ETDC
Educational Technology in Developing Countries

Vol. 1, No.1 (June 2026), 83-102
P-ISSN: xxxx-xxxx | E-ISSN: xxxx-xxxx
DOI: https://doi.org/10.xxxxx/lms.vxxx.xx
https://ejournal.tuah.or.id/index.php/etdc/index

Design and Validation of a Quality Assurance Framework for AI Educational Technology in Low-Resource Settings

Mohamad Haditia[1]*; Ismail Yau Abubakar[2], Rokimin[3], Anas Fauzi[3]

[1] STIT Al-Anshar Tanjung Selor, North Kalimantan, Indonesia
[2] British International School, Jahi, Nigeria
[3] Darunnajah University, Jakarta, Indonesia

*Correspondence: mohamad.haditia.2301218@students.um.ac.id

Received: 25-02-2026	Accepted: 14-05-2026	Accepted: 17-05-2026

Abstract. The diffusion of artificial intelligence-enabled educational technology in low- and middle-income countries has outpaced the development of context-appropriate quality assurance mechanisms, leaving procurement agencies without evidence-based instruments to evaluate the pedagogical, technical, governance, and equity dimensions of such products. This study develops, validates, and pilot-tests the Quality Assurance Framework for Artificial Intelligence Enabled Educational Technology (hereafter, AEQAF) through Design Science Research grounded in responsible artificial intelligence theory, critical educational technology scholarship, and decolonial perspectives on technology in the Global South. The framework was constructed across three iterative cycles over twenty-four months through four phases: landscape analysis of 147 products across 23 markets; co-design workshops with 68 stakeholders in five venues; a three-round modified Delphi with 24 experts across 14 countries; and pilot application on 18 products in Nigeria, Ghana, and Kenya using three trained evaluators. The AEQAF comprises six dimensions operationalised through 42 indicators. Delphi consensus at the 75% threshold was achieved for 39 of 42 indicators. The pilot application demonstrated substantial interrater reliability (Fleiss kappa = 0.81). A non-circular discriminant validity test using five dimensions yielded a Cohen's d of 1.42, supporting the framework's capacity to differentiate product quality. The framework reveals systematically low scores for data governance, equity, and evidence dimensions. The AEQAF is ready for multi-country validation and provides an openly licensed instrument bridging technical and educational evaluation for resource-constrained systems.

Keywords: artificial intelligence in education, quality assurance, low- and middle-income countries, design science research, decolonial perspective.

INTRODUCTION

The accelerating integration of artificial intelligence into educational technology has reshaped the global education technology landscape at a pace that has overwhelmed most national regulatory systems' capacity to respond (Bond et al., 2024; Holmes et al., 2022; Williamson et al., 2023). The global market for artificial intelligence in education was valued at USD 5.88 billion in 2024 and is projected to grow at a compound annual rate exceeding 30% through the end of the decade (Khan et al., 2024; Zapata-Rivera et al., 2024). In low- and middle-income countries, where learning poverty affects approximately seven in ten children,

artificial intelligence-enabled products are increasingly positioned as scalable solutions to persistent shortfalls in educational quality, teacher capacity, and equitable access (Nemorin et al., 2023; Okiri, 2024). The speed of product proliferation, combined with the relative absence of contextualised evaluation infrastructure, has produced a pronounced gap between supply and governance capacity. This gap is the empirical phenomenon that motivates the present study.

A recent live mapping by AI for Education identified 84 artificial intelligence-powered solutions deployed across Sub-Saharan Africa, India, and Pakistan, of which 43% were adaptive, personalised learning platforms and 24% were student-facing chatbots (Nemorin et al., 2023). Despite this volume, the proportion of products that have undergone rigorous pedagogical evaluation remains extremely small. Systematic reviews have repeatedly found that fewer than one in ten deployed educational technology products in low- and middle-income countries have independent evidence of learning impact, with evidence concentrated in a small subset of relatively well-resourced implementations (Hennessy et al., 2022; Major et al., 2021; Rodriguez-Segura, 2022). Adoption decisions are frequently made without reliable information about pedagogical alignment, technical performance under bandwidth constraints, governance posture, or equity implications of specific products (Luckin & Cukurova, 2019; Perrotta & Selwyn, 2020). The result is a procurement environment in which quality is asserted rather than evidenced, and in which the burden of evaluation falls on actors least equipped to perform it.

This procurement environment sits in tension with three increasingly mature scholarly conversations that the present framework must enter rather than merely cite. The first is the responsible artificial intelligence literature, which has developed robust general principles for algorithmic fairness, transparency, accountability, and privacy (Floridi et al., 2018; Jobin et al., 2019). The second is the critical educational technology scholarship that interrogates how datafication, automation, and corporate platform power reshape educational governance (Macgilchrist et al., 2024; Selwyn, 2024; Williamson et al., 2023). The third is the decolonial artificial intelligence scholarship, which examines how technologies developed in high-income contexts impose epistemic, economic, and political asymmetries on the Global South, and which calls for design processes that centre marginalised voices (Birhane, 2020; Costanza-Chock, 2020; Mohamed et al., 2020). These three literatures share a common observation that the present study extends: responsible artificial intelligence in education cannot be reduced to technical compliance; it requires the construction of an evaluation infrastructure that is itself context appropriate, participatory in its design, and accountable to the communities it serves.

Three connected gaps follow from this observation, and the AEQAF is designed to address all three. The first gap concerns the absence of an evaluation instrument that simultaneously attends to technical artificial intelligence performance and pedagogical educational quality in a single, coherent instrument calibrated to the realities of low- and middle-income countries. Model-level benchmarks such as the Massive Multitask Language Understanding benchmark and the Holistic Evaluation of Language Models suite (Hendrycks et al., 2020; Liang et al., 2023) assess isolated capabilities in abstraction from pedagogical context, cultural relevance, or classroom deployment conditions, and are largely developed against materials from high-income countries. Product level educational technology evaluation frameworks such as EdTech Tulna in India, EdTech Impact in the United Kingdom, and the Systems Approach for Better Education Results in Information and Communication Technology programme of the World Bank were developed before the widespread adoption

of generative artificial intelligence and do not systematically address artificial intelligence specific concerns including hallucination, algorithmic bias, model provenance, and the dynamic nondeterministic nature of generative outputs. The structured comparison in the Discussion grounds this argument empirically rather than solely conceptually.

The second gap concerns the data governance and equity dimensions of artificial intelligence in education in low- and middle-income country contexts. Recent landscape reviews indicate that only a small minority of countries have laws specifically protecting children's data privacy in educational settings, and that, where such laws exist, they typically regulate ministries rather than private educational technology providers (Holmes et al., 2022; Williamson et al., 2023). This governance asymmetry is sharpened in low and middle income countries where children data is frequently transferred to foreign jurisdictions for model training and analytics purposes without transparent consent mechanisms or equivalent legal protections, a practice that Birhane (2020) and Mohamed et al. (2020) explicitly frame as algorithmic colonisation when the value extracted from such data accrues to foreign technology corporations rather than the communities from which it originates. At the same time, emerging evidence indicates that artificial intelligence-enabled products are not equally accessible across gender, socioeconomic, and urban-rural dimensions (Molina et al., 2018; Muralidharan et al., 2019). A practical evaluation framework for low- and middle-income country contexts must therefore treat data governance and equity as first-class evaluation dimensions rather than peripheral concerns.

The third gap concerns the implementability of existing frameworks within the capacity profiles of ministries of education in low- and middle-income countries. Frameworks developed for high-income regulatory contexts frequently assume resource levels, technical staff, and procurement governance structures that are not typical in low-resource settings. A practically usable framework must combine conceptual rigour with operational feasibility: indicators must be scoreable by trained non-specialist evaluators, evidence requirements must be tiered rather than binary, and documentation must support transparent comparison across heterogeneous products. The present study addresses these three connected gaps through a Design Science Research approach (Hevner et al., 2004; Peffers et al., 2007), in which the research contribution takes the form of a validated, openly licensed artefact along with empirical evidence of its reliability, validity, and diagnostic utility. The study investigates three research questions. First, what dimensions and indicators should constitute an evidence-based quality assurance framework for evaluating artificial intelligence-enabled products in low-resource educational settings? Second, to what extent can a multinational multi-stakeholder consensus be achieved on the proposed framework's dimensions and indicators through iterative expert validation? Third, does the resulting framework demonstrate acceptable interrater reliability and discriminant validity when applied to evaluate currently deployed products? The study contributes to the Educational Technology in Developing Countries research agenda across four registers, as detailed in the Discussion: theoretical, methodological, empirical, and practical.

CONCEPTUAL FRAMEWORK

The AEQAF integrates three theoretical strands that share a common normative commitment to evaluating educational technology by reference to the communities it serves rather than solely to the systems it instantiates. The first strand operationalises responsible artificial intelligence principles (Floridi et al., 2018; Jobin et al., 2019) into the education

domain, translating abstract commitments to fairness, accountability, transparency, and privacy into scoreable indicators grounded in specific design features. The second strand draws on the Technological Pedagogical Content Knowledge tradition (Koehler et al., 2014) to position pedagogical alignment as inseparable from technical performance and contextual relevance, rejecting the dichotomy in which artificial intelligence products are evaluated either as technical systems or as pedagogical tools. The third strand builds on the capability and equity orientation in development scholarship and on its critical and decolonial extensions (Birhane, 2020; Costanza-Chock, 2020; Mohamed et al., 2020; Ramsarup et al., 2023), positioning equity as constitutive of rather than supplementary to product quality. These strands are not selected arbitrarily; they are the literatures that, together, account for what an evaluation framework operating in this domain must register: normative principles, pedagogical conditions, and the distributive politics of who benefits.

The integration of these three strands is generative rather than additive, and each strand does work that the others cannot. Responsible artificial intelligence principles supply the normative anchors against which features such as algorithmic transparency and data minimisation become scoreable. However, they say little about the pedagogical conditions under which such features matter. The Technological Pedagogical Content Knowledge tradition provides the pedagogical anchor, foregrounding the integration of content, pedagogy, and technology, but was originally developed to evaluate teacher knowledge rather than product design, and must be redefined at the product level. The present study performs this re-specification by treating the product as a sociotechnical artefact whose pedagogical model is embedded in its design choices. The capability and decolonial strand then provides the substantive view of what education is for: not the transfer of content but the expansion of human agency under conditions of substantive equality, including across the urban-rural, gender, and language dimensions along which low- and middle-income country educational systems often divide.

The resulting conceptual stance is that the quality of an artificial intelligence-enabled educational technology product for deployment in low- and middle-income countries is a multidimensional construct instantiated by six dimensions: Pedagogical Alignment, Technical Performance, Contextual Relevance, Data Governance, Equity and Inclusion, and Evidence of Impact. Each dimension is operationalised through indicators with explicit scoring rubrics, and the aggregate framework score is a profile rather than a single scalar, which supports diagnostic use. The choice to report a profile is itself a substantive decision aligned with the design justice tradition (Costanza-Chock, 2020). A single number obscures the trade-offs among dimensions that procurement officers, ministries, and developers need to see, whereas a profile makes them visible and contestable. The framework, therefore, functions as a diagnostic instrument that supports deliberation rather than as a ranking mechanism that displaces it.

METHOD

Research Design

The study adopted the six-phase Design Science Research Methodology (Peffers et al., 2007): problem identification, objective definition, design and development, demonstration, evaluation, and communication. Three iterative cycles were conducted between March 2024 and February 2026. Design Science Research was selected because the research objective required the construction of a purposeful artefact with both conceptual rigour and practical

utility (Hevner et al., 2004; vom Brocke et al., 2020). Two alternatives were considered and rejected: pure instrument development through systematic review and Delphi alone would have produced a validated instrument without the iterative co-design that the low- and middle-income country context requires; participatory action research would have foregrounded co-design but would not have produced the structured expert consensus needed for procurement use. Design Science Research combines both moves within a single iterative cycle. Evaluation was multimodal, following the Framework for Evaluation in Design Science (Venable et al., 2016), with both ex ante and ex post components. Figure 1 visualises the full process.

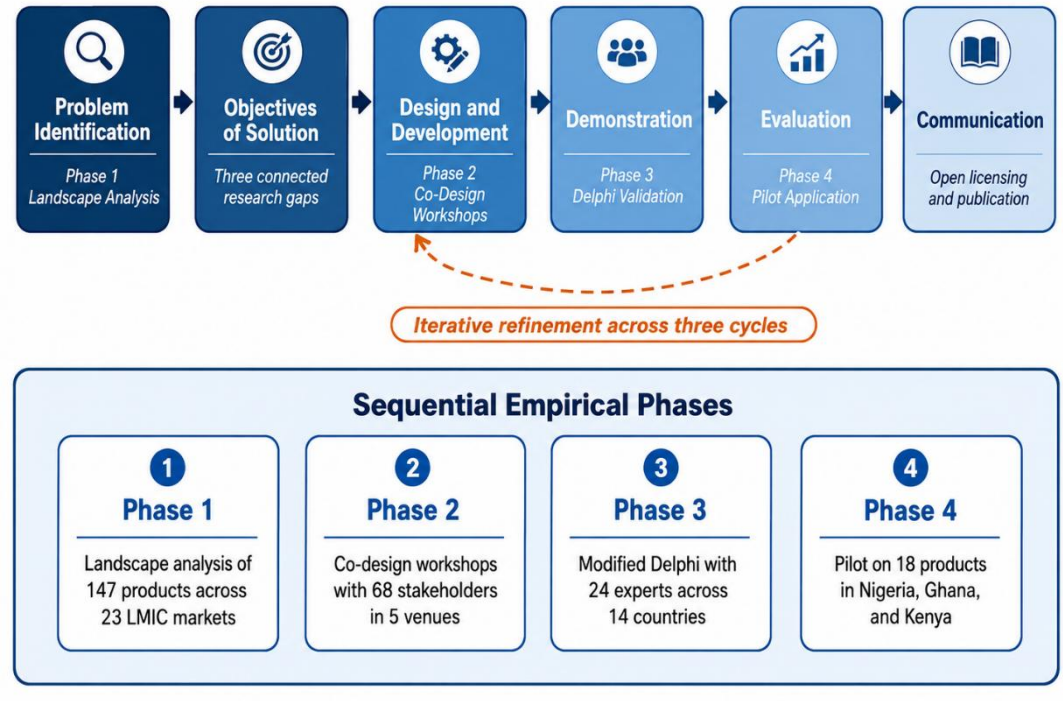

Figure 1. Design Science Research Process and Four Sequential Empirical Phases

Phases of Empirical Inquiry

The four sequential empirical phases differ in their objectives, participants, data sources, and analytic methods, but each contributes to the iterative refinement of the framework architecture (Eybers, 2023; Fahd et al, 2021; Steinherr et al., 2024). Table 1 presents a structured summary of inputs, procedures, and outputs across all phases. The remainder of this section provides only the procedural detail necessary to interpret the results that follow, with extensive documentation of search protocols, workshop synthesis, Delphi recruitment, and evaluator training preserved in the audit trail.

Table 1. Summary of Empirical Phases

Phase	Objective	Participants and Data	Procedure	Output
1	Map the ecosystem and identify gaps	147 products across 23 markets; 89 documents from databases and grey literature	Boolean search 2019 to 2025 in Scopus, Web of Science, and ERIC; product coding along six dimensions	Requirements specification for framework dimensions
2	Co-design framework architecture	68 stakeholders across four categories in five venues (Lagos, Accra, Nairobi, New Delhi, online)	Two-day workshops with card sorting, nominal group technique, and collaborative rubric calibration	Candidate dimensions and indicators with stakeholder grounding
3	Validate framework through expert consensus	24 experts across 14 countries with expertise in at least two of the four required domains	Three-round modified Delphi with relevance and clarity ratings; 75% consensus threshold defined a priori	Validated framework with 42 indicators and structured rubrics
4	Test reliability and validity in deployment	18 products across three categories in Nigeria, Ghana, and Kenya; three trained evaluators	Stratified purposive sampling; 20-hour evaluator training; independent scoring of each product on all 42 indicators	Interrater reliability estimates and discriminant validity evidence

Landscape Analysis and Requirements

Phase 1 mapped the artificial intelligence-enabled educational technology ecosystem in low- and middle-income countries and surveyed existing evaluation frameworks. Academic databases (Scopus, Web of Science, ERIC) and grey literature sources (World Bank, UNESCO, UNICEF Innocenti, EdTech Hub, AI for Education) were searched using Boolean combinations of terms from January 2019 through February 2025. The initial search yielded 487 records, reduced to 312 after duplicate removal. Application of inclusion criteria (peer-reviewed or formally published grey literature, English language, explicit focus on artificial intelligence in education or on educational technology evaluation in low- and middle-income country contexts) yielded 89 documents for thematic analysis. Two reviewers screened independently and resolved disagreements with a third reviewer (Cohen's kappa = 0.84) (Fleiss et al., 2003). In parallel, 147 products deployed across 23 markets were catalogued and coded along six dimensions: functional category, target population, artificial intelligence architecture type, evidence base, data practices, and target market. Product data were sourced from the AI for Education product dashboard, the EdTech Hub evidence library, and direct developer submissions (Nemorin et al., 2023).

Multinational Co-Design Workshops

Phase 2 conducted five co-design workshops between June and November 2024 in Lagos (Nigeria), Accra (Ghana), Nairobi (Kenya), New Delhi (India), and an online venue. A total of 68 participants were purposively recruited through partner nominations and open calls, spanning four stakeholder categories: government policymakers and ministry officials (n = 18), developers and engineers (n = 16), primary and secondary school educators and administrators (n = 19), and education researchers and evaluators (n = 15). Each workshop

lasted two days and followed a participatory design protocol comprising card sorting of candidate dimensions, indicator development through nominal group technique, and rubric calibration through collaborative scoring (Costanza-Chock, 2020; Holstein et al., 2019). Participants exerted significant influence over the framework's structure: in two workshops, they successfully argued for adding or merging dimensions that the research team had not anticipated. Artefacts were synthesised through reflexive thematic analysis (Braun & Clarke, 2019) by two researchers, with intercoder agreement examined on a 30% subset (Cohen's kappa = 0.79). Tensions between workshops were documented explicitly and resolved through cross-site reconciliation discussions.

Modified Delphi Validation

Phase 3 implemented a three-round modified Delphi procedure to achieve structured expert consensus (Diamond et al., 2014). Twenty-four international experts were purposively recruited through a multistage process combining bibliometric identification of highly cited authors (Scopus citation threshold above the eightieth percentile within the 2019 to 2024 window) and nominations from Phase 2 participants and international partner organisations. Because bibliometric identification tends to surface scholars publishing in English-language indexed journals (skewing toward high-income perspectives), supplementary nominations from low- and middle-income country partner institutions and non-governmental organisations were used to ensure that scholars working outside indexed channels were represented. Eligible panelists demonstrated expertise in at least two of the four domains: artificial intelligence and machine learning; educational technology evaluation; low- and middle-income country education systems; and data governance. The final panel represented 14 countries across Sub-Saharan Africa (n = 9), South Asia (n = 5), Southeast Asia (n = 3), Europe (n = 4), and North America (n = 3). Forty-two experts were invited; eighteen declined (twelve cited time constraints, six did not respond), with no systematic pattern in non-response across regions. Consensus was defined a priori as 75% of panelists rating an indicator at four or above on both relevance and clarity. Panel retention across rounds was 100%. Sensitivity analyses at 70% and 80% thresholds are reported alongside the primary results.

Pilot Application and Reliability Testing

Phase 4 piloted the validated framework on 18 products currently deployed in Nigeria, Ghana, and Kenya (six per country), selected through stratified purposive sampling across three functional categories: adaptive learning platforms (n = 7), tutoring chatbots (n = 6), and assisted assessment tools (n = 5). The sampling frame was constructed from the Phase 1 product inventory, restricted to products with active deployment in at least one of the three countries during 2024 to 2025, yielding 53 eligible products, of which 18 were sampled to balance functional category, developer origin (local, regional, global), and evidence-based maturity. Three evaluators were trained over a 20-hour programme that included framework orientation, rubric calibration using sample products not included in the pilot, and iterative resolution of disagreements. The three evaluators represented complementary backgrounds (educational psychology, computer science, and education policy) chosen to approximate the multidisciplinary teams that procurement agencies would typically deploy. Each evaluator independently scored each product; evaluations relied exclusively on publicly available information and developer disclosures. Interrater reliability was quantified using Fleiss kappa (Fleiss et al., 2003). Discriminant validity was evaluated using a five-dimensional residual score (dropping Evidence of Impact because it conceptually overlaps with the external criterion) and a Welch independent-samples t test with Cohen's d effect size; this design choice was

made a priori to prevent the partial circularity that would otherwise inflate the validity estimate (Gregor & Hevner, 2013).

Trustworthiness and Reflexivity

Trustworthiness procedures followed those of Lincoln and Guba (1985), as updated for mixed research by Nowell et al. (2017). Credibility was supported through methodological triangulation across the four phases and through member checking, in which preliminary findings were returned to a stratified subset of 14 participants for written comment. Transferability is supported by a detailed contextual description of the three pilot countries and by the analytic distinction (developed in the Discussion) between the framework's portable core and its context-sensitive periphery. Dependability is supported by a detailed audit trail accessible to qualified researchers on request. Each team member's reflexivity journaling supports confirmability. The author team includes researchers based in the Global South and the Global North, as well as from multilateral institutions, with backgrounds in computer science, education policy, learning engineering, and artificial intelligence governance. The team declares no current consulting relationships with the developers of the pilot products or with the agencies most likely to adopt the framework.

It is important to note that methods must be written in the same order in the results section. The order of writing methods must also be logical, depending on the type of research. The method for one type of research will differ significantly from that of other studies. For example, writing survey research methods is very different from writing laboratory test research methods, which involve a lot of equipment and materials. The method section can be organized into several separate subsections, such as materials, tools, and data collection procedures.

Very likely, a novelty in a study is in the method section, even when the topic is the same as in previous studies. New methods that are simpler yet equally capable of answering research questions are superior, as they can be replicated or applied by subsequent researchers. In addition, if the equipment has an accuracy tolerance for reading data, such as a thermocouple, transducer, or airflow meter, this must be clearly and honestly stated in the method section.

RESULT

Landscape Analysis Findings

The systematic analysis of 147 products across 23 markets revealed substantial gaps in quality and evidence. Only 23 products (15.6%) reported any form of learning outcomes evaluation; of these, only 8 (5.4%) had undergone independent, externally validated impact assessment. 91 products (61.9%) lacked publicly accessible data governance policies, and 104 products (70.7%) lacked documented alignment with national curricula in their target markets. Regarding artificial intelligence-specific features, 67 products (45.6%) utilised large language models without disclosing information about model provenance, training data composition, or hallucination mitigation strategies. These patterns were broadly consistent across the three regional subsamples, with Sub-Saharan African products showing slightly lower data governance disclosure rates (34.8% versus 41.2% in South Asia and 42.9% in Southeast Asia). The disclosure gaps are more pronounced for the dimensions specific to artificial intelligence than for general educational technology indicators, suggesting an artificial intelligence-specific governance deficit beyond the sector-wide pattern. These

findings confirmed the acute need for a structured evaluation framework and informed the prioritisation of data governance and evidence of impact as framework dimensions.

Framework Architecture and Delphi Consensus

Through three iterative cycles of co-design and expert validation, the AEQAF converged on six evaluation dimensions, operationalised through 42 indicators. Table 2 presents the framework architecture with dimension definitions and indicator counts. Table 3 presents the Delphi consensus outcomes across three rounds. Overall, 39 of 42 indicators (92.9%) achieved the a priori consensus threshold. Contextual Relevance achieved a unanimous consensus (100%). Three indicators in the Evidence of Impact dimension required substantive revision between Rounds 2 and 3. The original indicator requiring randomised controlled trial-level evidence was reframed as a tiered evidence classification (descriptive, correlational, quasi-experimental, experimental) after panelists from low- and middle-income country contexts argued that mandating randomised controlled trial evidence would exclude promising early-stage products. This revision was subsequently endorsed at the 83.3% threshold.

Table 2. AEQAF Framework Architecture: Dimensions, Indicator Counts, and Definitions

No.	Dimension	k	Definition and Key Indicators
D1	Pedagogical Alignment	9	Alignment between product pedagogy and established principles of effective instruction and national curricular standards. Indicators include learning objective clarity, age appropriateness, scaffolding, quality of formative feedback, and transparency of the pedagogical model.
D2	Technical Performance	7	Reliability, accuracy, and responsiveness under realistic deployment conditions. Indicators include output accuracy, latency, offline functionality, device compatibility, multilingual support, and system uptime.
D3	Contextual Relevance	6	Fit between product design and the realities of deployment in low- and middle-income countries. Indicators include cultural appropriateness, local language support, representation diversity, low bandwidth operability, infrastructure alignment, and teacher integration design.
D4	Data Governance	8	Mechanisms protecting the privacy, dignity, and agency of learners. Indicators include privacy policy transparency, consent mechanisms, data minimisation, storage jurisdiction, disclosure of third-party sharing, algorithmic transparency, child-specific protections, and data deletion provisions.
D5	Equity and Inclusion	6	Capacity to serve diverse learners equitably. Indicators include gender bias assessment, disability accessibility in line with the Web Content Accessibility Guidelines, socioeconomic access barriers, urban-rural equity, algorithmic fairness auditing, and marginalised population representation.
D6	Evidence of Impact	6	Strength and independence of evidence supporting learning and implementation outcomes. Indicators include tiered evidence classification, implementation fidelity, user satisfaction data, cost-effectiveness reporting, scalability evidence, and availability of independent evaluation.
	Total	**42**	

Table 3. Modified Delphi Consensus Results Across Three Rounds (N = 24 Panelists)

Dimension	R1 M (SD)	R2 M (SD)	R3 Con. 75%	Con. 70%	Con. 80%	Status	Failed Items
D1 Pedagogical	4.6 (0.5)	4.7 (0.4)	95.8%	100%	87.5%	Accepted	0
D2 Technical	4.4 (0.6)	4.5 (0.5)	91.7%	95.8%	83.3%	Accepted	0
D3 Contextual	4.7 (0.4)	4.8 (0.3)	100%	100%	100%	Accepted	0
D4 Data Governance	4.5 (0.7)	4.6 (0.5)	87.5%	95.8%	79.2%	Accepted	1
D5 Equity	4.3 (0.6)	4.5 (0.5)	91.7%	100%	83.3%	Accepted	0
D6 Evidence	4.2 (0.8)	4.4 (0.6)	83.3%	91.7%	75.0%	Accepted	2
Overall	4.45 (0.6)	4.58 (0.5)	92.9%	97.6%	85.7%	Accepted	3

Pilot Application: Reliability and Discriminant Validity

Interrater reliability across three evaluators on the 18 product sample was substantial (Fleiss kappa = 0.81, 95% CI 0.74 to 0.88), exceeding the conventional 0.75 threshold (Fleiss et al., 2003). Dimension level kappa ranged from 0.73 (95% CI 0.64 to 0.82) for Evidence of Impact to 0.89 (95% CI 0.82 to 0.96) for Technical Performance. The gradient is interpretable: dimensions requiring technical verification yielded higher agreement than dimensions involving evaluative judgements. These estimates reflect a best-case scenario in which evaluators were trained directly by the framework developers; the reliability achieved by independently trained ministry teams is a research question for the next pilot wave.

Discriminant validity was tested using a non-circular design that excluded Evidence of Impact from the validating score, because that dimension overlaps conceptually with the external criterion. Products with independently documented evidence (n = 7) scored significantly higher on the five-dimensional residual score (M = 3.4, SD = 0.43) than products without (n = 11; M = 2.5, SD = 0.58), Welch t(15.7) = 3.51, p = 0.003, Cohen d = 1.42, 95% CI for d from 0.41 to 2.41. For transparency, the original six-dimensional score yielded Cohen's d = 1.94 (95% CI 0.82 to 3.06), now interpreted as internal consistency between the framework as a whole and one of its constituent dimensions rather than as a valid discriminant test. The corrected estimate of 1.42 remains a large effect.

The pilot revealed a consistent pattern: the lowest aggregate scores were observed for Data Governance (M = 2.5, SD = 0.71), Equity and Inclusion (M = 2.4, SD = 0.66), and Evidence of Impact (M = 2.0, SD = 0.84). These are precisely the dimensions most critical for responsible deployment. Figure 2 visualises the dimension-by-category pattern; Table 4 reports means, standard deviations, and ranges. Within-category variability should temper category-level claims: the chatbot category (n = 6) showed a wide range in Pedagogical Alignment (1.8 to 3.8), with two of six reaching the adequate threshold and four falling below it. Category-level differences from one-way analysis of variance with Tukey honestly significant difference comparisons (chatbots scoring lower than adaptive platforms on Pedagogical Alignment and Contextual Relevance, both $p < 0.01$) should be interpreted as central tendencies masking within-category variation. Two design choices distinguish higher-scoring chatbots: explicit pedagogical scaffolding within the conversational flow and locally produced training data.

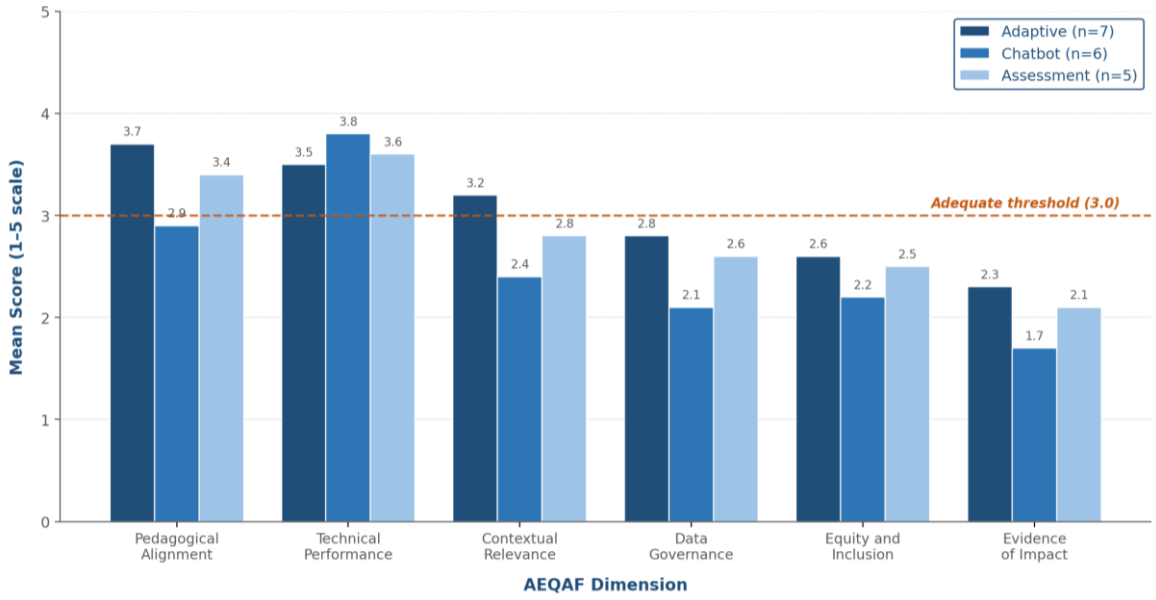

Figure 2. Mean AEQAF Scores by Product Category and Dimension

Table 4. Mean AEQAF Scores with Standard Deviations and Ranges by Product Category

Dimension	Adaptive (n = 7) M (SD) [range]	Chatbot (n = 6) M (SD) [range]	Assessment (n = 5) M (SD) [range]	Overall (N = 18) M (SD) [range]
D1 Pedagogical	3.7 (0.38) [3.1 to 4.2]	2.9 (0.71) [1.8 to 3.8]	3.4 (0.42) [2.9 to 3.9]	3.3 (0.61) [1.8 to 4.2]
D2 Technical	3.5 (0.40) [3.0 to 4.0]	3.8 (0.45) [3.2 to 4.4]	3.6 (0.37) [3.1 to 4.1]	3.6 (0.42) [3.0 to 4.4]
D3 Contextual	3.2 (0.47) [2.5 to 3.9]	2.4 (0.55) [1.5 to 3.1]	2.8 (0.40) [2.3 to 3.4]	2.8 (0.61) [1.5 to 3.9]
D4 Data Governance	2.8 (0.55) [2.0 to 3.6]	2.1 (0.62) [1.3 to 3.0]	2.6 (0.49) [2.1 to 3.3]	2.5 (0.71) [1.3 to 3.6]
D5 Equity	2.6 (0.48) [1.9 to 3.4]	2.2 (0.71) [1.4 to 3.2]	2.5 (0.41) [2.0 to 3.0]	2.4 (0.66) [1.4 to 3.4]
D6 Evidence	2.3 (0.78) [1.2 to 3.6]	1.7 (0.65) [1.0 to 2.7]	2.1 (0.83) [1.1 to 3.2]	2.0 (0.84) [1.0 to 3.6]
Overall AEQAF	3.0 (0.41) [2.4 to 3.7]	2.5 (0.58) [1.6 to 3.4]	2.8 (0.42) [2.2 to 3.4]	2.8 (0.55) [1.6 to 3.7]

DISCUSSION

Interpreting the Empirical Findings

Three findings warrant particular attention. First, the high Delphi consensus rate of 92.9%, combined with the unanimous endorsement of Contextual Relevance, indicates a shared recognition across a diverse multinational expert community that existing artificial intelligence and educational technology evaluation frameworks fail to capture the realities of deployment in low- and middle-income countries (Rodriguez-Segura, 2022; Okiri, 2024). The sensitivity analysis reported in Table 3 shows this conclusion is robust: even at the more demanding 80% threshold, the framework retains 85.7% of indicators, and Contextual Relevance remains at 100%. The Delphi consensus is therefore a substantive finding about

expert convergence on an underspecified evaluation problem rather than an artefact of a permissive threshold (Diamond et al., 2014).

Second, the findings on reliability and validity support the framework's readiness for the next stage of multi-country validation, while clearly indicating where the warrants end. Substantial interrater reliability (Fleiss' kappa = 0.81) was achieved by three evaluators trained directly by the framework developers, representing a best-case scenario rather than an estimate under ministry deployment conditions (Fleiss et al., 2003). The non-circular discriminant validity test (Cohen's d = 1.42, lower confidence bound at 0.41) provides a more honest estimate than the partly circular six-dimension figure of 1.94 originally computed, and the corrected estimate still supports the framework's ability to differentiate products with materially different quality profiles (Venable et al., 2016). Reporting both estimates transparently allows readers to see how the partial circularity in the original design inflated the apparent effect.

Third, the systematic observation that deployed products score lowest on the Data Governance, Equity and Inclusion, and Evidence of Impact dimensions, which are most critical for responsible deployment, provides an empirical baseline, to the authors' knowledge, the first of its kind for the artificial intelligence in education market in Sub-Saharan Africa (Birhane, 2020; Mohamed et al., 2020; Nemorin et al., 2023). Variability within categories qualifies this pattern. Chatbot products exhibit substantial dispersion: two of six achieve adequate Pedagogical Alignment, while four fall short. The category mean of 2.9 masks meaningful design heterogeneity that procurement processes should distinguish (Mollick & Mollick, 2024). The implication for procurement is not that all chatbots are unsuitable, but that adoption decisions need to be product-specific rather than category-wide. Two design choices appear to distinguish higher-scoring from lower-scoring chatbots in the pilot: the presence of explicit pedagogical scaffolding in the conversational flow and the use of locally produced training data for the content domains in which the product operates (Koehler et al., 2014; Luckin & Cukurova, 2019). Several alternative explanations for the pattern of low aggregate scores merit consideration.

A selection bias might suggest that the 18 pilot products are not representative; the stratified selection reduces but does not eliminate this risk. A maturity explanation would suggest that low scores reflect an early market stage, but the landscape analysis shows that some evaluated products have been deployed for 3 or more years without substantial improvement in data governance disclosure (Selwyn, 2024; Williamson et al., 2023). Two further explanations deserve attention. First, framework-induced bias: an instrument that operationalizes responsible artificial intelligence principles will naturally identify products lacking such features, especially those least incentivized by current procurement processes (Floridi et al., 2018; Jobin et al., 2019). The framework makes its analytic frame visible rather than concealed. Second, evaluator expectation: evaluators were trained in a context where these dimensions were identified as problem areas, and this priming may have influenced their scoring. Independent evaluators trained without exposure to the landscape findings would provide a stronger test.

Distinctive Contribution Relative to Existing Frameworks

The novelty of the AEQAF lies in its differentiation from three existing evaluation frameworks. Table 5 shows coverage across four frameworks along the six AEQAF dimensions. No existing framework systematically covers all six dimensions tailored to the realities of deployment in low- and middle-income countries (Holmes et al., 2022; Macgilchrist et al., 2024). EdTech Tulna offers comprehensive pedagogical alignment indicators but predates

generative artificial intelligence and does not address algorithmic transparency or model provenance (Hendrycks et al., 2020; Liang et al., 2023). EdTech Impact presents effectiveness evidence and user experience indicators, but considers data governance and equity secondary issues, and uses a regulatory baseline calibrated to UK conditions. The Systems Approach for Better Education Results in Information and Communication Technology program provides systemic policy and infrastructure indicators. However, it functions at the country system level rather than the product level. The AEQAF uniquely integrates all six dimensions at the product level and operationalizes them through 42 indicators with rubrics specific to low- and middle-income countries (Ramsarup et al., 2023).

Table 5. Comparative Dimension Coverage Across Four Evaluation Frameworks

Dimension	AEQAF (Present Study)	EdTech Tulna (India)	EdTech Impact (United Kingdom)	SABER-ICT (World Bank)
Pedagogical Alignment	Comprehensive (9 indicators)	Comprehensive	Partial	System level only
Technical Performance with AI Specifics	Comprehensive (7 indicators with AI provenance, hallucination)	Pre AI era; absent	Partial, pre-generative AI	Infrastructure focused
Contextual Relevance to LMIC	Comprehensive (6 indicators)	India specific	UK calibrated	Cross-country systemic
Data Governance for Children	Comprehensive (8 indicators with cross-border data flow)	Limited	General privacy only	Not addressed at the product level
Equity and Inclusion	Comprehensive (6 indicators)	Disability accessibility	Effectiveness across groups	Gender at the system level
Tiered Evidence of Impact	Comprehensive (6 indicators with tiered classification)	Effectiveness indicators	Strong; effectiveness-focused	Country-level outcomes
AI Specific Coverage	Comprehensive	Absent	Limited	Absent at the product level

A preliminary empirical comparison used the EdTech Tulna pedagogical rubric on a subset of six pilot products, alongside AEQAF pedagogical alignment scoring. The scores from Tulna and AEQAF showed a strong correlation (Spearman rho = 0.79), indicating convergent validity. AEQAF identified six additional concerns not captured by Tulna: three related to model provenance and hallucination management, two concerning cross-border data flow disclosure, and one on algorithmic fairness auditing (Bond et al., 2024; Holmes et al., 2022). These issues highlight features that pedagogical frameworks developed before the rise of generative artificial intelligence may overlook (Zapata-Rivera et al., 2024). Conducting a comprehensive comparative study applying all three frameworks to the 18 pilot products is a key next step.

The four contributions of AEQAF to the Educational Technology in Developing Countries research agenda are illustrated in Figure 3. The theoretical contribution integrates responsible AI principles, the Technological Pedagogical Content Knowledge (TPACK) framework, and a

decolonial, capability-based perspective into a unified scoring model that supports evaluation rather than only diagnosis (Floridi et al., 2018; Koehler et al., 2014; Mohamed et al., 2020). The methodological contribution involves a reproducible Design Science Research process that combines co-design, Delphi methodology, and pilot validation (Hevner et al., 2004; Peffers et al., 2007; vom Brocke et al., 2020). The empirical contribution provides the first systematic baseline of AI-enabled educational technology quality in three Sub-Saharan African markets (Hennessy et al., 2022; Rodriguez-Segura, 2022). The practical contribution is an openly licensed instrument that procurement agencies, development partners, and researchers can adapt to their specific contexts (Costanza-Chock, 2020).

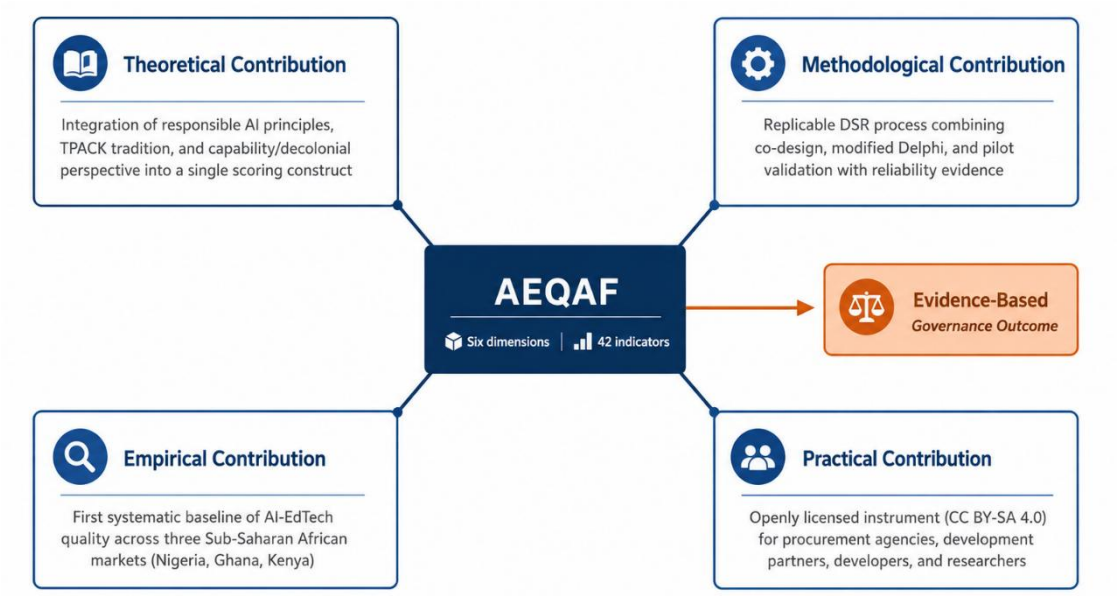

Figure 3. Four Register Contribution of the AEQAF to the Educational Technology in Developing Countries Agenda

Translating Findings into Practice for Stakeholders

The findings have practical implications for four stakeholder groups. For national education ministries in low- and middle-income countries, the AEQAF offers a structured procurement screening tool that can be incorporated into adoption, licensing, and periodic review processes (Molina et al., 2018; Okiri, 2024). This pathway is best understood as a chain with identifiable failure modes, such as regulatory capture (where framework governance is dominated by the actors it is meant to regulate) and metric gaming (where developers optimize scoreable indicators without improving core quality) (Williamson et al., 2023). The framework alone cannot prevent these failures; institutional safeguards, including independent governance, regular indicator updates, and a mix of quantitative and qualitative reviews, are essential complements that ministries should develop alongside any adoption process (Selwyn, 2024).

For development partners (the World Bank, UNESCO, UNICEF, regional development banks, and bilateral donors), the framework provides a due diligence tool for investment assessment, program design, and grant monitoring (Molina et al., 2018; Rodriguez-Segura, 2022). Its open license enables partners to adapt rubrics to their institutional priorities while maintaining comparability across programs. For developers, the framework functions not only as a quality benchmark but also as a design guide: each indicator highlights a feature relevant to quality, and developers can use the indicator set as a checklist during product design, not

just during external evaluation (Holstein et al., 2019; Luckin & Cukurova, 2019). This proactive use of the framework shifts it from a punitive to a constructive role within the ecosystem. Particular focus should be given to areas currently underinvested across deployed products, namely Data Governance, Equity and Inclusion, and Evidence of Impact, where targeted developer attention could yield significant improvements at relatively low cost (Khan et al., 2024).

For researchers, the framework supports a research agenda that has been challenging to pursue with ad hoc evaluation methods (Perrotta & Selwyn, 2020). Three key directions emerge. First, longitudinal tracking of AEQAF scores across product generations could reveal whether artificial intelligence in education products is improving in areas where it currently performs poorly. Second, linkage studies connecting AEQAF scores to learning outcomes, using randomized or quasi-experimental designs, would determine if the framework accurately predicts the outcomes it aims to justify, a question this study explicitly does not resolve (Major et al., 2021; Muralidharan et al., 2019). Third, comparative research applying the framework across different linguistic and regulatory contexts would produce calibration data for adaptation to other regions. The current findings are limited to Anglophone Sub-Saharan African settings; the core components dimension structure and pedagogical, technical, and evidence indicators are portable, but the context-specific aspects (such as data governance, equity, and relevance) will need translation and recalibration for other regulatory and linguistic environments (Ramsarup et al., 2023).

Scholarly Standing and Trajectories for Future Inquiry

The scholarly standing of the AEQAF rests on its capacity to enter conversations that have been advancing in parallel without sufficient mutual engagement. The responsible artificial intelligence literature has matured normative principles into governance guidelines, but translating those principles into evaluation instruments has been concentrated in high-income regulatory contexts (Floridi et al., 2018; Jobin et al., 2019). The critical educational technology scholarship has refined the analytic apparatus for understanding how educational technology shapes and is shaped by political economy, but has been largely diagnostic rather than constructive (Macgilchrist et al., 2024; Selwyn, 2024; Williamson et al., 2023). The decolonial artificial intelligence scholarship has named the patterns of extraction, dependency, and epistemic asymmetry that any framework operating in the Global South must reckon with, but has yet to produce widely adopted instruments (Birhane, 2020; Costanza-Chock, 2020; Mohamed et al., 2020). The AEQAF enters all three conversations by operationalising responsible artificial intelligence principles, registering the concerns of critical scholarship through its data governance and equity dimensions, and treating community contextual knowledge as evaluative authority (Ramsarup et al., 2023).

Several limitations should be acknowledged. First, the framework was developed and piloted in a specific geographic and linguistic envelope, and generalisability beyond Anglophone Sub-Saharan African contexts requires further validation (Lincoln & Guba, 1985; Nowell et al., 2017). Second, the Delphi panel, although international, overrepresented Sub-Saharan African perspectives at 37.5%, which may have biased dimension weighting toward concerns most salient in that region (Diamond et al., 2014). Third, the pilot application assessed 18 products, a meaningful but limited sample. Fourth, product evaluation was conducted at a single point in time, whereas products are dynamic systems with frequently updated content. Fifth, the framework evaluates product quality and evidence base rather than generating new evidence of learning impact; complementary impact studies are required

to link AEQAF scores to student outcomes (Major et al., 2021; Muralidharan et al., 2019). Sixth, lower agreement on Evidence of Impact (kappa = 0.73) indicates that evidence quality judgements retain meaningful subjectivity (Fleiss et al., 2003). Seventh, product evaluations relied on public information and developer disclosures, which may underrepresent products whose documentation is weaker than their actual implementation. Eighth, the cost of operating the framework in procurement at production scale has not been quantified.

Trajectories for future inquiry can be ordered by proximity to the framework's foundational validation needs. The two most pressing priorities are automated scoring tools for technically verifiable indicators and longitudinal validation against learning outcomes (Bond et al., 2024; Hendrycks et al., 2020; Liang et al., 2023). Automated scoring would reduce evaluator burden enough to make production use feasible at a ministry scale; longitudinal validation would establish whether AEQAF scores predict the educational outcomes that justify the framework's existence (Major et al., 2021). The next priority is adaptation and validation in additional linguistic and regulatory contexts (Francophone Africa, Latin America, Central Asia, Pacific Island contexts). Comparative studies across product generations and integration into national regulatory and procurement frameworks are downstream priorities (Okiri, 2024). The ordering reflects an analytic judgement that foundational validation precedes generalisation, which precedes downstream integration.

A note on positionality concludes the discussion. The author team holds no current consulting relationships with the developers of the 18 pilot products or with the agencies most likely to adopt the framework, and the framework will be released under a Creative Commons Attribution-ShareAlike 4.0 International licence to prevent exclusive control over commercial application. The framework is hosted at a public repository with version control and an independent steering committee whose composition reflects the geographic and stakeholder diversity that informed the original development. The work reported here is an opening onto a longer trajectory of community-informed refinement rather than the closure of an evaluation problem.

CONCLUSION

This Design Science Research study developed, validated, and pilot-tested the Quality Assurance Framework for Artificial Intelligence-Enabled Educational Technology, a six-dimensional, 42-indicator instrument for evaluating artificial intelligence-enabled educational technology in low-resource educational settings. The framework was developed through a 24-month iterative process that included systematic landscape analysis of 147 products across 23 markets, multinational co-design workshops with 68 stakeholders across five venues, three rounds of modified Delphi validation with 24 international experts across 14 countries, and a pilot application to 18 deployed products in Nigeria, Ghana, and Kenya. The framework achieved a Delphi consensus rate of 92.9% (robust across thresholds of 70%, 75%, and 80%), substantial interrater reliability (Fleiss kappa = 0.81), and substantial discriminant validity through a non-circular test on five framework dimensions (Cohen d = 1.42), supporting its readiness for the next phase of multi-country validation. The pilot revealed that currently deployed products in the three Sub-Saharan African markets score lowest on Data Governance, Equity and Inclusion, and Evidence of Impact, the dimensions most consequential for responsible deployment.

The study contributes to the Educational Technology in Developing Countries research agenda on four registers. Theoretically, it integrates responsible artificial intelligence

principles, the Technological Pedagogical Content Knowledge tradition, and the capability and decolonial perspectives into a coherent construct that supports scoring rather than only diagnosis. Methodologically, it provides a replicable Design Science Research process that combines co-design, Delphi, and pilot validation. Empirically, it provides the first systematic baseline of the quality of artificial intelligence-enabled educational technology across three Sub-Saharan African markets. Practically, it provides an openly licensed instrument that procurement agencies, development partners, and researchers can adapt to their own contexts. The framework is ready for multi-country deployment as a research tool, while falling short of the warrants required for unsupervised policy use, which the next phase of validation must address.

The framework is licensed under Creative Commons Attribution-ShareAlike 4.0 International, with version 1.0 released alongside this publication. Future research should prioritise, in order, two foundational priorities (automated scoring for technically verifiable indicators and longitudinal validation against learning outcomes), generalisability priorities (adaptation to Francophone Africa, Lusophone Africa, Latin America, Central Asia, and Pacific Island contexts), and downstream priorities (comparative studies across product generations and integration into national regulatory frameworks). The longer-term goal is the operationalisation of evidence-based governance for artificial intelligence in education across the Global South, a goal that no single research team can complete, and the open licensing of the AEQAF is intended to make it collaboratively tractable.

REFERENCES

Birhane, A. (2020). Algorithmic colonization of Africa. SCRIPTed, 17(2), 389–409. https://doi.org/10.2966/scrip.170220.389

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., ... & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International journal of educational technology in higher education*, *21*(1), 4. https://doi.org/10.1186/s41239-023-00436-z

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Costanza-Chock, S. (2020). Design justice: Community-led practices to build the worlds we need. MIT press. https://doi.org/10.7551/mitpress/12255.001.0001

Diamond, I. R., Grant, R. C., Feldman, B. M., Pencharz, P. B., Ling, S. C., Moore, A. M., & Wales, P. W. (2014). Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. *Journal of clinical epidemiology*, *67*(4), 401-409. https://doi.org/10.1016/j.jclinepi.2013.12.002

Eybers, S. (2023, August). Design Science Research in information systems as educational technology in teaching and learning environments: a systematic literature review. In *International Conference on Innovative Technologies and Learning* (pp. 385-402). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-40113-8_38

Fahd, K., Miah, S. J., Ahmed, K., Venkatraman, S., & Miao, Y. (2021). Integrating design science research and design based research frameworks for developing education support systems. *Education and Information Technologies*, *26*(4), 4027-4048.https://doi.org/10.1007/s10639-021-10442-1

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). Statistical methods for rates and proportions (3rd ed.). Wiley.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People: An ethical framework for a good artificial intelligence society. *Minds and Machines*, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2), 337–355. https://doi.org/10.25300/MISQ/2013/37.2.01

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*. https://doi.org/10.48550/arXiv.2009.03300

Hennessy, S., D'Angelo, S., McIntyre, N., Koomar, S., Kreimeia, A., Cao, L., ... & Zubairi, A. (2022). Technology use for teacher professional development in low-and middle-income countries: A systematic review. *Computers and education open*, *3*, 100080. https://doi.org/10.1016/j.caeo.2022.100080

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. https://doi.org/10.2307/25148625

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., ... & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, *32*(3), 504-526. https://doi.org/10.1007/s40593-021-00239-1

Holstein, K., McLaren, B. M., & Aleven, V. (2019). Co designing a real time classroom orchestration tool to support teacher artificial intelligence complementarity. *Journal of Learning Analytics*, 6(2), 27–52. https://doi.org/10.18608/jla.2019.62.3

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of artificial intelligence ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Khan, M. S., Umer, H., & Faruqe, F. (2024). Artificial intelligence for low income countries. *Humanities and Social Sciences Communications*, *11*(1), 1-13. https://doi.org/10.1057/s41599-024-03947-w

Koehler, M. J., Mishra, P., Kereluik, K., Shin, T. S., & Graham, C. R. (2014). The technological pedagogical content knowledge framework. In *Handbook of research on educational communications and technology* (pp. 101-111). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-3185-5_9

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Re, C., Acosta-Navas, D., Hudson, D. A., … Koreeda, Y. (2023). Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023(November). https://doi.org/10.48550/arXiv.2211.09110

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

Luckin, R., & Cukurova, M. (2019). Designing educational technologies in the age of artificial intelligence: A learning sciences driven approach. *British Journal of Educational Technology*, 50(6), 2824–2838. https://doi.org/10.1111/bjet.12861

Macgilchrist, F., Potter, J., & Williamson, B. (2024). Challenging the inequitable impacts of educational technology. Learning, Media and Technology, 49(2), 147–150. https://doi.org/10.1080/17439884.2024.2350117

Major, L., Francis, G. A., & Tsapali, M. (2021). The effectiveness of technology supported personalised learning in low and middle income countries: A meta analysis. *British*

Journal of Educational Technology, 52(5), 1935–1964. https://doi.org/10.1111/bjet.13116

Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial artificial intelligence: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy and Technology*, 33(4), 659–684. https://doi.org/10.1007/s13347-020-00405-8

Molina, E., Pushparatnam, A., Rimm-Kaufman, S. E., & Wong, K. K. Y. (2018). Evidence-based teaching: Effective teaching practices in primary school classrooms. *World Bank Policy Research Working Paper*, (8656). https://doi.org/10.1596/1813-9450-8656

Mollick, E., & Mollick, L. (2024). Instructors as innovators: A future-focused approach to new AI learning opportunities, with prompts. *arXiv preprint arXiv:2407.05181*. https://doi.org/10.2139/ssrn.4802463

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology aided instruction in India. *American Economic Review*, 109(4), 1426–1460. https://doi.org/10.1257/aer.20171112

Nemorin, S., Vlachidis, A., Ayerakwa, H. M., & Andriotis, P. (2023). Artificial intelligence hyped? A horizon scan of discourse on artificial intelligence in education and development. *Learning, Media and Technology*, 48(1), 38–51. https://doi.org/10.1080/17439884.2022.2095568

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods*, *16*(1), 1609406917733847. https://doi.org/10.1177/1609406917733847

Okiri, P. O., & Hercz, M. (2024). Distributed pedagogical leadership practice for sustainable pedagogical improvement: A literature review (2010–2023). *European Journal of Education*, *59*(4), e12723.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45-77. https://doi.org/10.2753/MIS0742-1222240302

Perrotta, C., & Selwyn, N. (2020). Deep learning goes to school: Toward a relational understanding of AI in education. *Learning, media and technology*, *45*(3), 251-269. https://doi.org/10.1080/17439884.2020.1686017

Ramsarup, P., McGrath, S., & Lotz-Sisitka, H. (2023). Reframing skills ecosystems for sustainable and just futures. *International Journal of Educational Development*, *101*, 102836. https://doi.org/10.1016/j.ijedudev.2023.102836

Rodriguez-Segura, D. (2022). EdTech in developing countries: A review of the evidence. *The World Bank Research Observer*, *37*(2), 171-203. https://doi.org/10.1093/wbro/lkab011

Selwyn, N. (2024). On the limits of artificial intelligence (AI) in education. *Nordisk tidsskrift for pedagogikk og kritikk*, *10*(1). https://doi.org/10.23865/ntpk.v10.6062

Steinherr, V. M., Brehmer, M., Stöckl, R., & Reinelt, R. (2024, May). Design science research as a guide for innovative higher education teaching: towards an application-oriented extension of the proficiency model. In *International Conference on Design Science Research in Information Systems and Technology* (pp. 213-228). Cham: Springer Nature Switzerland. https://www.springerprofessional.de/en/design-science-research-as-a-guide-for-innovative-higher-educati/27136128

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: a framework for evaluation in design science research. *European journal of information systems*, *25*(1), 77-89. https://doi.org/10.1057/ejis.2014.36

vom Brocke, J., Winter, R., Hevner, A., & Maedche, A. (2020). Accumulation and evolution of design knowledge in design science research: A journey through time and space. *J. Assoc. Inf. Syst*, *21*(9). https://doi.org/10.17705/1jais.00611

Williamson, B., Macgilchrist, F., & Potter, J. (2023). Re-examining AI, automation and datafication in education. *Learning, media and technology*, *48*(1), 1-5. https://doi.org/10.1080/17439884.2023.2167830

Zapata-Rivera, D., Torre, I., Lee, C. S., Sarasa-Cabezuelo, A., Ghergulescu, I., & Libbrecht, P. (2024). Generative AI in education. *Frontiers in Artificial Intelligence*, *7*, 1532896. https://doi.org/10.3389/frai.2024.1532896