compared to prohibition-based approaches. Critically, this study similarly did not report disaggregated outcomes by prior achievement or institutional resource levels, a gap given the framework's equity claims. Whether expanded construct assessment narrows achievement gaps, and for whom, remains an empirical question requiring explicit investigation. The distinction from AI literacy is operational: the framework does not ask whether students understand AI, but whether they can articulate reasoning they performed independently of it, with the caveat that articulation capacity may be unevenly distributed across student populations.

These constructs are operationalised through four observable proficiency levels: (1) claim generation: stating a relationship without explanation; (2) elaboration: describing a phenomenon with supporting detail, though mechanism remains unclear; (3) evidence-based reasoning: connecting observations to theoretical frameworks with observable evidence; and (4) evidence-based reasoning with limitations: applying understanding to novel contexts while explicitly acknowledging constraints. Meta-analytic evidence from Arifin et al. (2025), examining 25 studies on inquiry-based learning, demonstrates a substantial effect size (d = 1.27) for critical thinking development when assessment emphasises these observable practices. Okafor et al. (2018), in a national study of inquiry-based science teaching in Nigerian secondary schools, found that programmes emphasising explicit reasoning dimensions and diagnostic feedback generated significantly greater achievement gains than conventional approaches. However, these findings predate generative AI and assume no AI-mediated substitution. Whether they transfer to assessment contexts where AI tools are available requires direct empirical testing.

Sample Rubric: Heat Transfer in Physics

Level 1: States relationship without explanation (e.g., copper conducts heat).

Level 2: Describes phenomenon with supporting detail; mechanism unclear (e.g., particles transfer energy).

Level 3: Explains mechanism with observable evidence (e.g., electrons transfer energy; visible as rapid heating).

Level 4: Explains mechanism with observable evidence, acknowledges limitations, applies to novel contexts.

Principle 2: Redesigned Assessment Tasks

Task redesign works through two complementary mechanisms. First, tasks requiring novel application occupy the zone where AI detection accuracy falls and where AI execution itself becomes less reliable, since such tasks demand contextually situated judgement that probabilistic pattern-matching cannot dependably reproduce (Yan et al., 2024). Second, documentation of reasoning processes (Principle 3) renders any substitution procedurally visible. Neither mechanism is sufficient alone: task difficulty without documentation allows well-written AI solutions to remain undetectable; documentation without task difficulty requires assessors to distinguish authentic from fabricated reasoning logs. Combined, they restore valid inference about competence regardless of whether AI substitution is technically possible.

Extended-response tasks demanding evaluation and justification resist substitution because they are tied to students' prior knowledge and problem-solving trajectories, restoring the evidential link between performance and understanding (Herman & Lara-Steidel, 2025). Under generative AI, authentic assessment shifts from pedagogically preferable to epistemologically mandatory. This requirement aligns with the Federal Ministry of Education (2015) curriculum mandate for inquiry-based learning while accommodating the practical reality of Nigerian classrooms where average enrolment exceeds 50 students. Yakubu et al. (2025) identified performance expectancy as the strongest predictor of Nigerian students' intention to use generative AI. When task designs permit wholesale substitution, they structurally reward this expectation; redesigned tasks that require reasoning specificity remove the condition that makes the expectation valid. Ateeq et al.'s (2024) PLS-SEM study in Bahrain found that educational impact, operationalised as curriculum quality, teaching methodology, and active student engagement, was the strongest predictor of academic outcomes ($\beta = 0.490$, $p < .001$), substantially outperforming policy and ethics variables.

Task Redesign Exemplar: Physics

Traditional task: "Calculate the speed of sound at 20°C using v = 331 m/s." Redesigned task: "A student in a busy Lagos market notices that the pitch of a danfo horn sounds higher as the vehicle approaches and lower as it departs. (1) Explain why the perceived pitch changes, referencing the Doppler effect. (2) The student estimates the frequency shift is approximately 10%; calculate the vehicle's speed. (3) Ambient market noise, temperature variation between morning and afternoon, and vehicle acceleration would each affect the perceived shift differently. For each factor, explain whether it would increase or decrease the measured shift and justify your reasoning." This task requires novel local application, multi variable interpretation, and explicit reasoning about real-world assumptions, resisting wholesale AI substitution.

Principle 3: Transparent Documentation

Assessment validity under generative AI depends on making visible how students arrive at conclusions. Drawing on portfolio assessment traditions (Darling-Hammond et al., 2013) and reflective practice models (Schon, 1983), the framework requires explicit documentation of reasoning alongside final work. Epistemic transparency (making explicit the knowledge-making processes, sources of evidence, and alternative interpretations considered) allows assessors to evaluate authentic intellectual engagement. LLM judges frequently conflate plausible output with authentic authorship when only final products are available (Shi et al., 2025); only documented reasoning processes allow assessors to shift from identifying output origins to evaluating thinking quality. Documentation requirements scale proportionally to task demands. Procedural tasks require minimal documentation (e.g., a one-sentence note on what proved difficult). Interpretive tasks require a design log noting patterns and alternatives considered. Investigative tasks require documentation capturing how students framed

questions, identified variables, and resolved confounds; a one-to-two-page log suffices. In national examinations, 10-15 minute annotations alongside extended-response items are sufficient.

Responding directly to Bennett's (2011) concern that formative gains are frequently shallow and unsustained, the framework treats documentation not as a parallel activity appended to assessment, but as an intrinsic scored component evaluated against the same rubric as the substantive response. Pérez-Pérez et al.'s (2026) PRISMA 2020 scoping review validates this design choice: transparency requirements operationalised as mere declarations, without being embedded in assessable criteria or rubrics, fail to produce the epistemic clarity that valid inference requires, and generate incentives for strategic concealment rather than honest engagement. The review also confirms that proportional, expectation-aligned documentation requirements are operationally feasible, a finding corroborated in Nigerian classrooms by Adesokan et al. (2025), whose students offered detailed, coherent accounts distinguishing AI-generated responses from their own synthesis. Potential gaming of epistemic logs could take several forms: post-hoc rationalization of poorly-reasoned work disguised as authentic thinking-aloud, verbatim copying of AI-generated reasoning explanations, or overly polished documentation mismatched to final product quality. Teacher training should include practice distinguishing authentic from fabricated logs using exemplar student work, with standards calibrated through peer moderation.

Principle 4: Equitable Access

Assessment systems must ensure every student can demonstrate competence regardless of digital access, technological skill, or learning context (Universal Design for Learning, 2023). Under generative AI, equity carries an additional validity dimension: when assessment requires digital pathways without non-digital alternatives, it measures AI access rather than scientific competence. In Nigeria, where AI access is sharply stratified by geography and socioeconomic status, this problem is concrete. The 65% of rural schools lacking consistent electricity (Ibitoye et al., 2025), combined with average internet penetration of 35-40% (World Bank, 2023), means that assessment requiring technology-dependent pathways would systematically disadvantage the majority of Nigerian secondary students, not because they lack reasoning capacity, but because they lack infrastructure. Performance differences under such conditions measure access, not competence, invalidating assessment at its inferential root.

The framework's solution is twofold. First, multiple valid pathways exist: written investigative reports with annotated reasoning, recorded oral explanations, visual concept maps, and group projects with individual reflection, each evaluated against identical scientific reasoning criteria. No pathway requires technology ownership. Second, transparent documentation of reasoning makes actual intellectual engagement visible regardless of AI access: a student without ChatGPT who reasons incrementally and documents that process, and a student with full AI access who carefully annotates and explains modifications to AI-generated content, can both demonstrate equivalent competence through equivalent reasoning. Process journals can be handwritten, documented through classroom conversation, or captured through low-bandwidth methods. Note: Whether scientific reasoning constructs are truly equivalent across modalities (written, oral, visual) requires empirical investigation through measurement models to ensure rubric equivalence across expression formats.

Critically, restricting institutional resources does not prevent AI use. Yakubu et al. (2025) found that facilitating conditions, whether students had institutional technological support, had no significant effect on intention to use generative AI ($\beta = 0.003$, $p = .963$). Students require only an internet-capable personal device, making access effectively independent of institutional provision. Restricting school-level resources therefore widens the equity gap between students who access AI privately and those who cannot, without reducing

the validity threat. Only transparent documentation of reasoning makes this equity gap irrelevant to assessment validity.

Integrated Application: Worked Example

Consider a Senior Secondary 2 chemistry class in Delta State with 65 students and unreliable internet. Under Principle 1, the learning target becomes reasoning through unfamiliar redox reactions, not equation memorisation. Under Principle 2, the teacher presents a locally observable scenario: rusting of an iron water tank on school grounds. Students identify oxidising and reducing agents, explain electron transfer, propose an intervention, and justify their reasoning with reference to specific observable features. Under Principle 3, students submit documentation recording what evidence they considered, uncertainties they encountered, what prior knowledge they used, and where they consulted AI tools (for those with device access). Teachers use these epistemic records to distinguish students who constructed reasoning from those reproducing text without engagement. Under Principle 4, both device and non-device pathways use the same rubric evaluating chemical reasoning quality, justification specificity, and accuracy. Technological access does not determine the opportunity to demonstrate competence. This example requires no new infrastructure: only shifted expectations about what tasks demand, what epistemic documentation accompanies responses, and how such documentation reveals authentic understanding.

Limitations, Implementation Barriers, and Recommendations

Three critical barriers constrain implementation, with the largest being structural rather than pedagogical. First, national examination dominance creates a ceiling effect for school-level reform. While this framework proposes extended-response assessment with epistemic transparency, most Nigerian science assessment remains Multiple Choice, True/False, and Short Answer formats in JAMB, WAEC, and NECO examinations. These formats are neither designed for nor compatible with reasoning documentation. WAEC's Paper 3 in Chemistry and Physics does employ extended-response items, but these represent perhaps 30-40% of final grade weighting across states; the bulk of assessment remains selected-response. Teachers implementing this framework in schools face a structural disincentive: students coached in reasoning documentation for school-based assessment must nonetheless prepare for national examinations that do not reward such skills. This is not a pedagogical problem amenable to rubric design; it is an institutional misalignment requiring examination board restructuring. Piloting in four to six states would generate feasibility evidence, but national scale adoption requires prior commitment from WAEC, JAMB, and NECO leadership to gradually shift high-stakes examinations toward extended-response and documented reasoning formats: a multi-year structural change beyond any single institution's capacity.

Second, marking burden increases substantially when documentation accompanies extended-response tasks. Spot-check oral viva sampling of 10-15 students per assessment cycle at 4-5 minutes per student requires approximately 50-75 minutes per round. However, if oral vivas replace rather than supplement written examinations, net burden reduces to approximately 30-45 additional minutes per term, contingent on examination schedule restructuring within WAEC's authority. This remains manageable only if not layered atop existing grading loads; teachers in Nigerian schools averaging 50+ students per class report that current marking demands are already unsustainable.

Third, teacher assessment literacy remains limited. Research on assessment reform confirms that one-time in-service workshops produce minimal sustained change (Ibitoye et al., 2025); effective models employ monthly coaching cycles over one academic year, enabling teachers to practice rubric application using actual student work and calibrate standards through peer moderation. TRCN, NTI, and university-based teacher education programmes provide existing infrastructure for integration.

On empirical claims: Several 2025 citations require methodological transparency. Adesokan et al. (2025) and Yakubu et al. (2025) are peer-reviewed studies with adequate sample sizes (150 and 289 students respectively), though they lack disaggregated outcome reporting by achievement level. Nwali & Udumukwu (2025) surveyed 150 academics across three Nigerian universities and reported 80% expressed concern about data privacy in AI detection systems; this represents opinion on specific concerns rather than a claim about detection tool performance. Ya'u & Mohammed (2025) examined 245 university students, reporting that 75% used AI primarily for surface-level tasks (summarisation, editing), though this pattern may not generalise to secondary students with lower AI proficiency. These are presented as indicative patterns from Nigerian contexts rather than universal claims.

The broader recommendation requires sequential institutional action. If students are driven by performance expectancy rather than policy compliance, institutional change occurs only when assessment designs restructure which performances receive reward. Detection and prohibition have failed because they do not change the expectancy that AI use appears to fulfil. Redesigned assessment that makes authentic reasoning demonstrably more rewarding than substitution directly addresses this structural driver. However, school-level redesign without examination board alignment creates the misalignment barrier described above. National adoption therefore requires: (1) examination board commitment to phased extended-response expansion; (2) pilot in 4-6 states across 2-3 academic cycles generating feasibility and outcome data; (3) integration into WAEC, JAMB, and NECO protocols with staged rollout; (4) concurrent intensive teacher coaching infrastructure through TRCN and NTI. Each stage requires evidence generation from prior implementation to build political and institutional buy-in.

References

Abbas, M. Y. (2025). Distinguishing artificial intelligence-generated text from human text: Evaluating GPT-4o's self-detection capacity. Computers and Education, 45(2), 103-118.

Adesokan, A., Salman, O., & Raheem, O. (2025). Generative AI use and metacognitive awareness among Nigerian secondary students: A mixed-methods study. Journal of Educational Technology and Society, 28(1), 45-62.

Arifin, S., Chen, L., Thompson, M., & Williams, R. (2025). Inquiry-based learning and critical thinking development: A meta-analytic review. Review of Educational Research, 95(1), 1-38.

Ateeq, M., Al-Balushi, S., Al-Harthi, M., & Hassan, M. (2024). Educational impact on student achievement in higher education: A PLS-SEM approach. Education and Information Technologies, 29(3), 2847-2870.

Banta, T. W., & Palomba, C. A. (2014). Assessment essentials: Planning, implementing, and improving student learning (2nd ed.). Jossey-Bass.

Bennett, R. E. (2011). Formative assessment: A critical review. Assessment in Education: Principles, Policy & Practice, 18(1), 5-25.

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). Chatting with ChatGPT: How may I assist you? Journal of Academic Language & Learning, 17(1), 180-193.

Darling-Hammond, L., Ancess, J., & Falk, B. (2013). Authentic assessment in action: Studies of schools and students at work. Teachers College Press.

Erol, Y., Coban, C., & Yasar, S. (2025). Reliability and bias in AI-generated text detection systems. Computational Linguistics Review, 51(2), 234-256.

European Network for Academic Integrity. (2024). Guidance on institutional approaches to academic integrity in the age of artificial intelligence. ENAI Publications.

Eze, P. O., & Obi, T. U. (2019). Assessment in Nigerian secondary science education: Current practices and policy implications. African Journal of Educational Assessment, 7(2), 112-130.

Federal Ministry of Education. (2015). National curriculum for secondary education in Nigeria. FME Publications.

Garzon, J., Baldiris, S., & Fabregat, R. (2025). Generative AI and virtual reality: Emerging applications in educational assessment. Computers & Education, 216, 104897.

Gill, B. P., Zhang, M., & Chen, W. (2023). ChatGPT and academic integrity: An empirical investigation of detection evasion. International Review of Research in Open and Distributed Learning, 24(4), 289-310.

Hargreaves, A. (2005). Educational change takes ages: Life, career and generational factors in teachers' emotional responses to educational change. Teaching and Teacher Education, 21(8), 967-983.

Herman, J. L., & Lara-Steidel, A. P. (2025). Authentic assessment for deeper learning in science education. Science Education Review, 34(1), 45-68.

Holmes, K., Bailik, S., & Fadel, C. (2019). Artificial intelligence in education: Promises and implications. Center for Curriculum Redesign.

Holmes, W., Persson, J., & Chounta, I. A. (2022). Artificial intelligence and tutoring systems: Advances, opportunities, and challenges. Journal of Educational Computing Research, 60(5), 1104-1126.

Holmes, W., Tuomi, I., & Kamarainen, A. (2022). State-of-the-art and practice in AI-enhanced learning environments. European Commission, JRC Publications Repository. https://doi.org/10.2760/137379

Ibitoye, M., Okafor, C., Nnamdi, O., & Adebayo, J. (2025). Information communication technology integration in Nigerian secondary schools: A systematic review of capacity and constraints. Computers and Education, 217, 104923.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 17-64). Praeger.

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

Kind, P. M., & Osborne, J. F. (2017). Styles of science reasoning: A student typology informed by epistemology. International Journal of Science Education, 39(6), 731-748.

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. Sage Publications.

Luckin, R. (2023). Machine learning and human intelligence: The future of education in the age of AI. UCL Institute of Education Press.

Macgilchrist, F. (2021). Postdigital heterogeneity: Digital living in the midst of smart devices, algorithms and data. Postdigital Science and Education, 3(2), 441-461. https://doi.org/10.1007/s42438-020-00166-9

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). Macmillan.

National Research Council. (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. National Academies Press.

Nwali, C. U., & Udumukwu, O. E. (2025). Academic staff perspectives on AI authorship and institutional policy gaps in Nigerian universities. Higher Education Research & Development, 44(2), 234-252.

Okafor, C. E., Nwoke, B. I., & Eze, S. O. (2018). Inquiry-based science teaching and student achievement in Nigerian secondary schools. African Educational Research Journal, 6(3), 127-142.

Ojimba, D. P. (2023). National examination systems and science curriculum implementation in Nigeria: Alignment and challenges. Journal of Curriculum Studies in Nigeria, 15(1), 89-107.

Osborne, J. F., & Dillon, J. (2008). Science education in Europe: Critical reflections. King's College London, Educational Resource Information Center.

Pearce, J. M., & Chiavaroli, N. (2020). Practical considerations for scaling oral assessment. Assessment & Evaluation in Higher Education, 45(2), 156-174.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for understanding and improving science assessment. Measurement: Interdisciplinary Research and Perspectives, 14(2), 35-56.

Pérez-Pérez, M., López-García, C., & Rodríguez-Martínez, E. (2026). Transparency requirements in AI governance: A PRISMA 2020 scoping review of disclosure mandates and compliance outcomes. Science, Technology & Human Values, 51(1), 112-142.

Peterson, J. (2025). The detection signature problem: How probabilistic generation defeats static AI detection systems. Artificial Intelligence and Society, 40(1), 189-207.

Salaudeen, Y., Pitan, O. S., & Aminu, M. A. (2025). Measurement and validity in assessment: Conceptual distinctions and implications for AI-assisted evaluation systems. International Journal of Assessment and Evaluation, 33(2), 78-96.

Scarfe, E., Watchirn, E., & Gibbs, G. (2024). Can we detect where text is written by language models? A comparative analysis of detection tools' performance. Journal of Academic Integrity, 7(2), 45-62.

Schon, D. A. (1983). The reflective practitioner: How professionals think in action. Basic Books.

Selwyn, N. (2019). Automation, education, and the surveillance society. Oxford University Press.

Selwyn, N. (2022). The digital education myth: Trajectories of technology in education. Polity Press.

Shi, Z., Chen, H., Wang, Y., & Liu, Q. (2025). Large language models and authorship attribution: Understanding conflation of plausibility and authenticity. Natural Language Processing Review, 28(3), 234-260.

Stribling, C. M., Clifton, J. W., & Maclean, R. (2024). Large language models in graduate-level science assessment: Performance, competency mapping, and implications for credentialing. Computers and Education, 210, 104816.

Supovitz, J. A. (2009). Can high stakes testing leverage educational improvement? Prospects from the last decade of testing and accountability reform. Journal of Educational Change, 10(2-3), 211-227.

Touretzky, D. S., Martin, C., & Seehorn, D. (2019). Envisioning AI for K-12: Reflections on the national artificial intelligence research resource task group recommendations. ArXiv Preprint. https://arxiv.org/abs/2010.08892

UNESCO. (2023). Global status report on artificial intelligence in education. UNESCO Publications.

Universal Design for Learning. (2023). UDL guidelines 2.2: Supporting all learners in all contexts. CAST.

University College Cork. (2025). Institutional guidance on academic integrity policies in an era of generative AI. UCC Publications.

Usman, I., Gada, A., & Musa, B. (2025). Tiered AI involvement frameworks and student learning outcomes: Evidence from Nigerian higher education institutions. Technology, Pedagogy and Education, 34(1), 56-74.

Wang, J., Xia, M., & Ye, Z. (2025). The reliability problem in AI-generated text detection: Performance variation across task types. ACM Computing Surveys, 58(3), 1-42.

West, S., Shebab, E., & Williamson, M. (2023). ChatGPT and laboratory reports: Evaluating authenticity and coherence in student submissions. Assessment & Evaluation in Higher Education, 48(7), 1058-1076.

Wiggins, G. P. (1998). Educative assessment: Designing assessments to improve student performance. Jossey-Bass.

World Bank. (2023). Digital development overview: Sub-Saharan Africa connectivity and infrastructure. World Bank Publications.

Ya'u, Y. Z., & Mohammed, H. I. (2025). Generative AI use patterns and academic performance among Nigerian university students. Computers and Education, 214, 104879.

Yakubu, A., David, O., & Abubakar, M. (2025). Generative AI use intentions among Nigerian university students: A structural equation modelling approach. Computers and Education, 215, 104902.