Journal of Emerging Technologies (JET)
Volume 6, Issue 1 (2026)
ISSN: 2710-0189 (Print) | 2710-0197 (Online)
Website: https://journals.jozacpublishers.com/jet/
https://dx.doi.org/10.4314/jet.v6i1.6

Cultural and ethical responsiveness of AI-driven assessment: Gender and geographical equity, adolescent mental health, and inclusive educational outcomes in Calabar, Nigeria

Ekaette Mfon Useh[*1], Christiana Onwubali[2]

[1*]*Department of Educational Measurement and Evaluation, University of Calabar, Nigeria, ekaetteuseh4@gmail.com*
[2]*Department of Psychology, University of Chester, United Kingdom, christianaonwubali25@gmail.com*

*Corresponding author: ekaetteuseh4@gmail.com

Abstract

The increasing integration of artificial intelligence (AI) into educational and mental-health assessment raises concerns about cultural relevance, psychometric fairness, accessibility, and ethical governance, particularly in resource-constrained contexts such as Nigeria. This paper examined the cultural, ethical, and psychometric implications of AI-driven assessment, focusing on differential item functioning (DIF), adolescent mental-health screening, community-centred assessment, and AI companionship. This integrative literature review synthesised published research on cultural and ethical assessment practices, gender- and location-based DIF in AI-generated and West African Examinations Council (WAEC) English Language items, AI-based mobile screening of adolescent mental-health conditions, community-centred assessment, and AI companionship among Nigerian university freshmen. The synthesis incorporated four directly relevant Calabar-focused studies, a Nigerian AI companionship study, international empirical literature, and child-centred AI guidance. The reviewed evidence indicated that culturally responsive and ethical assessment practices support inclusive assessment; both AI-generated and WAEC English Language items exhibited gender- and location-related DIF, although lower DIF magnitudes were reported for AI-generated items. Community-centred approaches strengthened ownership and accountability, while AI-based adolescent screening demonstrated potential for early identification alongside concerns about privacy, demographic bias, validation, and clinical follow-up. These findings informed an integrated framework connecting accessibility, cultural and linguistic validity, subgroup fairness, data governance, human oversight, and post-assessment support. AI-driven assessment can advance equitable educational and mental-health outcomes when culturally grounded, psychometrically validated, ethically governed, and supported by professional judgement and community participation. AI should complement rather than replace human expertise in assessment and intervention.

Keywords: Adolescent mental health, Artificial intelligence, Cultural responsiveness, Educational assessment, Gender equity, Inclusive education

1. Introduction

Assessment is not a culturally neutral activity. Decisions about what knowledge counts, how questions are phrased, how responses are scored, and what follows from a score are shaped by curriculum, language, institutional expectations, and the social context of learners. These concerns become more consequential when artificial intelligence (AI) is used to generate test items, score responses, classify learning needs, or screen for possible mental-health concerns. While AI can accelerate assessment and expand access to feedback, it can

Research Article: This article is published by ***Jozac Publishers*** in the ***Journal of Emerging Technologies (JET)***. This article is distributed under a Creative Commons Attribution International License (CC BY-SA 4.0). **Conflict of Interest:** The author/s declared no conflict of interest.

also reproduce or intensify inequalities when its data, assumptions, or deployment conditions do not represent the population being assessed (Ho, 2024; Idowu, 2024).

Calabar, in Cross River State, provides a particularly relevant context for examining the intersection of assessment, culture, ethics, gender, geography, and inclusion. Recent research conducted in the area has considered cultural and ethical dimensions of assessment in public secondary schools (Ita et al., 2025), gender- and location-based differential item functioning in AI-generated and WAEC English Language items (Uchegbue et al., 2025), and community-centred assessment as a pathway to inclusive social transformation (Otu et al., 2026). A related systematic review has examined AI-based screening of adolescent mental-health conditions through mobile health applications (Isaiah et al., 2026a). Together, these studies provide a useful evidence base for evaluating AI assessment as both a technical practice and a social process.

The relevance of adolescent mental health is substantial. The World Health Organization (WHO, 2025) estimates that approximately one in seven adolescents aged 10–19 experiences a mental-health condition. Screening technologies may help identify concerns earlier, particularly where specialist services are limited. However, an automated screening score is not a diagnosis. It must be interpreted alongside the adolescent's circumstances, appropriate validated measures, and qualified professional assessment. Privacy, informed participation, the risk of stigma, and the availability of referral pathways are therefore central to ethical implementation (UNICEF, 2025; UNESCO, 2023).

Recent work by Isaiah et al. (2026b) on AI companionship among Nigerian university freshmen further illustrates that the impact of AI is contextual rather than uniformly beneficial or harmful. In that qualitative study, students described AI as emotional scaffolding in some circumstances and as a possible substitute for human interaction in others; offline social support appeared to shape whether AI use was complementary or displacing. Although companionship is distinct from formal assessment, the study reinforces a broader principle: digital tools operate within existing social environments and should be evaluated in relation to human support, institutional practices, and unequal access to resources.

This paper therefore brings together the local assessment studies and related AI research to address four questions: (1) What cultural and ethical conditions are necessary for AI-driven assessment to be valid and fair in Calabar? (2) What do existing findings on gender- and location-based DIF reveal about the limits of assuming AI-generated items are automatically unbiased? (3) How should AI-based adolescent mental-health screening be governed? and (4) How can community-centred assessment support inclusive educational outcomes? The paper's contribution is an integrated, context-sensitive interpretation of existing evidence rather than a claim to have conducted new fieldwork.

2. Literature review

2.1. AI-driven Assessment and Educational measurement

AI-driven assessment encompasses several distinct applications, including test-item generation, adaptive task selection, automated scoring of written or spoken responses, personalised feedback, performance prediction, and the identification of learners who may require additional support. Although these applications can improve assessment efficiency, consistency, and responsiveness, each requires a separate validity argument demonstrating that the resulting scores and decisions are appropriate for their intended purpose. Ho (2024) cautions that artificial intelligence presents both opportunities and threats for educational measurement, particularly when automated systems introduce new sources of construct-irrelevant variance or make assessment decisions that are difficult to interpret. Similarly, Idowu (2024) demonstrates that algorithmic debiasing requires attention to training data, model design, evaluation criteria, and institutional decision-making rather than relying on technical adjustments alone.

Recent evidence also indicates that the use of AI to generate assessment items does not automatically guarantee fairness. In a study involving 400 Senior Secondary II students in Calabar, Uchegbue et al. (2025) compared AI-generated English Language test items with WAEC items and identified gender- and location-based differential item functioning (DIF) in both sets of items. Although the AI-generated items showed lower DIF magnitudes, the findings suggest that they were not entirely free from group-related differences. This distinction is important because an item may function differently across groups even when the overall test demonstrates acceptable reliability. Therefore, reliability coefficients alone cannot establish that assessment results are equitable across gender, geographical location, or cultural backgrounds.

Beyond item generation, fairness assessment should consider whether AI systems perform consistently across demographic groups and whether their outputs lead to equitable educational opportunities. Idowu's

(2024) systematic review identifies several approaches to evaluating and mitigating algorithmic bias, including sample reweighting, adversarial learning, and group-based performance comparisons. However, no single fairness metric is suitable for every educational context; the choice should reflect the assessment purpose, the characteristics of the learners, and the consequences of incorrect decisions. Ho (2024) similarly emphasises the need for ongoing monitoring of bias and changes in measurement performance over time.

In the Calabar context, these findings support a culturally responsive approach that considers gender, urban–rural location, language background, disability, and differences in access to digital resources. AI-generated questions should therefore be evaluated for curriculum alignment, clarity, cultural relevance, reliability, and differential item functioning before they are used to make consequential decisions. Human review remains essential, but it must be systematic and supported by empirical evidence rather than subjective judgement alone. Combining expert review with pilot testing, subgroup analysis, and periodic monitoring offers a stronger basis for developing AI-driven assessments that are valid, inclusive, and educationally equitable (Ho, 2024; Idowu, 2024; Uchegbue et al., 2025).

2.2. Cultural and ethical dimensions of assessment in Calabar

Ita et al. (2025) examined the challenges and opportunities associated with integrating cultural and ethical dimensions into assessment and evaluation in public secondary schools in Calabar. Their descriptive survey involved 400 educators and evaluators drawn from a reported population of 2,455. Using the Cultural and Ethical Dimensions in Assessment and Evaluation Questionnaire (CEDAEQ), the study found that cultural and ethical factors significantly influenced assessment implementation. The authors highlighted professional training, community participation, and effective policy implementation as important strategies for improving assessment practice. These findings suggest that cultural responsiveness and ethical conduct are integral to assessment quality, particularly in culturally diverse educational settings.

For AI-driven assessment, cultural responsiveness extends beyond translating questions into local languages. It requires evaluating whether test items, examples, idiomatic expressions, assumed background knowledge, response formats, and scoring criteria are appropriate for the learners being assessed. An item that appears neutral may disadvantage students if answering it depends on cultural experiences, linguistic conventions, or technological familiarity that are unrelated to the intended learning outcome. Consequently, assessment developers should pilot AI-generated items with diverse learners and examine whether particular groups experience systematic differences in item performance.

Ethical responsiveness further requires clear communication about how assessment scores are generated and used, secure handling of learners' information, procedures for correcting errors, and identifiable responsibility for consequential decisions. UNESCO (2023) recommends a human-centred approach to generative AI in education that protects privacy, promotes inclusion, and respects linguistic and cultural diversity. More recently, UNICEF (2025) emphasised children's rights to privacy, non-discrimination, transparency, accountability, and inclusion in AI systems affecting them. These principles are particularly relevant when AI-generated scores influence students' progression, placement, discipline, or access to educational and psychological support.

2.3. Gender and location-based differential item functioning

Differential item functioning (DIF) occurs when learners from different groups who have comparable levels of the underlying ability have different probabilities of answering an assessment item correctly. Unlike simple differences in average scores, DIF examines whether individual items operate differently across groups after accounting for the ability being measured. It is therefore an important indicator of potential assessment bias, although detecting DIF does not automatically establish that an item is unfair. Further investigation is necessary to determine whether the observed difference reflects construct-relevant factors or irrelevant cultural, linguistic, or contextual influences (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; Chen et al., 2025). Recent methodological research similarly emphasises that identifying DIF is only the first step: researchers must interpret its magnitude and investigate possible causes before deciding whether an item should be revised or removed (Li et al., 2025).

Uchegbue et al. (2025) provide context-specific evidence from Calabar. Their study compared AI-generated English Language multiple-choice items with 2025 WAEC items among 400 Senior Secondary School Two students in Calabar Municipality. The researchers reported gender- and location-based DIF in

both test types, although the magnitudes were generally lower for the AI-generated items. This finding suggests that AI-assisted item development may offer opportunities to reduce some group-related differences. However, it does not establish that AI-generated items are universally unbiased or consistently superior to conventional examination items. Instead, it reinforces the need to evaluate each set of items empirically with the intended student population.

Recent research further demonstrates why DIF analysis should account for the diversity within demographic groups. In a systematic review of 83 studies, Chen et al. (2025) identified methods including Rasch-based analysis, the Mantel–Haenszel procedure, item response theory (IRT), logistic regression, and SIBTEST as established approaches to investigating DIF in second-language assessment. The review also identified limitations in reporting practices and in the empirical investigation of the causes of DIF, highlighting the need for more rigorous analysis and interpretation.

Aryadoust et al. (2024) also found that ignoring heterogeneity within broad groups could lead to under-detection of DIF in listening-comprehension assessment. Their analysis identified 13 items with statistically significant DIF at the broad-group level and 15 at the subgroup level, although most flagged items did not raise practical concerns. This distinction illustrates why researchers should examine both statistical significance and the practical importance of item differences. Where sample sizes permit, assessment developers should also investigate intersections between gender and geographical location rather than assuming that all members of either group share the same assessment experiences.

Evidence concerning AI-generated assessment questions adds another important consideration. Bhandari et al. (2024) compared ChatGPT-generated algebra questions with textbook questions using responses from 207 participants and found comparable item-difficulty and discrimination parameters. Although this supports the potential psychometric usefulness of AI-generated questions, acceptable psychometric properties do not independently demonstrate cultural or demographic fairness. Item performance must still be examined across relevant groups.

2.4. AI-based screening of adolescent mental health

Isaiah et al. (2026a) reviewed 45 studies published between 2020 and 2024 on AI-based screening of adolescent mental-health conditions through mobile health applications. The review identified machine learning, deep learning, transformer models, chatbots, multimodal approaches, and passive sensing as emerging screening methods. Depression and anxiety were the most frequently investigated conditions, while privacy, demographic bias, limited longitudinal validation, interpretability, and clinical validation remained significant concerns.

Supporting these findings, Zhou et al. (2024) developed a machine-learning model for identifying depression risk among adolescents and reported an area under the receiver operating characteristic curve (AUC) of 0.892. Their findings demonstrated the potential of AI to identify adolescents at risk of depression, although the model's sensitivity indicated that some cases could be missed. Similarly, Beames et al. (2024) reviewed smartphone-based approaches for detecting depression and anxiety among young people aged 12–25 years. Their findings highlighted the potential of smartphone sensor data, including movement and device-use patterns, to complement conventional mental-health assessments. However, differences in age, technology access, and usage patterns may limit the generalisability of these approaches.

Furthermore, Ignashina et al. (2026) investigated large language models for identifying depressive symptoms in clinical notes concerning adolescents. Their findings suggested that language models could support automated symptom identification, although variations in model performance highlighted the need for further validation.

Collectively, these studies indicate that AI can support early identification of adolescent mental-health difficulties. However, model performance in other populations cannot automatically be generalised to adolescents in Calabar, Nigeria. Differences in language, culture, socioeconomic circumstances, and access to digital technology may influence screening accuracy. Additionally, passive sensing raises privacy concerns because location, sleep, and device-use data may reveal sensitive personal information. Therefore, AI-based screening should complement validated psychological assessments, incorporate human oversight, and provide appropriate safeguards for adolescents' confidentiality and informed consent.

2.5. Community-centred assessment and inclusion

Otu et al. (2026) examined community-centred assessment approaches and inclusive social transformation in Calabar Municipality. Using a survey and the Community-Centred Assessment and Social Transformation Questionnaire (CCASTQ), the study reported that community-based evaluation strategies were associated with relevance, ownership, accountability, and sustainability in development initiatives. These findings highlight the importance of involving community members in identifying needs, interpreting evidence, and evaluating intervention outcomes.

Supporting this perspective, Jagosh et al. (2012), in a realist review of participatory research, found that trusting relationships, collaborative partnerships, and community participation could contribute to positive research outcomes. Similarly, Domecq et al. (2014) reported that involving patients in research planning, design, and implementation could improve research relevance and the quality of research processes. Shippee et al. (2015) further emphasised that meaningful engagement helps researchers identify outcomes that matter to intended users and connect research findings with practical needs.

These findings have important implications for AI-supported assessment. Involving students, parents, teachers, counsellors, disability advocates, and community representatives in developing and evaluating AI tools can help identify culturally inappropriate questions, accessibility barriers, and potential ethical risks. Such participation is particularly important when assessing adolescent mental health, as differences in language, cultural beliefs, socioeconomic circumstances, and experiences of stigma may influence how young people understand questions and disclose psychological difficulties.

However, community participation should complement rather than replace psychometric testing, professional expertise, and independent validation. For adolescents in Calabar, a community-centred approach could improve the cultural relevance, accessibility, and acceptability of AI-based assessment while ensuring that screening results inform appropriate support. Thus, combining community knowledge with scientific validation offers a more inclusive approach to developing AI assessment systems that respond to local needs.

2.6. AI, emotional support, and the social context of young people

Isaiah et al. (2026b) investigated AI companionship among 30 Nigerian university freshmen who used AI relationally for more than 30 minutes daily and six student-affairs counsellors. The qualitative case study identified emotional scaffolding, displacement, authenticity, institutional inclusion, and the conditional nature of offline social capital as key themes. The authors interpreted AI companionship as potentially complementary when students had offline support and potentially substitutive when human support was limited. This study concerns relational AI rather than test generation or mental-health screening; its relevance lies in demonstrating how young people's experiences of AI are shaped by their social environments and access to human support.

Supporting this perspective, Herbener and Damholdt (2025) examined chatbot use and social connectedness among 1,599 Danish high-school students. They found that students who used chatbots for social support reported greater loneliness and lower perceived social support than other students. These findings suggest that some young people may turn to AI companionship when experiencing unmet social needs, although the cross-sectional evidence does not establish that chatbot use causes loneliness.

Similarly, Al-Zahrani (2025) surveyed 819 higher-education students and identified both perceived benefits of AI chatbots, including personalised assistance and access to information, and concerns about diminished human connection and emotional support. This finding highlights the possibility that AI may provide accessible interaction while raising questions about its relationship with existing human support networks.

Evidence from intervention research also demonstrates the potential benefits of AI-mediated emotional support. Feng et al. (2025) synthesised 31 randomised controlled trials involving 29,637 participants and found that AI chatbots produced small-to-moderate improvements in mental distress among adolescents and young adults. However, effectiveness varied across outcomes and intervention designs, while engagement difficulties and technical limitations remained concerns.

Further evidence comes from Yang et al. (2026), who conducted a randomised controlled trial involving 99 university students experiencing digital stress. Their findings showed that emotional-support and problem-solving chatbots had different benefits, with both reducing anxiety symptoms. However, neither

produced a significant improvement in depressive symptoms. This suggests that AI support should be matched to users' needs rather than assumed to provide uniform benefits across psychological outcomes.

Collectively, these studies suggest that the effects of AI on young people's emotional well-being depend partly on their social circumstances, the type of AI interaction, and the availability of human support. Applied to assessment, this means that technological accuracy alone is insufficient. Schools and mental-health services must also consider whether young people understand and trust assessment results, can obtain clarification, and have access to appropriate follow-up support. AI systems that identify psychological difficulties without providing accessible referral or support pathways may have limited practical value and could create additional risks. The present paper therefore connects evidence on AI companionship, emotional support, assessment fairness, and community inclusion to develop an integrated understanding of culturally valid, equitable, and adolescent-centred AI assessment. These studies address complementary dimensions of the issue; their samples and findings should not be combined as though they represent a single dataset or a pooled estimate of AI effectiveness.

2.7. An integrative theoretical framework

Sociotechnical systems perspective

The sociotechnical systems perspective views artificial intelligence (AI) assessment as the outcome of interactions among algorithms, datasets, assessment tasks, users, technological infrastructure, and institutions. From this perspective, assessment bias cannot be attributed solely to the AI model because the surrounding social and organisational environment also influences how results are generated, interpreted, and used. For example, poor internet connectivity may prevent some students from completing digital assessments, language differences may affect their understanding of questions, and institutional limitations may prevent students identified as needing support from accessing appropriate services.

This perspective is consistent with Sloane et al. (2022), who argued that responsible AI development requires attention to social context, power relations, and the participation of affected communities. Similarly, Selbst et al. (2019) demonstrated that fairness problems can arise when algorithmic systems are evaluated without considering the broader social systems in which they operate. Their analysis highlights the limitations of treating fairness as a purely technical property that can be resolved independently of institutional practices and social circumstances.

More recent evidence strengthens this argument. Dolata et al. (2022), in a systematic analysis of 310 articles, conceptualised algorithmic fairness as a sociotechnical phenomenon shaped by interactions between technological systems and social structures. Similarly, Starke et al. (2022) synthesised 58 empirical studies of perceived algorithmic fairness and highlighted the importance of human perceptions, decision-making procedures, and social context. Their review also identified a geographical limitation in the evidence, with studies concentrated largely in Western democratic settings, underscoring the need for context-specific research.

In education, Baker and Hawn (2022) reviewed algorithmic bias in educational systems and identified potential disparities associated with gender, race, nationality, socioeconomic status, and disability. Their findings indicate that unfairness can emerge at different stages of educational technology development and implementation, supporting the need to evaluate both technical performance and the circumstances in which assessment systems are used.

This perspective is directly relevant to the present study, which examines cultural validity, fairness, and inclusion in AI-supported assessment. The studies by Ita et al. (2025), Uchegbue et al. (2025), Isaiah et al. (2026a), and Otu et al. (2026) address complementary dimensions of this issue, including cultural appropriateness, differential item functioning, AI-based mental-health screening, and community participation. The sociotechnical perspective provides a framework for integrating these findings without treating them as interchangeable evidence. It directs attention to whether assessment items are culturally understandable, whether AI systems perform equitably across student groups, whether digital infrastructure permits inclusive participation, and whether schools can provide appropriate support following assessment. Consequently, it helps establish that fair and effective AI assessment depends not only on algorithmic accuracy but also on cultural relevance, institutional readiness, accessibility, and meaningful human oversight.

Validity and Fairness in educational measurement

Validity-centred measurement provides a second theoretical foundation for evaluating AI-supported educational assessment. It emphasises that an assessment is defensible when sufficient evidence supports the interpretation and use of its scores for the intended population and purpose. Reliability is important, but consistent results do not necessarily mean that an assessment measures the intended construct fairly across different student groups.

Kane (2013) explained that validation involves evaluating the assumptions and inferences underlying score interpretation and use. Similarly, the Standards for Educational and Psychological Testing emphasise that validity evidence should address the intended interpretation of scores, the assessment process, and the consequences of test use (American Educational Research Association et al., 2014).

More directly, Baker and Hawn (2022) reviewed algorithmic bias in educational systems and demonstrated the importance of examining disparities that may arise across demographic groups when educational technologies are developed and deployed. Their analysis supports evaluating AI assessment systems beyond overall predictive accuracy.

These perspectives establish the relevance of differential item functioning (DIF) analysis, cultural and linguistic review, accessibility testing, and examination of assessment consequences. DIF analysis can help identify items that behave differently for students from different groups who have comparable underlying ability. However, statistical evidence of DIF requires substantive interpretation to determine whether the differences reflect construct-irrelevant bias or legitimate differences in the knowledge or skills being assessed.

For the present study, this framework supports examining whether AI-supported assessments are culturally appropriate, linguistically comprehensible, and fair to students from different backgrounds. It also helps distinguish technical consistency from defensible score interpretation, ensuring that assessment fairness is evaluated alongside reliability and predictive performance.

Child-Rights and Human-Centred AI

The third foundation is a child-rights and human-centred approach, which emphasises that AI systems affecting children and adolescents should protect their safety, privacy, dignity, autonomy, and equal opportunities. UNICEF (2025), in its updated Guidance on AI and Children, identifies ten requirements for child-centred AI, including safety, data privacy, non-discrimination, transparency, accountability, well-being, and inclusion. Similarly, UNESCO (2023) recommends a human-centred approach to generative AI in education that protects learners' data, promotes human agency, and respects cultural and linguistic diversity.

These principles are particularly important when AI is used to assess educational performance or adolescent mental health. An assessment system may generate technically accurate predictions but still cause harm if it exposes sensitive information, produces unexplained classifications, reinforces existing inequalities, or denies students opportunities to challenge decisions. UNICEF's (2025) guidance therefore provides a relevant framework for considering children's rights throughout the development and implementation of AI systems, while UNESCO's (2023) guidance reinforces the need for age-appropriate use, human oversight, and institutional responsibility.

For the present study, this perspective supports examining whether AI-supported assessment protects students' confidentiality, accommodates cultural and linguistic differences, and ensures that assessment results are interpreted by qualified professionals. It also highlights the importance of involving adolescents in decisions affecting them and ensuring that students identified as needing assistance can access appropriate support. Accordingly, the child-rights perspective complements validity and fairness by asking not only whether an assessment produces defensible results, but also whether its use respects students' rights and promotes their well-being.

3. Methodology

3.1. Research design

This study adopts an integrative evidence-synthesis design. It critically compares published studies that address cultural and ethical assessment in Calabar, gender- and location-based DIF in AI-generated and WAEC English Language items, AI-based adolescent mental-health screening, community-centred assessment, and AI companionship among Nigerian freshmen. The design is appropriate because the study's objective is to connect findings from complementary research areas and derive an integrated framework,

rather than to collect new primary data or calculate a pooled statistical effect (Whittemore & Knafl, 2005; Torraco, 2016).

3.2. Evidence sources and selection

The evidence base includes the four named Calabar-focused or adolescent-screening studies supplied for this paper, the attached study on AI companionship among Nigerian university freshmen, and selected peer-reviewed literature and policy guidance on educational measurement, algorithmic fairness, adolescent mental health, and child-centred AI. The local studies were selected because their stated aims and findings directly address cultural responsiveness, gender and geographical equity, adolescent screening, or community inclusion. International literature and policy documents were used to contextualise the local evidence and identify ethical standards.

3.3. Analytical procedure

The studies were compared thematically according to their research purpose, population, design, principal findings, and implications for assessment practice. The synthesis focused on five domains: cultural and ethical responsiveness; gender and location-based fairness; adolescent mental-health screening; community participation; and the relationship between digital outputs and human support. Findings were attributed to their original studies. No new effect sizes were calculated, the original samples were not pooled, and no additional participant data were generated.

3.4. Scope and limitations of the evidence base

This is an integrative synthesis, not a registered systematic review. It is limited by differences among the included studies in population, method, outcome, and purpose. A survey of educators, a DIF analysis of test items, a systematic review of mHealth applications, a community assessment study, and a qualitative study of AI companionship cannot be treated as methodologically equivalent. Accordingly, the paper compares convergent themes and implications but does not claim causal conclusions or generalise individual study estimates beyond their stated settings.

4. Result and Discussion

4.1. Cultural and ethical responsiveness is foundational to assessment quality

The first theme is that cultural and ethical dimensions are central to credible assessment. Ita et al. (2025) reported that cultural and ethical factors significantly influenced assessment implementation in public secondary schools in Calabar. The study also highlighted professional training, community participation, and policy implementation as practical strategies. These findings align with the broader argument that an assessment's quality depends not only on the technical instrument but also on whether procedures are transparent, contextually meaningful, and responsibly implemented.

For AI-driven assessment, cultural responsiveness should be built into item development, scoring, interpretation, and review. A model may generate grammatically correct questions that nevertheless rely on unfamiliar examples or culturally specific assumptions. Ethical responsiveness requires clear explanations of how AI is used and safeguards against treating an automated score as unquestionable evidence. Thus, cultural and ethical review should occur before deployment and continue as tools are updated.

4.2. AI-generated items may reduce but do not eliminate DIF

Uchegbue et al. (2025) reported gender- and location-based DIF in both AI-generated and WAEC English Language test items among SSII students in Calabar Municipality, with lower DIF magnitudes reported for the AI-generated items. This finding is a practical contribution because it challenges two opposing assumptions: that established standardised items are necessarily free from group bias, and that AI-generated items are necessarily biased. Neither assumption is justified without evidence.

The result instead supports a validation-first approach. AI-generated items may offer flexibility to revise wording, replace culturally loaded examples, and generate alternative versions. However, model outputs must be screened by subject experts and tested empirically with learners. DIF should be examined alongside item difficulty, discrimination, reliability, content coverage, and the intended interpretation of scores. A finding of lower DIF in one test comparison should not be described as proof of universal fairness.

4.3. Gender and geography must be evaluated across the full assessment pathway

Gender and location equity extend beyond item-level DIF. Students may differ in device access, connectivity, opportunities to practise, privacy while completing a task, and access to assistance when a system fails. These conditions can influence who participates and whose responses are available for scoring. If missing responses are interpreted as lack of ability or motivation, infrastructural disadvantage can be misclassified as an individual deficit.

For Calabar, future audits should therefore examine access, completion, missingness, item functioning, score accuracy, feedback quality, and subsequent decisions by gender and location where sample sizes permit. The analysis should also consider disability and socioeconomic circumstances without assuming that these characteristics map neatly onto urban or rural categories. Reporting uncertainty and protecting small groups' privacy are essential.

4.4. Adolescent mental-health AI requires additional safeguards

Isaiah et al. (2026a) identified promising applications of AI in mobile adolescent mental-health screening, including transformer-based text models, multimodal approaches, and passive sensing. At the same time, the review identified concerns about privacy, demographic bias, limited longitudinal validation, and clinical implementation. These limitations matter because the consequences of a false positive or false negative may include stigma, unnecessary distress, missed care, or inappropriate decisions.

A screening output should therefore be treated as an indication for further review, not a diagnosis. Any local use should establish who can access the data, how long data are retained, how the adolescent is informed, how urgent risk is escalated, and what professional support is available. Passive data such as GPS, app use, and sleep patterns should not be collected simply because they are technically available. Collection must be necessary, proportionate, transparent, and supported by robust safeguards.

4.5. Community participation strengthens legitimacy and conclusion

Otu et al. (2026) reported that community-centred assessment strategies supported relevance, ownership, accountability, and sustainability in social transformation initiatives in Calabar Municipality. This finding complements the emphasis on community participation in Ita et al. (Ita et al., 2025). In AI assessment, participatory processes can help identify language barriers, culturally inappropriate items, inaccessible interfaces, and unintended consequences that technical testing may overlook.

Participation should be meaningful rather than symbolic. Learners and caregivers should be able to influence design decisions and understand how their feedback is used. Teachers and counsellors can identify operational constraints; disability advocates can evaluate accessibility; and community representatives can help assess whether the outputs are relevant to local priorities. Such involvement complements rather than replaces independent psychometric and ethical review.

4.6. Human support determines whether digital assessment produces inclusive outcomes

The attached AI-companionship study (Isaiah et al., 2026b) found that AI could function as a supplement to human relationships for some students and as a potential substitute where offline social support was limited. Although the study did not examine formal educational assessment, it reinforces the contextual nature of AI use. A technically accurate output may not produce an inclusive outcome unless people can understand it, challenge it, and access appropriate assistance. However, the ethical application of AI and its scientific integrity remain major concerns (Chukwuere, 2026), but it can contribute to some cognitive skills like critical thinking, analytical reasoning, and many others for users (Chukwuere & Chukwuere, 2025).

This distinction is especially important in adolescent mental-health screening and educational support. Detection without referral is incomplete; feedback without reasonable accommodation may not improve learning; and an automated label without a review mechanism can harden into an institutional judgement. Human oversight must therefore involve trained personnel with time, authority, and clear protocols to question AI outputs and arrange appropriate follow-up.

5. Conclusion

This paper examined cultural and ethical responsiveness in AI-driven assessment by integrating evidence on assessment practices in Calabar, gender- and location-based DIF, AI-based adolescent mental-health

screening, community-centred evaluation, and AI companionship among Nigerian freshmen. The evidence indicates that AI can create opportunities for more flexible assessment and earlier identification of needs, but it does not remove existing challenges of validity, group fairness, access, privacy, and human support.

The Calabar DIF study reported that AI-generated English Language items exhibited lower DIF magnitudes than the compared WAEC items, while still identifying DIF in both. The cultural and ethical assessment study highlighted the importance of training, community participation, and policy implementation. The adolescent mHealth review underscored both technical promise and substantial ethical and validation requirements, while the community-centred study emphasised local ownership and accountability. Read together, these findings support a model of AI assessment in which equitable access, contextual validity, subgroup auditing, child-centred data governance, and meaningful human oversight are mandatory.

The central recommendation is that AI should support rather than replace teachers, assessment specialists, counsellors, and community knowledge. High-stakes decisions should not rely solely on AI-generated scores or mental-health flags. Before deployment, systems should be tested with intended users, assessed for gender and location-based differences, reviewed for cultural and disability accessibility, and linked to transparent appeal and support pathways. This approach offers a defensible basis for improving inclusive educational outcomes in Calabar without overstating what the current evidence can establish.

6. Funding

No external financial support was received for the preparation of this paper.

References

1. Al-Zahrani, A. M. (2025). Exploring the impact of artificial intelligence chatbots on human connection and emotional support among higher education students. SAGE Open, 15(2). https://doi.org/10.1177/21582440251340615
2. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.
3. Aryadoust, V., Min, S., & Chen, X. (2024). Investigating differential item functioning across interaction variables in listening comprehension assessment. Studies in Educational Evaluation, 80, Article 101322. https://doi.org/10.1016/j.stueduc.2024.101322
4. Beames, J. R., Han, J., Shvetcov, A., Zheng, W. Y., Slade, A., Dabash, O., Rosenberg, J., O'Dea, B., Kasturi, S., Hoon, L., Whitton, A. E., Christensen, H., & Newby, J. M. (2024). Use of smartphone sensor data in detecting and predicting depression and anxiety in young people (12–25 years): A scoping review. *Heliyon, 10*(15), Article e35472. https://doi.org/10.1016/j.heliyon.2024.e35472
5. Bhandari, S., Liu, Y., Kwak, Y., & Pardos, Z. A. (2024). Evaluating the psychometric properties of ChatGPT-generated questions. Computers and Education: Artificial Intelligence, 7, Article 100284. https://doi.org/10.1016/j.caeai.2024.100284
6. Chen, X., Aryadoust, V., & Zhang, W. (2025). A systematic review of differential item functioning in second language assessment. Language Testing, 42(2), 193–222. https://doi.org/10.1177/02655322241290188
7. Chukwuere, J. E. (2026). The Ethical Concerns and Scientific Integrity Questions That Arise in the Use of ChatGPT in Higher Education. *International Journal of Applied Research in Business and Management*, 7(6). https://doi.org/10.51137/wrp.ijarbm.788
8. Chukwuere, J. E., & Chukwuere, G. C. J. (2025). The impact of generative artificial intelligence on higher education students in developing cognitive skills: A mini literature review. *Journal of Community Psychosocial Research*, 2(1).
9. Domecq, J. P., Prutsky, G., Elraiyah, T., Wang, Z., Nabhan, M., Shippee, N., Brito, J. P., Boehmer, K., Hasan, R., Firwana, B., Erwin, P., Eton, D., Sloan, J., Montori, V., Asi, N., Abu Dabrh, A. M., & Murad, M. H. (2014). Patient engagement in research: A systematic review. BMC Health Services Research, 14, Article 89. https://doi.org/10.1186/1472-6963-14-89

10. Feng, X., Hang, Y., Wu, W., Xiao, X., Dong, F., & Qiao, Z. (2025). Effectiveness of AI-driven conversational agents in improving mental health among young people: Systematic review and meta-analysis. *Journal of Medical Internet Research, 27*, e69639. https://doi.org/10.2196/69639

11. Herbener, A. B., & Damholdt, M. F. (2025). Are lonely youngsters turning to chatbots for companionship? The relationship between chatbot usage and social connectedness in Danish high-school students. *International Journal of Human-Computer Studies, 196*, Article 103409. https://doi.org/10.1016/j.ijhcs.2024.103409

12. Ho, A. D. (2024). Artificial intelligence and educational measurement: Opportunities and threats. *Journal of Educational and Behavioral Statistics, 49*(5). https://doi.org/10.3102/10769986241248771

13. Idowu, J. A. (2024). Debiasing education algorithms. *International Journal of Artificial Intelligence in Education, 34*, 1510–1540. https://doi.org/10.1007/s40593-023-00389-4

14. Ignashina, M., Bondaronek, P., Shams, P., Osuchukwu, J., Santel, D., Pestian, J., & Ive, J. (2026). Assisting screening of adolescent depression with large language models as symptom detectors: Algorithm development and validation. *JMIR Formative Research, 10*, Article e71614. https://doi.org/10.2196/71614

15. Isaiah, U. S., Naibo, A. T., Ugondo, M. T., Ogwuche, S., Iziengbuwa, S. J., Abiama, E. E., & Ineme, M. E. (2026). The role of AI companions in supporting adaptive emotion regulation among university students with social anxiety. *Discover Mental Health*, *6*(1), 53. https://doi.org/10.1007/s44192-026-00387-7

16. Isaiah, U. S., Useh, E. M., George, I., Franklin, C. C., & Adili-George, C. F. (2026). AI-Based Screening of Adolescent Mental Health Conditions Using Mobile Health Applications: A Systematic Review. *J Clin Med Health Care*, *3*(2), 1-15.https://doi.org/10.61440/JCMHC.2026.v3.57

17. Isaiah, U., Chukwuemelie, E. J., Effiong, S. S., Inwang, W. C., & Adili-George, C. (2026). When AI becomes a friend: Emotional support, social displacement, and university belonging among Nigerian freshmen. *Journal of Emerging Technologies*, *6*(1), 45-60. https://doi.org/10.4314/jet.v6i1.4

18. Ita, C. I., Anele, E., Otu, B. D., Otu, S. D. B., Useh, E. M., Ogba, U. F., Ekereke, A. B., Ekong, E. K., & Beshel, I. A. (2025). Challenges and opportunities in integrating cultural and ethical dimensions into assessment and evaluation in public secondary schools in Calabar, Nigeria. *African Journal of Theory and Practice of Educational Assessment, 14*(1), 39–50. https://doi.org/10.83151/r19t-fg54

19. Jagosh, J., Macaulay, A. C., Pluye, P., Salsberg, J., Bush, P. L., Henderson, J., Sirett, E., Wong, G., Cargo, M., Herbert, C. P., Seifer, S. D., Green, L. W., & Greenhalgh, T. (2012). Uncovering the benefits of participatory research: Implications of a realist review for health research and practice. The Milbank Quarterly, 90(2), 311–346. https://doi.org/10.1111/j.1468-0009.2012.00665.x

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

21. Li, Z., Shin, J., Kuang, H., & Huggins-Manley, A. C. (2025). Exploring the evidence to interpret differential item functioning via response process data. Educational and Psychological Measurement, 85(4), 783–813. https://doi.org/10.1177/00131644241298975

22. National Centre for Artificial Intelligence and Robotics. (2024). National artificial intelligence strategy. Retrieved from https://ncair.nitda.gov.ng/wp-content/uploads/2024/08/National-AI-Strategy_01082024-copy.pdf

23. Otu, B. D., Uchegbue, H. O., Anele, E., Otu, S. D. B., Ogba, U. F., Ekereke, A. B., Useh, E. M., Beshel, I. A., & Ekong, E. K. (2026). Community-centred assessment approaches and inclusive social transformation in Calabar Municipality Local Government Area of Cross River State, Nigeria. https://doi.org/10.83151/4ne4-hx73

24. Shippee, N. D., Domecq Garces, J. P., Prutsky Lopez, G. J., Wang, Z., Elraiyah, T. A., Nabhan, M., Brito, J. P., Boehmer, K., Hasan, R., Firwana, B., Erwin, P. J., Montori, V. M., & Murad, M. H. (2015). Patient and service user engagement in research: A systematic review and synthesized framework. Health Expectations, 18(5), 1151–1166. https://doi.org/10.1111/hex.12090

25. Torraco, R. J. (2005). Writing integrative literature reviews: Guidelines and examples. Human Resource Development Review, 4(3), 356–367. https://doi.org/10.1177/1534484305278283

26. Uchegbue, H. O., Otu, B. D., Ita, C. I., Useh, E. M., Otu, S. D. B., Ogba, U. F., Ekereke, A. B., Beshel, I. A., Ekong, E. K., & Anele, E. (2025). Comparative analysis of gender and location-based differential item functioning in AI-generated and WAEC English Language test items in Calabar, Nigeria. *African Journal of Theory and Practice of Educational Assessment, 14*(1). https://doi.org/10.83151/62kt-rf92

27. UNESCO. (2023). Guidance for generative AI in education and research. Retrieved from https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

28. UNICEF. (2025). Guidance on AI and children (Version 3.0). UNICEF Office of Research–Innocenti. Retrieved from https://www.unicef.org/innocenti/reports/policy-guidance-ai-children

29. United Nations Children's Fund. (2025). Guidance on AI and children: Version 3.0-Recommendations for AI policies and systems that uphold child rights. UNICEF. Retrieved from https://www.unicef.org/innocenti/reports/policy-guidance-ai-children

30. Whittemore, R., & Knafl, K. (2005). The integrative review: Updated methodology. Journal of Advanced Nursing, 52(5), 546–553. https://doi.org/10.1111/j.1365-2648.2005.03621.x

31. World Health Organization. (2025, September 1). Mental health of adolescents. Retrieved from https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health

32. Yang, Q., Wu, H., Du, X., Sun, Q., & Ren, Z. (2026). Emotional support or problem-solving? A randomized controlled trial of function-specific AI chatbots for digital stress in college students. *Applied Psychology: Health and Well-Being, 18*(4), Article e70208. https://doi.org/10.1111/aphw.70208

33. Zhou, Y., Zhang, X., Gong, J., Wang, T., Gong, L., Li, K., & Wang, Y. (2024). Identifying the risk of depression in a large sample of adolescents: An artificial neural network based on random forest. *Journal of Adolescence, 96*(7), 1485–1497. https://doi.org/10.1002/jad.12357