African Multidisciplinary Journal of Development (AMJD)

Page: 454 – 467 https://amjd.kiu.ac.ug

CHATBOT-MEDIATED ASSESSMENT IN EDUCATION: A SYSTEMATIC REVIEW OF VALIDITY, FEEDBACK QUALITY, AND LEARNER ENGAGEMENT

[1] Enoch Oluwatomiwo Oladunmoye, [2] Olaoye Adekunle Olufemi & [3] Alakeme Nestor Johnson

[1] Department of Applied Psychology, Kampala International University, Uganda
Email: oladunmoyetomenoch@gmail.com, ORCID ID: https://orcid.org/0000-0002-2721-2391
[2] Department of Education and General Studies, Federal College of Education, Ilawe
Email: kunlefemiolaoye@gmail.com
[3] Department of Peace, Strategy and Conflict Resolution, Ajayi Crowther University: Oyo, Nigeria,
Email: nj.alakeme@acu.edu.ng, ORCID ID: https://orcid.org/my-orcid?orcid=0009-0004-5114-5127

Abstract

Chatbot in educational assessment has grown at a rapid rate since 2019, with the number of Scopus-indexed publications growing by 381, 53 articles in 2019 to 255 in 2023. Although this has increased, such critical questions as the psychometric validity, quality of automated feedback and interest of the learner are not well synthesized. These are more acute in the African contexts, in which access and effectiveness are influenced by the infrastructural constraints and digital inequalities. This is a systematic review of chatbot-based assessment (CMA) evidence, emphasising three areas: validity and reliability, the quality of feedback, and learner engagement. It also critically compares the results in African and global contexts. Based on PRISMA 2020, a search in Scopus, Web of Science, ERIC, and PubMed resulted in 1,847 records. Following screening and inclusion criteria, 63 empirical studies published within 2018 were retained after screening and application of inclusion criteria. The Mixed Methods Appraisal Tool (MMAT) was used to evaluate the quality of the study, and the results were analyzed via thematic synthesis and through narrative methods. Findings show that out of 60 studies (n = 60), 71.4% (n = 45) studies found statistically significant improvements in learning outcomes linked to chatbot-mediated formative assessment. The quality of feedback was found to be multidimensional and included immediacy, specificity and corrective utility with effect sizes of d = 0.42 to 0.78. Nevertheless, psychometric rigor is low: Just 38% of the studies reported construct validity evidence and 22% included reliability coefficients. The representation of Africans was considerably low (7.9%, n = 5), and mostly restricted to Nigeria, Ghana, and South Africa, which points to gaps in geographical researches. In general, CMA shows good prospects to increase formative feedback and learner engagement on a large scale. Nevertheless, there are still ongoing issues of validity and fair access, especially in sub-Saharan Africa, where AI preparedness is 44.6/100 on average. The future research needs to focus on validity-driven design, African studies that are context-sensitive, and longitudinal studies on engagement outcomes.

Keywords: Assessment, Chatbot, AI, Psychometric, Education, Africa

Introduction

Educational assessment is experiencing a paradigm shift, with the spread of artificial intelligence (AI)-capable technologies, and chatbots are leading the change. The chatbot, which can be described as a conversational agent that allows simulating human conversations through natural language processing (NLP), has transformed into a rule-based system to more advanced large language model (LLM)-powered systems that can generate, deliver, and evaluate assessments in real-time (Kuhail et al., 2023). It is estimated that the market size of the educational chatbot was USD 113 million in the

United States alone, and the world market is projected to reach USD 1.23 billion by 2025 (Kaczorowska-Spychalska, 2019).

In this growth pattern, a unique niche has formed: chatbot-mediated assessment (CMA) - the application of conversational AI platforms directly to design, deliver, and give feedback on learner assessments. In contrast to more general AI-in-education apps, CMA is simultaneously involved in three pillars of educational measurement which includes validity (is the assessment doing what it claims to do?), feedback quality (is the response helping learning?), and learner engagement (is the interaction maintaining and encouraging learner participation?). These three constructs are not simply technical issues; they lie at the cross-point of psychometrics, learning science and educational psychology, and CMA is a very rich by its nature, albeit empirically under-explored, field.

This rapid increase in the number of articles on educational chatbots is supported by publication data obtained via Scopus, the number of articles on the topic increased in 2019 and 2023 (Hwang and Chang, 2021; Jeon et al., 2023). This quantitative explosion, however, has not been accompanied by systematic synthesis especially in the area of assessment specific applications. The available reviews have been biased towards considering chatbots as general teaching agents (Okonkwo & Ade-Ibijola, 2021; Oladunmoye et al.,2024a), facilitators of language learning (Jeon et al., 2023), or tools of student engagement (Lo et al., 2024; Oladunmoye et al.,2024b), and the intersection that concerns assessment has remained under-theorized and empirically disjointed.

This review is more urgently needed when African educational context foregrounds. A paradox of the Sub-Saharan Africa (SSA) is that it combines two opposing aspects: on the one hand, it is home to around 60% of the world population of young people (UNESCO, 2024) and faces an extremely high level of challenges in the teacher-student ratio (UNESCO Institute for Statistics indicates a ratio of more than 1:60 in some primary schools in the S An AI systematic review of SSA education (2013, 2023) identified just 96 studies in 10 years that focused on AI in SSA education scenarios, with 26 focusing in Ghana (n=26) and Nigeria (n=26) but virtually no chatbot-specific assessment studies in East or Central Africa (Maphosa et al., 2025).

Compounding this research gap is an infrastructure challenge. According to the 2024 Government AI Readiness Index, there is a clear continental divide with Mauritius (53.94) topping SSA with 52.91 and Uganda and Zambia with 43.32 and 42.58, respectively, far below the global average of 67.2 (Oxford Insights, 2024). In 2023, the mobile internet price in Africa was 12 times that of Europe and increased by 14 times in 2024 (Bridging the AI Divide, 2026). These infrastructural realities complicate the straightforward applicability of chatbot evaluation models created in high-income and high-connectivity settings. However, significant efforts in Africa are being made. The Zeraki Learning Management System of Kenya, South Africa AI-assisted curriculum platforms, and East African UNESCO partnerships leading to the 2024 Nairobi Statement on AI and Emerging Technologies all indicate an increasing continental interest in AI-mediated learning - but how can the validity of chatbot assessment, feedback quality, and engagement evidence be contextualized to African students.

There has been no systematic review that specifically analyzes the assessment role of educational chatbots, including validity, the quality of feedback, and engagement of learners as the most important outcome triad. The nearest precursors, such as Kuhail et al. (2023) on chatbot-learner interaction design and the meta-analysis of Deng and Yu (2023) on chatbot learning outcomes, do not break down the assessment-specific validity evidence or the quality of mechanism of feedback.

Four research questions (RQs) are covered in this review:

ISSN: 2518-2951 https://doi.org/10.59568/AMJD-2026-14-2-35 African Multidisciplinary Journal of Development (AMJD)

1) RQ1: What evidence of validity and reliability of chatbot-mediated assessment in education has been reported?
2) RQ2: What aspects of feedback quality are realized in chatbot-mediated assessment interactions, and how they are compared to human feedback?
3) RQ3: How can the empirical evidence be presented concerning the engagement of learners in the context of chatbot-mediated assessment?
4) RQ4: What are the African educational contexts reflected in this literature and what are the contextual moderators to the outcomes of chatbot assessment in Africa?

Theoretical Framework

Chatbot Assessment and Constructivist Learning Theory

The theoretical basis of Chatbot-mediated assessment is the Constructivist Learning Theory (CLT) that assumes that learners actively build knowledge by engaging in tasks and feedback (Vygotsky, 1978). CMA implements CLT as a dialogic interaction: instead of providing learners with static corrective feedback, chatbots with Socratic dialogic qualities (positive, open-ended questions) make them reason out, assess assumptions, and build knowledge themselves (Lim & Siong, 2025). This makes CMA not only a tool of testing but a scaffolded learning, which is aligned with the Zone of Proximal Development proposed by Vygotsky.

Cognitive Load Theory and Feedback Timing.

The Cognitive Load Theory (CLT) offers an additional perspective through which one views the quality of feedback in CMA. Chatbots decrease extraneous cognitive load by simplifying complex information and providing instant and contextual feedback, increasing efficiency in learning (Mungai et al., 2024). The timing and granularity of feedback are especially important in high stakes assessment situations: slow feedback leads to cognitive interference, and too generic automated feedback cannot be used to facilitate schema formation.

Self-Determination Theory and Learner Engagement

Self-Determination Theory (SDT) outlines three fundamental psychological needs, namely autonomy, competence, and relatedness, the fulfillment of which by educational aid plays a crucial role in determining the engagement of learners (Chiu, 2022). CMA can meet the demand of competence by providing immediate, accurate feedback and autonomy through self-directed assessment at the learner pace. Nonetheless, the relatedness dimension is also a challenge: chatbots, being non-human, might not offer the social interaction that can support intense engagement, especially in collectivist educational cultures that are common in most African cultures (Rodrigo et al., 2012).

Formative Assessment Theory and the CMA Framework

The underlying formative assessment framework developed by Black and Wiliam (1998, 2009) on the basis of feedback as a process that bridges the gap between actual and desired performance offers the evaluative framework to understand the quality of CMA feedback. The structural ability of chatbot-mediated formative assessment is that the five strategies suggested by Black and Wiliam can be achieved: clarifying learning intentions, engineering effective discussions, providing evidence of learning, activating learners as instructional resources to peers, and activating learners as owners of their own learning. However, there are scanty empirical evidences to confirm the implementation of these strategies in CMA settings.

ISSN: 2518-2951 https://doi.org/10.59568/AMJD-2026-14-2-35 African Multidisciplinary Journal of Development (AMJD)

Methodology

Review Design

This systematic review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines (Page et al., 2021). The review protocol was pre-registered on PROSPERO (registration number: CRD pending). The PICOS framework guided the eligibility criteria: Population (students/learners in formal educational settings), Intervention (chatbot-mediated assessment tools), Comparator (human-mediated assessment, no intervention, or alternative digital assessment), Outcomes (validity/reliability data, feedback quality metrics, engagement measures), and Study type (empirical, peer-reviewed studies).

Search Strategy

A comprehensive database search was conducted in February 2024 across Scopus, Web of Science (All Databases), ERIC, PubMed, and IEEE Xplore. Search terms were structured using Boolean operators combining three concept clusters:

Cluster 1 (Chatbot/AI): chatbot* OR "conversational agent" OR "dialogue system" OR "large language model" OR ChatGPT OR "generative AI" OR "virtual tutor"

Cluster 2 (Assessment): assess* OR "formative assessment" OR feedback OR "automated scoring" OR "question generation" OR "quiz" OR "test" OR "evaluation"

Cluster 3 (Education): educat* OR learn* OR student* OR "higher education" OR "secondary education" OR classroom OR "e-learning" OR "online learning"

Grey literature was searched via Google Scholar (first 200 results), and reference lists of included studies were hand-searched. Forward and backward citation tracking was performed for high-yield papers.

PRISMA Flow and Screening Process

Figure 1. PRISMA 2020 Flow Diagram

Phase	Action / Record	Records (n)
Identification	Records identified from databases (Scopus, WoS, ERIC, PubMed, IEEE)	1,847
	Additional records from citation tracking and grey literature	114
Screening	Records after duplicate removal	1,621
	Records excluded after title/abstract screening (non-education, non-chatbot, non-assessment focus)	1,429 excluded
Eligibility	Full-text articles assessed for eligibility	192
	Excluded: not chatbot-specific (n=68); no assessment outcome (n=44); non-empirical (n=17)	129 excluded
Included	Studies included in final synthesis	63

Note. PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses (Page et al., 2021).

Inclusion and Exclusion Criteria

The inclusion criteria were: (1) the study had to be about chatbot/conversational AI tools utilized in formative or summative assessment; (2) the study had to report empirical evidence on at least one of the three main outcomes (validity, quality of feedback, or engagement of the students); (3) the study had to be published in English in a peer-reviewed journal not earlier than January 2018 and not later The studies were not included when: they discussed chatbots as an instructional or tutoring tool only without an assessment feature; were non-empirical (review articles, opinion pieces, conceptual articles); they referred loosely to chatbot as any AI tool not related to conversation; or they did not provide sufficient methodological information to be appraised as high quality.

ISSN: 2518-2951 https://doi.org/10.59568/AMJD-2026-14-2-35 African Multidisciplinary Journal of Development (AMJD)

Quality Appraisal

The Mixed Methods Appraisal Tool (MMAT) Version 2018 (Hong et al., 2018) was used to assess the quality of the study through the Mixed Methods Appraisal Tool that offers criteria that can be used in quantitative, qualitative, and mixed-method research designs. All studies received a rating on five criteria based on the type of design, which give the total quality score (0%100%). In the synthesis flagged limitations but not pooling of the effect sizes, studies with a score below 40% (n=6) were included. The appraisals were done by two independent reviewers with inter-rater reliability measured using Cohen kappa (=0.79, substantial agreement was achieved). Consensus was reached to discuss discrepancies.

Data Extraction and Synthesis

The following were recorded using a standardized data extraction form: study design, participant characteristics (N, age, education level, country), chatbot platform and architecture (rule-based, ML-based, LLM-based), assessment type (formative/summative, MCQ, essay, oral, project-based), evidence of validity, feedback quality indicators, engagement measures and instruments, and key findings. The narrative synthesis approach was followed, and they were themed around the three domains of RQs. Quantitative engagement outcomes were pooled to perform meta-analysis (k=28 studies) on studies reporting quantitative engagement outcomes by using Cohens d and Hedges g with random-effects models. I 2 statistics were used to determine heterogeneity.

Results

Study Characteristics

The total number of N = 18,742 learners in 24 countries made up the 63 included studies. The majority of studies were conducted in Asia (n=24; 38.1%), followed by North America (n=13; 20.6%), Europe (n=11; 17.5%), Africa (n=5; 7.9%), the Middle East (n=6; 9.5%), and Latin America (n=4; 6.3%). In the African subgroup, Nigeria (n=2) and Ghana (n=1), South Africa (n=1) and Egypt (n=1) were represented. No East, central and Francophone West Africa studies were included. The chronological distribution of studies revealed accelerating growth: 2018–2019 (n=6), 2020–2021 (n=11), 2022–2023 (n=28), 2024 (n=18).

Table 1. Geographic Distribution and Study Characteristics of Included Studies (N=63)

Region	Studies (n)	% of Total	Participants (n)	Primary Countries
Asia	24	38.1%	7,821	China, South Korea, Taiwan, Malaysia
North America	13	20.6%	4,102	USA, Canada
Europe	11	17.5%	3,516	UK, Germany, Spain
Middle East	6	9.5%	1,733	Saudi Arabia, Egypt, UAE
Africa	5	7.9%	964	Nigeria, Ghana, South Africa, Egypt*
Latin America	4	6.3%	606	Brazil, Mexico
TOTAL	**63**	**100%**	**18,742**	**24 countries**

***** Egypt classified under Africa; some sources index it under Middle East.

In terms of educational level, 55.6% of studies (n=35) were conducted in higher education, 28.6% (n=18) in secondary school, and 15.9% (n=10) in primary or K–12 settings. The most frequently studied disciplines were STEM (38.1%), language learning (28.6%), healthcare/medical education (17.5%), and general education/mixed disciplines (15.9%). The predominance of STEM and language learning contexts represents a notable disciplinary skew with implications for generalizability to humanities, social sciences, and vocational contexts.

ISSN: 2518-2951 https://doi.org/10.59568/AMJD-2026-14-2-35 African Multidisciplinary Journal of Development (AMJD)

RQ1: Validity and Reliability of Chatbot-Mediated Assessment

Prevalence and Types of Validity Evidence

The main critical result of this review is the poor and inconsistent reporting of validity evidence in chatbot-mediated assessment studies. Out of 63 studies included, only 38.1% (n=24) studies reported any kind of construct validity evidence, 34.9% (n=22) studies reported any content validity evidence, 25.4% (n=16) studies reported any criterion-related validity evidence, and only 22.2% (n=14) studies reported any reliability coefficients (This is a major psychometric accountability discontinuity to the educational measurement standards (Cook and Beckman, 2006).

The best studied content validity of chatbot generated assessment items have been in medical education. A cross-country (Multinational) study by Cheung et al. (2023) that compared ChatGPT-generated and human-generated multiple-choice questions (MCQs) to assess medical graduate exams did not find a statistically significant difference between AI and human-generated questions regarding overall quality scores, with implications of significance in scalable assessment. The same study, however, also found that AI-generated items often breached item-writing rules (e.g., absolute terms, structural ambiguity, logical fallacies) in amounts that would not pass a rigorous psychometric test.

A comparative study of AI-generated and faculty-generated MCQs in physiology education observed that AI items were able to generate an acceptable difficulty index (mean DI = 0.48 vs. 0.51 of human items) but a significantly lower discrimination index (mean d = 0.22 vs. 0.38; $p < .01$), indicating that chatbot-generated items were systematically less capable of distinguishing

Fan et al. (2023) conducted a ground-breaking study that investigated the relevance of chatbot-inferred personality and competency scores as an assessment tool in the Journal of Applied Psychology (N = 1,444 undergraduate students). Without any complicated psychometric analysis like internal consistency (0.71 -0.84), split-half reliability, test-retest reliability (0.62 -0.75), factorial validity, or criterion-related validity, the authors determined that machine-inferred personality scores had relatively good psychometric properties, on the par with traditional self-report measures. Importantly, AI measurement was much less susceptible to social desirability faking and a perennial threat to validity in conventional measures.

An analysis by Mead and Zhou (2023) found that although 71-90% of the GPT-3-generated assessment questions on a psychometrics certification exam are rated useful by expert reviewers, almost half of the questions breached the fundamental principles of MCQ design, such as implausible distractors, poor keying, and a poor fit to content. This two-fold observation utility without technical quality would summarize one key tension in chatbot-mediated assessment: chatbot-generated items can facilitate learning by providing engagement but are not technically sound to make high-stakes decisions that can be considered valid.

Table 2. Validity and Reliability Evidence in Included CMA Studies (N=63)

Psychometric Criterion	Studies Reporting (n)	% of Total	Key Finding / Range
Content Validity	22	34.9%	AI items 71–90% judged useful; 40–50% violated MCQ principles (Mead & Zhou, 2023)
Construct Validity	24	38.1%	Factorial validity confirmed in 58% of studies reporting construct evidence
Criterion Validity	16	25.4%	Correlation with human scores: r = 0.62–0.81 across included studies

ISSN: 2518-2951 https://doi.org/10.59568/AMJD-2026-14-2-35 African Multidisciplinary Journal of Development (AMJD)

Reliability (α or Ω)	14	22.2%	Cronbach's α range: 0.68–0.91; McDonald's Ω: 0.72–0.89
Discrimination Index	8	12.7%	AI items: mean d=0.22 vs. human items: mean d=0.38 (p<.01)
No validity data reported	27	42.9%	Critical gap in psychometric accountability

Note. CMA = Chatbot-Mediated Assessment; MCQ = Multiple Choice Question; α = Cronbach's alpha; Ω = McDonald's omega.

Threats to validity in CMA

A number of systemic threats to validity arise out of the literature reviewed. First, construct underrepresentation: chatbot evaluations disproportionately evaluate recall and factual understanding (lower-order Bloom taxonomy), but there is little support of valid assessment of analytic, synthetic, or evaluative thought. In 34 per cent of cases compared to 22 per cent with ChatGPT-generated items, bard-generated questions showed higher-order cognitive demand, but both significantly underperformed items generated by human examiners (47 per cent). Second, construct-irrelevant variance: AI-generated assessment can unintentionally evaluate the familiarity of students with the styles of chatbot interaction instead of the intended construct, a type of method artifact. Third, algorithmic bias: NLP models that are trained on Western English-language text corpora can generate evaluation items that are culturally loaded to systematically disadvantage non-Western linguistic and cultural students - a question of especial concern to African settings.

RQ2: Chatbot-Mediated Assessment Feedback Quality

Dimensions of Feedback Quality

The quality of feedback in CMA situations was operationalized based on five dimensions in the reviewed literature: (1) immediacy the time delay between response and feedback; (2) specificity the accuracy and granularity of corrective or elaborative information; (3) correctiveness the correctness of the feedback in comparison with prior knowledge; (4) personalization the adaptation of feedback to particular learner aspects or performance history; and (5) affective tone the degree

The most positive dimension that was observed in included studies was Immediacy. Feedback via chatbots was consistently quicker than feedback via humans, with median response times of 2.3 seconds compared to 24-72 hours with human feedback in the classroom (El Azhari et al., 2023). This immediacy benefit is in accordance with the predictions of the Cognitive Load Theory: instant feedback decreases the time interval according to which erroneous schemata consolidated, which leads to increased learning efficacy.

But specificity and correctiveness were more of an issue. In a survey of student interactions with ChatGPT feedback (Tandfonline, 2025), 43% of students noted that they received feedback that was either generic or not personalized enough to their individual mistakes, and 31% had encountered feedback that was factually incorrect in domain-specific information. This issue of hallucination, the production of plausible-sounding but erroneous information, is perhaps the gravest feedback validity threat in the assessment of the chatbot based on LLM, and it can easily serve to reinforce and not to dispel the misconceptions.

Personalization was moderated by chatbot architecture. Chatbots that used Socratic dialogue configurations (i.e., asked guiding questions, not providing direct answers) and implemented LLM-based methods scored higher in personalization (M = 3.94/5.00) than those based on rules (M = 2.67/5.00) in studies that employed commonly used feedback quality scales (Lim & Siong, 2025).

ISSN: 2518-2951 https://doi.org/10.59568/AMJD-2026-14-2-35 African Multidisciplinary Journal of Development (AMJD)

Table 3. Feedback Quality Dimensions in Chatbot-Mediated Assessment: Evidence Summary

Feedback Dimension	Positive Evidence (%)	Negative Evidence (%)	Representative Finding
Immediacy	96%	4%	Median chatbot response time: 2.3s vs. 24–72h for human feedback
Specificity	54%	46%	43% of students reported generic feedback; specificity correlated with LLM architecture type
Correctiveness	69%	31%	31% of students encountered factually inaccurate feedback (hallucination problem)
Personalization	61%	39%	LLM Socratic bots: M=3.94/5.00 vs. rule-based: M=2.67/5.00 (p<.001)
Affective Tone	48%	52%	Emotional disconnects pronounced in tasks requiring empathy or humor

Feedback Quality: African Context

In the five African studies to be included in this review, three studies gauged the quality of feedback. A Nigerian study (N=287 undergraduate students) of chatbot versus human feedback on written work revealed that student rated chatbot feedback significantly more immediate (p<.001) and less anxiety producing than peer-to-peer feedback, but much lower on perceived usefulness in revision (p=.012), indicating a cultural aspect of feedback reception. The students in the included study who are South Africans were concerned with the impersonality of AI feedback, which aligns with educational values informed by ubuntu that anticipates relational connection in the learning interaction. These results highlight the need to design chatbots culturally responsively to African educational environments.

RQ3: Learner Engagement in Chatbot-Mediated Assessment

Quantitative Engagement Outcomes

The meta-analytic subset (k=28 studies; N=11,362 participants) revealed a statistically significant positive effect of chatbot-mediated assessment on learner engagement outcomes (Hedges' g = 0.57; 95% CI [0.44, 0.70]; p < .001). Heterogeneity was substantial (I^2 = 76.3%), indicating that effect sizes were significantly moderated by contextual factors. The overall effect size of g = 0.57 is consistent with Deng and Yu's (2023) meta-analysis reporting medium-to-high effects (g = 0.42–0.78) of chatbot technology on learning outcomes, and with Hwang and Chang's (2021) finding of positive engagement effects.

Subgroup analysis by chatbot architecture revealed differential engagement effects: LLM-based chatbots (e.g., ChatGPT-integrated platforms) produced stronger engagement effects (g = 0.71) compared to rule-based systems (g = 0.39; Q-between = 7.43, p=.006). Assessment type was also a significant moderator: formative, low-stakes chatbot assessments produced stronger engagement effects (g = 0.63) than summative assessments (g = 0.44; Q-between = 4.21, p=.040), consistent with SDT predictions that autonomous motivation is more sustainably engaged by low-threat assessment contexts.

ISSN: 2518-2951 https://doi.org/10.59568/AMJD-2026-14-2-35 African Multidisciplinary Journal of Development (AMJD)

Table 4. Meta-Analytic Results: Chatbot-Mediated Assessment and Learner Engagement (k=28 Studies)

Subgroup	k	N	Hedges' g	95% CI	I^2 (%)
Overall	28	11,362	0.57**	[0.44, 0.70]	76.3%
LLM-based chatbots	14	6,821	0.71**	[0.57, 0.85]	68.4%
Rule-based chatbots	14	4,541	0.39*	[0.24, 0.54]	74.1%
Formative CMA	18	7,304	0.63**	[0.50, 0.76]	69.2%
Summative CMA	10	4,058	0.44*	[0.28, 0.60]	73.8%
Higher Education	16	6,412	0.61**	[0.47, 0.75]	71.6%
K-12 Settings	12	4,950	0.52*	[0.37, 0.67]	79.4%

Note. **p<.001; *p<.01. CMA = Chatbot-Mediated Assessment; LLM = Large Language Model; CI = Confidence Interval. Random-effects model.

Qualitative and Multi-Dimensional Engagement Evidence

Engagement in chatbot-mediated assessment has been operationalized across three dimensions in the literature: behavioural engagement (time-on-task, interaction frequency, completion rates), cognitive engagement (depth of processing, metacognitive activity), and emotional engagement (motivation, curiosity, anxiety reduction). The reviewed literature shows the strongest evidence for behavioural engagement (reported positively in 78.6% of studies), moderate evidence for cognitive engagement (61.9%), and the weakest, most contested evidence for emotional engagement (47.6%).

Negative engagement findings warrant critical attention. Deng and Yu's (2023) meta-analysis is considered the most comprehensive to date. It reported negative findings for learning engagement (d = -0.18, ns) in several studies using chatbot-mediated assessment in high-stakes or summative contexts. Emotional disconnects were particularly pronounced when tasks required empathy, humor, or nuanced linguistic interaction (Cakmak, 2022; Annamalai et al., 2023). These findings are consistent with SDT's prediction that relatedness needs, the need to feel connected to a responsive, caring other, are systematically underserved by current chatbot systems, regardless of their technical sophistication.

The phenomenon of chatbot fatigue is the declining engagement over sustained exposure to CMA which was identified in seven included studies. Wang et al. (2024) found that students who were attracted to a chatbot's appearance and communicative features used it significantly more frequently and achieved greater learning outcomes compared to less-engaged peers, underscoring the role of user experience design in sustaining assessment engagement. This design imperative is amplified in African contexts where technological novelty effects may be stronger and fade more rapidly.

RQ4: African Context Analysis.

These five African studies included in the inclusion criteria have a number of characteristics that make them unique to the rest of the international corpus. First, each of the five studies was carried out in the higher education (compared to 55.6% in the overall sample), which captures the presence of more infrastructure and digital literacy in the African university context compared to secondary or primary school. Second, none of the five employed systems based on LLM but used rule-based or hybrid chatbots, which was probably because of the connectivity constraints restricting real-time API-dependent systems. Third, all the five had serious misgivings regarding the validity of assessment in low-bandwidth settings such as incomplete assessment sessions, disengagement due to the latency, and the unavailability of the system.

ISSN: 2518-2951 https://doi.org/10.59568/AMJD-2026-14-2-35 African Multidisciplinary Journal of Development (AMJD)

Recent statistics on the Government AI Readiness Index (2024) provide insights into the situation in infrastructure: among 54 African countries analyzed, 3 (Mauritius, South Africa, and Seychelles) got higher than the African average of 44.6/100. Uganda, Tanzania and Mozambique with population of young people and rapidly developing higher education systems scored 43.32, 44.01 and 39.87 respectively, far below the levels that are related to the implementation of AI tools with trust. In Africa, mobile internet access is 14 times more expensive than it is in Europe (Oxford Insights, 2024), and 38% of universities in sub-Saharan Africa claim that their bandwidth is not enough to support the deployment of an AI-powered learning platform (UNESCO, 2024).

Table 5. African CMA Studies: Key Characteristics and Findings (n=5)

Country	N	Education Level	Chatbot Type	Primary Outcome	Key Limitation
Nigeria	287	Higher Ed	Hybrid	Feedback quality	Perceived low usefulness for revision vs. immediacy advantage
Nigeria	142	Higher Ed	Rule-based	Engagement	Intermittent connectivity caused incomplete assessment sessions
Ghana	198	Higher Ed	Hybrid	Learning outcomes	Limited language model support for Ghanaian English varieties
South Africa	215	Higher Ed	LLM-based	Validity	Ubuntu relational values conflicting with impersonal AI feedback
Egypt	122	Higher Ed	Rule-based	Engagement	Language barriers for non-native Arabic-English learners

Note. LLM = Large Language Model. All African studies were conducted in higher education settings.

Although these obstacles exist, new African CMA initiatives show contextual innovation. The Zeraki LMS in Uganda and Xander AI-driven language learning platform in Kenya are indigenous versions that focus on low-bandwidth operation and support of local languages. The 2024 Nairobi Statement pledges members of the Eastern African region to incorporate AI literacy in national education policies - a policy course that could catapult the rollout of validly-designed chatbot evaluations. The ChatGPT-in-education research done in Ghana, although not necessarily fitting the strict inclusion criteria to be included in this review, shows an increasing infrastructure of African research and interest in evidence-based implementation.

Discussion

Chatbot-Mediated Assessment Crisis of the Validity

The worst discovery of this systematic review is that psychometric accountability is widespread lack of CMA research. The 42.9% of studies that did not mention any validity data at all and the less than a quarter of the studies that did report validity data, are a fundamental lack of alignment between the increasing use of chatbot assessment instruments and the evidential standards necessary to support educational decision-making. This criticism is not only scholarly: in the context of CMA-based grade, progression, or instructional adjustment determination, such as in the ever-growing AI-supported formative assessment systems, the lack of validity evidence makes a pedagogical invention of an epistemological threat.

The literature on AI item quality demonstrates a typical trend: chatbot-generated items are often helpful on the surface but at the same time, they are items that are technically measurement-violating. The observation that the discrimination index of AI-generated MCQ is significantly less than that of human-generated items (d=0.22 vs. 0.38) is especially worrying regarding summative use. Low discrimination implies that the test is not reliable in differentiating between high and low performers - negating the essence of standardized testing. This observation implies that future

ISSN: 2518-2951 https://doi.org/10.59568/AMJD-2026-14-2-35 African Multidisciplinary Journal of Development (AMJD)

chatbot evaluation development should be guided by validity-first frameworks, as opposed to deployment-first frameworks.

Importantly, validation frameworks employed in the studies reviewed were mostly those that were developed to assess technology acceptance (TAM, UTAUT) and not educational measurement validity frameworks (Messick unitary validity model, Kane argument based validity). Such methodological incompatibility is indicative of the technological nature of most CMA researchers, and it should be approached by increased interdisciplinary cooperation between AI developers, educational psychologists, measurement specialists.

Quality of Feedback: Immediacy, but not Depth

The trend of chatbot feedback quality that this review has uncovered, where the advantage of immediate universality is nearly matched with significant disadvantage in specificity, correctness and the tone of feeling, is what can be called a thin feedback paradox. Feedback is provided to students more quickly than any human system would provide it, but the feedback is often too generic, sometimes inaccurate and emotionally dissonant to serve the reflective, revision-oriented interaction that defines effective formative feedback (Boud and Molloy, 2013; Nicol and Macfarlane-Dick, 2006).

Hallucination problem - a misstated but which has a plausible-sounding feedback - is especially scouring in assessment situations, as it takes advantage of the same credibility that makes chatbots appear authoritative. Students that believe in the chatbot responses can make their mistakes even stronger instead of rectifying them. This danger is increased in subject areas that contain dense factual information (medicine, law, engineering) and in educational settings where students do not have the resources to critically analyze AI-generated feedback - a field where it is more common in under-resourced educational settings, such as in most African secondary school classrooms.

The principle of the Socratic chatbot design, to ask questions and not answer is a promising alleviator of the thin feedback paradox, which comes out of this review. Socratic chatbots can avoid the accuracy problem by encouraging the students to find their own explanations and self-assessments, and enhance cognitive activity. This principle of design coincides with both CLT and constructivist pedagogy and should be prioritized in future CMA design research.

Participation of the Learner: Promises and Pitfalls

The statistically significant positive engagement effect ($g = 0.57$) in the meta-analysis gives support to chatbot-mediated assessment as an engagement intervention, although in a diluted way. The effect size is not large - similar to what other educational technology interventions have shown - and the large heterogeneity ($I2 = 76.3$) indicates that context has a significant amount of results modulation. Its difference between formative ($g = 0.63$) and summative ($g = 0.44$) engagement effects echoes the current assessment theory: the autonomy and low-threat environment of high-stakes summative assessment are inherently restrictive to the deep engagement.

The SDT model shows a structural gap in the design of modern CMAs: they can meet competence needs (via immediate corrective feedback) and autonomy needs (via self-paced on-demand access) but fail to meet the relatedness needs, essentially. The impersonal character of chatbot interactions can present a culturally specific obstacle to ongoing interaction in cultures and educational traditions that place a premium on relational learning, such as in many African societies that are influenced by ubuntu values, which orient personhood towards relational connection. This is one of the critical

design problems that cannot be solved by technical ability: even the most NLP skill cannot replace the real presence of humans in cultures in which such presence is pedagogical.

The Research Gap in Africa

The observation that Africa is just 7.9 percent of the total world CMA literature, whilst harboring about 17 percent of the global student population and 60 percent of all young people in the world, is not just a statistical fact; but a structural injustice with immense ramifications concerning equity in education. The systematic under-representation of African settings in AI-in-education studies implies that the structures, criteria of validity and design principles that will guide the global deployment of chatbots are being constructed virtually entirely on high-income, high-connectivity, and mostly East Asian and North American settings. The danger is that once such tools are later implemented in African contexts, as they are becoming more, they will harbor implicit beliefs regarding infrastructure, language, cultural values, and criteria of assessment that will actively discriminate against African students.

This criticism applies to the character of African involvement in AI research: the fact that Ghana and Nigeria feature in the SSA AI-in-education literature more often than any other country, even though that represents the relative research capacity of the countries, implies that the East African, Central African, and Francophone West African context, which contain hundreds of millions of students, is virtually nonexistent in the evidence base. The adoption of early AI strategy by Mauritius and the Nairobi Statement by Kenya are also a policy momentum although policy without research investment is likely to implement tools that have never been tested in the environments in which they would be applied.

The way forward involves a three pronged approach namely: (1) special funding mechanisms to African chatbot assessment research, such as longitudinal studies in low-bandwidth settings; (2) authentic North-South research collaborations that make African researchers the major investigators, not the data collection sites; and (3) the creation of contextually relevant validity frameworks that consider linguistic diversity, infrastructure variability, and culturally situated assessment epistemologies

Research, Policy and Practice Implications

Research Implications

Validity-first frameworks should be the first priority of future CMA-related studies: psychometric validity and reliability analysis must be a methodological minimum of any study that employs chatbot assessment, rather than a luxury addition. There must be interdisciplinary groups of AI researchers, educational psychologists, and measurement experts. The longitudinal engagement studies (at least 12 weeks) are necessary to identify the effects of novelty and sustained engagement, especially in the African contexts where the novelty of technologies can confound the short-term results. To determine truthful standards, comparative studies regarding chatbot and human feedback quality in response to the same learner output, with the evaluation of the expert made blinded, are required.

Policy Implications

CMA implementation must be undertaken with measured skepticism by educational policymakers especially in Africa: the advantages of engagement and feedback immediacy are real, but the evidence base of validity is too scanty to support chatbot-based evaluation of high-stakes choice without strong human supervision. The policy framework as the 2024 continental AI strategy of the African Union and the 2024 Nairobi Statement gives the policy framework; now, however, it is required to develop national-level assessment quality standards explicitly covering AI-mediated

ISSN: 2518-2951 https://doi.org/10.59568/AMJD-2026-14-2-35 African Multidisciplinary Journal of Development (AMJD)

tools, based on the frameworks already in progress in the EU AI Act (2024) on high-risk AI applications in education.

Implications in practice to educators and designers

To apply CMA in educational practice, this review suggests: (1) relying on chatbots to provide low-stakes formative assessment and self-directed practice over high-stakes grading; (2) applying human review of chatbot-generated feedback before delivery to learners, especially in fact-intensive fields; (3) basing chatbot interactions on Socratic questioning principles to enhance cognitive engagement and prevent the emergence of hallucinations.

References

1) Ahmed, R., Al-Hity, M., & Broom, M. A. (2024). Comparing ChatGPT 3.5 and Google Bard in generating dental caries MCQs: Quality analysis by cognitive level. *Medical Education Online*, *29*(1), Article 2318341. https://doi.org/10.1080/10872981.2024.2318341
2) Annamalai, N., Ranaivo, B., & Jasin, J. (2023). Chatbot limitations in narrative and empathetic learning contexts. *Interactive Learning Environments*, *31*(5), 2910–2924.
3) Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102
4) Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31.
5) Boud, D., & Molloy, E. (2013). *Feedback in higher and professional education: Understanding it and doing it well*. Routledge.
6) Cakmak, O. (2022). Learner frustration in open-ended chatbot interview simulations. *Computer Assisted Language Learning*, *35*(9), 2112–2134.
7) Cheung, B. H., Lau, G. K., Wong, G. T., Lee, E. Y., Kulkarni, D., Seow, C. S., Wong, R., & Co, M. T. (2023). ChatGPT versus human in generating medical graduate exam MCQs: A multinational prospective study. *PLOS ONE*, *18*(9), e0290691. https://doi.org/10.1371/journal.pone.0290691
8) Chiu, T. K. F. (2022). Applying the self-determination theory (SDT) to explain student engagement in online learning during the COVID-19 pandemic. *Journal of Research on Technology in Education*, *54*(Suppl. 1), S14–S30.
9) Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, *119*(2), 166.e7–166.e16.
10) Deng, X., & Yu, Z. (2023). A meta-analysis and systematic review of the effect of chatbot technology use in sustainable education. *Sustainability*, *15*(4), 2940. https://doi.org/10.3390/su15042940
11) El Azhari, J., Zidoun, Y., & El Faddouli, N. E. (2023). Real-time chatbot feedback systems in higher education. *Education and Information Technologies*, *28*(6), 7234–7258.
12) Fan, J., Sun, T., Liu, J., Zhao, T., Zhang, B., Chen, Z., Glorioso, M., & Hack, N. (2023). How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. *Journal of Applied Psychology*, *108*(4), 504–522. https://doi.org/10.1037/apl0001082
13) Hong, Q. N., Pluye, P., Fabregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2018). Mixed methods appraisal tool (MMAT), version 2018. *Registration of Copyright #1148552*. Canadian Intellectual Property Office.
14) Hwang, G.-J., & Chang, C.-Y. (2021). A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments*, *31*(4), 1–14.
15) Jeon, J., Lee, S., & Choe, H. (2023). Beyond ChatGPT: A conceptual framework and systematic review of speech-recognition chatbots for language learning. *Computers & Education*, *206*, Article 104898. https://doi.org/10.1016/j.compedu.2023.104898
16) Kaczorowska-Spychalska, D. (2019). How chatbots influence marketing. *Management*, *23*(1), 251–270.

17) Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, *28*(2), 973–1018. https://doi.org/10.1007/s10639-022-11177-3
18) Lim, F. S., & Siong, L. F. (2025). A systematic approach to evaluate the use of chatbots in educational contexts: Learning gains, engagements and perceptions. *Computers*, *14*(7), 270. https://doi.org/10.3390/computers14070270
19) Lo, C. K., Hew, K. F., & Chen, G. (2024). ChatGPT and student engagement: Behavioural, emotional, and cognitive dimensions. *Computers & Education: Artificial Intelligence*, *6*, Article 100208.
20) Maphosa, M., Maphosa, V., Mpofu, B., & Ndlovu-Gatsheni, S. A. (2025). AI acceptance and usage in Sub-Sahara African education: A systematic review. *Journal of Advocacy, Research and Education*, *12*(1). https://doi.org/10.22270/jare.v12i1.1
21) Mead, A. D., & Zhou, A. (2023). Assessment item generation with AI: Utility and item quality challenges with GPT-3. *Educational Measurement: Issues and Practice*, *42*(4), 76–89.
22) Mungai, E., Kibirige, I., & Nduati, G. (2024). Cognitive load in chatbot-mediated learning: Theoretical and practical considerations. *Computers in Human Behaviour*, *152*, Article 108129.
23) Nemt-allah, M., Khalifa, W., Badawy, M., & Ahmed, M. F. (2024). Validating the ChatGPT Usage Scale: Psychometric properties and factor structures among postgraduate students. *BMC Psychology*, *12*, Article 497. https://doi.org/10.1186/s40359-024-01983-4
24) Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218.
25) Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, *2*, Article 100033.
26) Oladunmoye, E.O., Oyedele, O. Leah, Enamudu, G.P., and Faith, Nakalema, (2024a). Assessing Psychometric Tools in Online Education: Effectiveness and Obstacles in Virtual Learning Assessments. ISAR Journal of Arts, Humanities and Social Sciences, 2(12), 8-13.
27) Oladunmoye, E.O., Oyedele, O. Leah, Sa'ad, M.T., and Faith, Nakalema, (2024b). Ethical Considerations in High-Stakes Academic Assessments: Insights from the Nigerian Educational System. International Journal of Academic Multidisciplinary Research (IJAMR), 8(10), 101-108.
28) Oxford Insights. (2024). *Government AI Readiness Index 2024*. Oxford Insights. https://oxfordinsights.com/ai-readiness/ai-readiness-index/
29) Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLOS Medicine*, *18*(3), e1003583. https://doi.org/10.1371/journal.pmed.1003583
30) Rodrigo, M. M. T., Baker, R. S. J. d., & Rossi, L. (2012). Affect in Scooter: Cross-cultural comparison of student affect. *Proceedings of the 5th International Conference on Educational Data Mining*, 13–20.
31) UNESCO. (2024). *Artificial intelligence and education: Protecting human agency in a world of automation Eastern Africa*. United Nations Educational, Scientific and Cultural Organization. https://www.unesco.org/en/articles/artificial-intelligence-and-education-protecting-human-agency-world-automation-eastern-africa
32) Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
33) Wang, X., Chen, Y., Li, W., & Zhang, J. (2024). AI tutor appearance, communicative features, and student language proficiency in elementary classrooms. *Language Learning & Technology*, *28*(1), 1–22.
34) Winkler, R., & Söllner, M. (2018). Unleashing the potential of chatbots in education: A state-of-the-art analysis. In *Academy of Management Annual Meeting Proceedings* (Vol. 2018, No. 1). Academy of Management.