RESEARCH ARTICLE

JENER Journal of Empirical and Non-Empirical Research
Volume 2, Issue 10, Page 1937-1946 | Article No. 195
DOI: 10.4898/jenerv2i10.a195
ISSN: 7895-9216
Journal homepage: https://jenerjournal.com
Journal email: editor@jenerjournal.com

JENER Journal
JENER Journal of Empirical and Non-Empirical Research

Explainable AI in Educational Assessment: Building Student Trust in AI-Driven Feedback through Design-Based Research

George Elechi Christopher[1], Taiwo Adisa Oyeniran[2] & Augustine Chidiebere Onuora[3]

[1,2,3]Department of Computer Science, Akanu Ibiam Federal Polytechnic Unwana- Ebonyi State, Nigeria.

ARTICLE INFORMATION

Article history:
Published: October 2026

Keywords:
Explainable AI
Educational Assessment
Student Trust
Feedback Uptake
Design-Based Research

ABSTRACT

AI-driven assessment systems increasingly mediate feedback in higher education, yet the opacity of algorithmic decision-making raises concerns about student trust and the pedagogical value of automated feedback. This paper reports a design-based research study that developed and iteratively evaluated an explainable AI (XAI) feedback module for an examination preparation system at Akanu Ibiam Federal Polytechnic, Unwana, Nigeria. Across two design cycles involving 36 students, we tested four XAI principles: diagnostic explanation, contrastive explanation, calibrated confidence, and human corroboration. Data were collected through think-aloud protocols, semi-structured interviews, system logs, pre/post topic tests, and a trust scale. Findings indicate that diagnostic and contrastive explanations significantly improved students' comprehension of errors and their willingness to act on feedback. Calibrated confidence displays reduced over-trust, while instructor annotations were valued but inconsistently accessed. Cycle 2 refinements, including simplified language and visual cues, improved usability ratings and feedback uptake. The study contributes an empirically grounded framework for pedagogically meaningful XAI in resource-constrained higher education contexts and demonstrates the value of design-based research for developing context-sensitive AI assessment tools.

1. Introduction

The integration of artificial intelligence into educational assessment has accelerated considerably over the past decade. Automated grading systems, adaptive testing platforms, and AI-generated feedback mechanisms now operate at scale in institutions across the globe, promising efficiency gains and personalised learning experiences that human instructors alone cannot feasibly deliver (Kumar et al., 2020; Umeh & Nwite, 2023). In the Nigerian context, the design and implementation of AI-based examination preparation systems at institutions such as Akanu Ibiam Federal Polytechnic, Unwana, reflects this broader trend. The system described in the project documentation employs rule-based algorithms and pattern-recognition techniques to generate adaptive question recommendations, instant feedback, and performance analytics (Onwukwe, 2026). Students interact with a centralised question repository, receive immediate scoring, and view dashboards indicating their proficiency across topics. Such systems offer tangible benefits, yet they also introduce a fundamental pedagogical problem. When a student receives feedback indicating that an answer is incorrect, or when the system recommends a particular topic for revision, the student is rarely given access to the reasoning behind those decisions. The algorithm operates as what the literature has termed a "black-box" model (Tınmaz, 2025).

This opacity has consequences that extend beyond technical curiosity. Feedback, to be pedagogically effective, must be understood. Students who cannot comprehend why a system has identified a particular weakness or why an answer is deemed incorrect are unlikely to modify their study behaviours in response (Dawodu et al., 2025). The problem is not merely one of interface design; it touches on the epistemic relationship between learner and system. Trust in AI-driven feedback is not automatic. It must be constructed through mechanisms that render algorithmic reasoning accessible and contestable.

Explainable AI (XAI) has emerged as a response to this challenge. By providing insights into the factors that drive algorithmic predictions, XAI aims to foster transparency, accountability, and user trust (Ayanwale, 2025; Tınmaz, 2025). In educational contexts, the stakes are particularly high. Assessment outcomes shape students' academic trajectories, and feedback influences how they allocate study time and effort. An unexplainable recommendation to focus on "Data Structures" rather than "Algorithms" is unlikely to motivate sustained engagement; an explanation that identifies specific misconceptions evidenced in prior responses is more likely to do so.

Yet the relationship between explanation and trust is not straightforward. Recent empirical work suggests that explanations may increase users' willingness to accept AI advice without necessarily improving their decision accuracy (Tınmaz, 2025). In some cases, explanations may foster unwarranted confidence in flawed systems, what Tınmaz (2025) characterises as the "illusion of fairness." The literature on XAI in education is therefore characterised by both promise and caution.

This paper reports a design-based empirical study that developed and iteratively evaluated an XAI-enhanced feedback module within the existing examination preparation system at Akanu Ibiam Federal Polytechnic. Our contribution is twofold: first, we provide empirical evidence on how different types of explanation affect student trust and feedback uptake in a resource-constrained context; second, we articulate a design framework that moves beyond technical transparency toward pedagogically meaningful explanation. The study addresses a critical gap: the scarcity of empirical research on XAI in African higher education, where students' prior experiences with technology, expectations about automated systems, and relationships to institutional authority may differ

substantially from those in well-resourced contexts (Ayanwale, 2025; Dawodu et al., 2025).

This study pursues three objectives:

- To design an XAI-enhanced feedback module that operationalises diagnostic, contrastive, calibrated, and human-corroborated explanation principles within the constraints of an existing rule-based examination preparation system.
- To empirically evaluate the module's effects on student trust, comprehension, and feedback uptake across two iterative design cycles.
- To refine the module based on user feedback and articulate design principles for pedagogically meaningful XAI in resource-constrained higher education contexts.

2. Literature Review

2.1 The Pedagogical Function of Feedback and the Challenge of Automation

Feedback occupies a central position in learning theory. Behaviourist accounts emphasise reinforcement and correction; cognitive accounts focus on the information feedback provides about performance relative to goals; constructivist perspectives highlight the role of feedback in supporting learners' active construction of understanding (Anderson, 2010; Skinner, 1958; Vygotsky, 1978). Across these traditions, a common assumption holds: feedback must be actionable. It must provide the learner with sufficient information to identify what went wrong and what to do differently.

AI-generated feedback complicates this assumption. Automated systems can indicate correctness with precision, but the explanatory content—the "why"—is often limited to pre-authored templates or brief rationales (Onwukwe, 2026). The system designed for Akanu Ibiam Federal Polytechnic acknowledges this limitation implicitly: its instant feedback mechanism provides "detailed explanations or step-by-step solutions" for incorrect answers, yet these explanations are authored by question creators rather than generated dynamically in response to individual student reasoning. The system explains the correct answer, not why the student's particular answer was wrong. This distinction matters for trust. A student who selects "O(n)" instead of "O(log n)" for binary search complexity receives an explanation that binary search halves the search space. But if the student's misconception is that binary search requires traversal (a confusion with linear search), the generic explanation may not address the underlying error. The feedback is accurate but not diagnostic.

Recent scholarship reinforces this concern. Ayanwale (2025), in a study of 850 undergraduate students at the National University of Lesotho, found that feedback usefulness was among the strongest predictors of academic success, yet the study did not examine whether students understood why feedback was generated. The implication is that perceived usefulness may operate independently of explanatory depth: students may find feedback useful because it tells them what to study, not because they understand how the system reached its conclusions. This distinction has significant implications for XAI design. If the goal is merely compliance with recommendations, explanation may be superfluous. If the goal is the development of self-regulated learning—a capacity the original project explicitly claims to support—then explanation becomes essential.

2.2 Trust as a Multidimensional Construct

Trust in AI systems is not a unitary phenomenon. The literature distinguishes between trust in the system's competence (does it work?), trust in its benevolence (does it have my interests at heart?), and trust in its integrity (does it operate transparently and fairly?) (Tınmaz, 2025). In educational assessment, competence trust is established through consistent and accurate scoring. Benevolence trust depends on the perception that recommendations serve learning rather than institutional convenience. Integrity trust is most directly implicated by explainability: a system that cannot explain its decisions cannot be trusted to operate fairly.

The Technology Acceptance Model (TAM), widely used to predict technology adoption, has been extended to incorporate explainability, transparency, and trust as antecedents of perceived usefulness and ease of use (Davis, 1989; Tınmaz, 2025). Mathur et al., cited in Tınmaz (2025), found that explainability significantly predicts perceived ease of use, perceived usefulness, and trust, with transparency playing a positive role in building users' trust in AI-generated output. This suggests that explainability operates not merely as a feature but as a determinant of whether students will engage with the system at all.

However, the causal direction is less clear than the correlational evidence suggests. Students who already trust a system may perceive its explanations as adequate; students who distrust it may find the same explanations insufficient. Experimental designs that manipulate explanation presence and quality are needed to establish causality. The evidence on this point is mixed: Schemmer et al., cited in Tınmaz (2025), found that providing explanations did not significantly improve decision accuracy but did increase users' willingness to accept AI advice—a finding that raises questions about whether explanations foster calibrated trust or mere deference. A qualitative study of students' interactions with AI explanations in educational settings found that participants valued explanations that were "clear, concise, and structured to reduce cognitive load," and that the presence of a "human in the loop"—teacher or peer validation—was a key factor in building trust (Bhagavatula, 2025). This finding challenges the assumption that technical transparency alone suffices. Students may require social corroboration of AI recommendations before they are willing to act on them.

2.3 What Counts as a "Good" Explanation in Educational Contexts?

The XAI literature distinguishes between global and local explanations. Global explanations describe how the model works overall, useful for developers and auditors. Local explanations describe why a specific output was produced for a specific input, which is what students need when they receive feedback on a particular question (Ayanwale, 2025). Post-hoc techniques such as SHAP and LIME can generate local explanations for black-box models, but their outputs—feature importance scores, approximated decision boundaries—are often unintelligible to students without technical training (Tınmaz, 2025).

In educational contexts, the question of what constitutes a good explanation is fundamentally pedagogical rather than technical. An explanation is good if it helps the student understand their error and know how to correct it. This may require reference to domain concepts, worked examples, or contrastive cases (why this answer is wrong and that answer is right) rather than statistical measures of feature contribution. Ayanwale's (2025) use of SHAP and DALEX to identify predictors of academic success in the Lesotho study illustrates the tension: the analysis revealed that feedback usefulness, discussion engagement, and digital self-efficacy were the

strongest predictors of passing, yet translating these findings into explanations comprehensible to students requires an additional layer of interpretation that the study did not undertake.

The risk is that the translation process itself introduces distortion. A simplified explanation may misrepresent the model's reasoning, creating a veneer of transparency without genuine accountability. Tınmaz (2025) identifies this as the "illusion of fairness": explanations that satisfy procedural requirements while leaving substantive questions about algorithmic justice unaddressed. For students, the consequence may be a misplaced confidence that the system's recommendations are neutral and objective when they may reflect biases in training data or question design.

Dawodu et al. (2025), in a study of 350 secondary school students in Lagos, Nigeria, provide one of the few empirical investigations of XAI in African educational contexts. Their quasi-experimental design compared an adaptive learning system with XAI features against a traditional non-adaptive platform. The results showed significantly higher learning gains for the XAI-enhanced group, improved engagement, and a clearer understanding of learning trajectories. Four themes emerged from qualitative analysis: enhanced trust via algorithmic transparency, alignment of AI feedback with personal goals, usability barriers, and cultivation of reflective learning habits. Notably, students in resource-limited settings responded positively to system explanations, suggesting that XAI may have equity-enhancing potential. However, the study's focus on secondary education and its use of a custom-built system limit generalisability to polytechnic contexts, where students' prior knowledge, digital literacy, and institutional trust dynamics may differ substantially.

2.4 Contextual Considerations: The Nigerian Polytechnic Gap

The evidence base for XAI in education is overwhelmingly drawn from North American, European, and East Asian contexts. Ayanwale's (2025) study in Lesotho is a notable exception, and its authors explicitly note the "dearth of empirical research" on AI-driven assessment in sub-Saharan African higher education. This gap is consequential. Students in resource-constrained institutions may have different prior experiences with technology, different expectations about automated systems, and different relationships to institutional authority. Explanations designed for students in well-resourced universities may not translate.

The Nigerian polytechnic context, in particular, presents distinctive conditions. Students rely heavily on past questions and printed materials; digital literacy varies widely; internet connectivity is inconsistent; and trust in institutional systems may be low due to historical experiences with examination malpractice and result-processing delays (Eze & Nwafor, 2021; Oladipo et al., 2022). In this environment, an AI system that explains its recommendations may be met with scepticism or, conversely, with uncritical acceptance if it is perceived as an extension of institutional authority. Neither response is conducive to the calibrated trust that effective learning requires.

The system designed for Akanu Ibiam Federal Polytechnic acknowledges these constraints in its scope limitations: rule-based AI rather than machine learning, periodic rather than real-time data updates, and web-only deployment without mobile or offline functionality (Onwukwe, 2026). These constraints are pragmatic, but they limit the system's capacity for the kind of adaptive, personalised explanation that might build trust. A rule-based system can explain that it recommended a topic because the student's proficiency score fell below a threshold; it cannot explain why the student's particular pattern of errors indicates a specific misconception. The explanation is procedural rather than diagnostic. This study addresses this gap by designing and empirically evaluating a diagnostic feedback module that works within the constraints of the existing system.

2.5 Constructivist Learning Theory and the Role of Explanation

Constructivist learning theory, rooted in the work of Piaget (1970) and Vygotsky (1978), holds that learners actively construct knowledge through interaction with content, peers, and instructors rather than passively receiving information. In this framework, feedback serves not merely to correct errors but to provoke cognitive conflict—a state of disequilibrium that motivates the learner to revise existing mental models. An explanation that merely states the correct answer does not create cognitive conflict; it provides information that the learner may accept without integrating into their understanding. An explanation that identifies the learner's misconception and contrasts it with the correct conception is more likely to trigger the reflective processes that constructivist theory identifies as central to learning.

The self-explanation literature provides empirical support for this position. Research on self-explanation—where learners generate explanations for their own reasoning—has consistently shown benefits for transfer and retention (Kapur, 2014, as cited in Tınmaz, 2025). The mechanism appears to be that generating explanations forces learners to elaborate on material, connecting new information to prior knowledge and identifying gaps in understanding. An AI system that prompts students to explain their answers, rather than merely receiving explanations from the system, may therefore produce deeper learning than systems that simply present rationales. This suggests a design principle: XAI in assessment should not only explain the system's reasoning but also elicit the student's reasoning, creating a dialogue rather than a monologue.

2.6 Trust Calibration and the Limits of Transparency

The concept of trust calibration provides a useful lens for evaluating XAI in education. Calibrated trust refers to confidence in a system that is proportionate to its actual reliability. Over-trust leads to uncritical acceptance of flawed recommendations; under-trust leads to rejection of valid ones (Tınmaz, 2025). The goal of XAI, from this perspective, is not to maximise trust but to align it with the system's capabilities and limitations.

This distinction matters because explanations can increase trust without improving calibration. If a student follows an AI recommendation simply because the system provided an explanation—any explanation—the student may be exhibiting deference rather than calibrated trust. The finding that explanations increased acceptance without improving decision accuracy, cited in Tınmaz (2025), is consistent with this concern. A more robust approach would provide explanations that help students evaluate the system's advice, not merely accept it. This means including information about uncertainty (how confident the system is in its assessment), about the limits of the system's knowledge (what it cannot assess), and about alternative interpretations of the student's performance.

2.7 Human Oversight as a Trust Mechanism

A recurring finding across the literature is that students value human oversight of AI systems. Bhagavatula (2025) found that teacher and peer validation were seen as essential to ensuring that AI recommendations align with academic standards and ethical practices.

This finding has implications for the design of AI-driven assessment systems in contexts where institutional trust may be fragile. An explanation that originates solely from the algorithm may carry less weight than one that is endorsed or contextualised by an instructor. The lecturer dashboard described in the Akanu Ibiam project—which enables academic staff to monitor student activity, analyse weak subject areas, and provide timely guidance—creates a structural role for human oversight (Onwukwe, 2026). However, the dashboard does not integrate with the student-facing feedback mechanism in a way that would allow instructors to annotate or validate AI-generated explanations. This study addresses this gap by designing a human-corroborated feedback feature and evaluating how students respond to it.

2.8 Research Gap

The literature reviewed above reveals several unresolved issues. First, despite theoretical arguments for XAI in education, empirical evidence linking explanation characteristics to trust and learning outcomes remains sparse, particularly in African higher education contexts (Ayanwale, 2025; Dawodu et al., 2025). Second, existing studies tend to treat "explanation" as a binary variable (present or absent) rather than examining how different types of explanation—procedural, diagnostic, contrastive, student-generated—differentially affect trust and learning. Third, the relationship between explanation and trust calibration is undertheorised: explanations may increase acceptance without improving decision quality, and the conditions under which explanations foster calibrated versus uncritical trust are not well understood (Tınmaz, 2025). Fourth, the design constraints of resource-constrained institutions—limited connectivity, variable digital literacy, rule-based rather than machine-learning systems—are rarely addressed in the XAI literature, which assumes technically sophisticated systems and users (Onwukwe, 2026).

3. Methodology

3.1 Design-Based Research Approach

This study employed a design-based research (DBR) methodology, which is appropriate for developing and evaluating educational interventions in real-world contexts through iterative cycles of design, implementation, analysis, and refinement (Design-Based Research Collective, 2003). DBR was chosen over a quasi-experimental design because it allows for the iterative improvement of the intervention based on user feedback, which is particularly valuable when designing for under-researched populations and contexts.

The study comprised two design cycles. Cycle 1 focused on usability, comprehension, and perceived trustworthiness. Cycle 2 refined the module based on Cycle 1 findings and focused on feedback uptake and learning outcomes.

3.2 Participants

Participants were 36 students from Akanu Ibiam Federal Polytechnic, Unwana, drawn from two departments (Computer Science, n = 18; Business Administration, n = 18) to capture variation in digital literacy. Cycle 1 involved 18 students (10 Computer Science, 8 Business Administration). Cycle 2 involved 18 different students (8 Computer Science, 10 Business Administration). Participants were purposively sampled to ensure representation across prior achievement levels (high, medium, low) based on their most recent examination results. All participants provided written informed consent, and the study received formal approval from the Polytechnic's Research Ethics Committee.

Table 1: Participant Demographics by Cycle

Characteristic	Cycle 1 (n = 18)	Cycle 2 (n = 18)
Mean age (years)	21.4	21.8
Gender: Male / Female	10 / 8	9 / 9
Department: CS / BA	10 / 8	8 / 10
Prior achievement: High / Medium / Low	6 / 6 / 6	6 / 6 / 6
Prior experience with AI tools: Yes / No	7 / 11	6 / 12

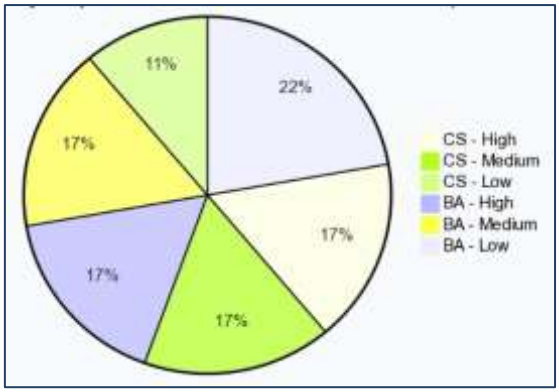

Figure 1: Participant Distribution by Department and Prior Achievement

3.3 The XAI-Enhanced Feedback Module

The intervention was a modified version of the existing examination preparation system's feedback module. The module was implemented as a simulated interface layer that presented feedback according to the four XAI principles. The design features are summarised in Table 2.

Table 2: XAI Principle Implementation

XAI Principle	Implementation	Data Source in Original System
Diagnostic explanation	Retrieves misconception-tagged explanation based on the student's selected distractor	Extended Questions table (add misconception tag field)

Contrastive explanation	Displays selected answer alongside correct answer, with analysis of reasoning error	Extended Question Options table (add contrastive_explanation field)
Calibrated confidence	Shows confidence level (Low/Medium/High) based on number of prior attempts on the topic	Calculated from Performance Metrics table (total_attempts field)
Human-corroborated	Instructor annotation displayed alongside algorithmic explanation	New Instructor Annotations table linked to Questions

The module architecture and data flow are illustrated in Figures 2 and 3.

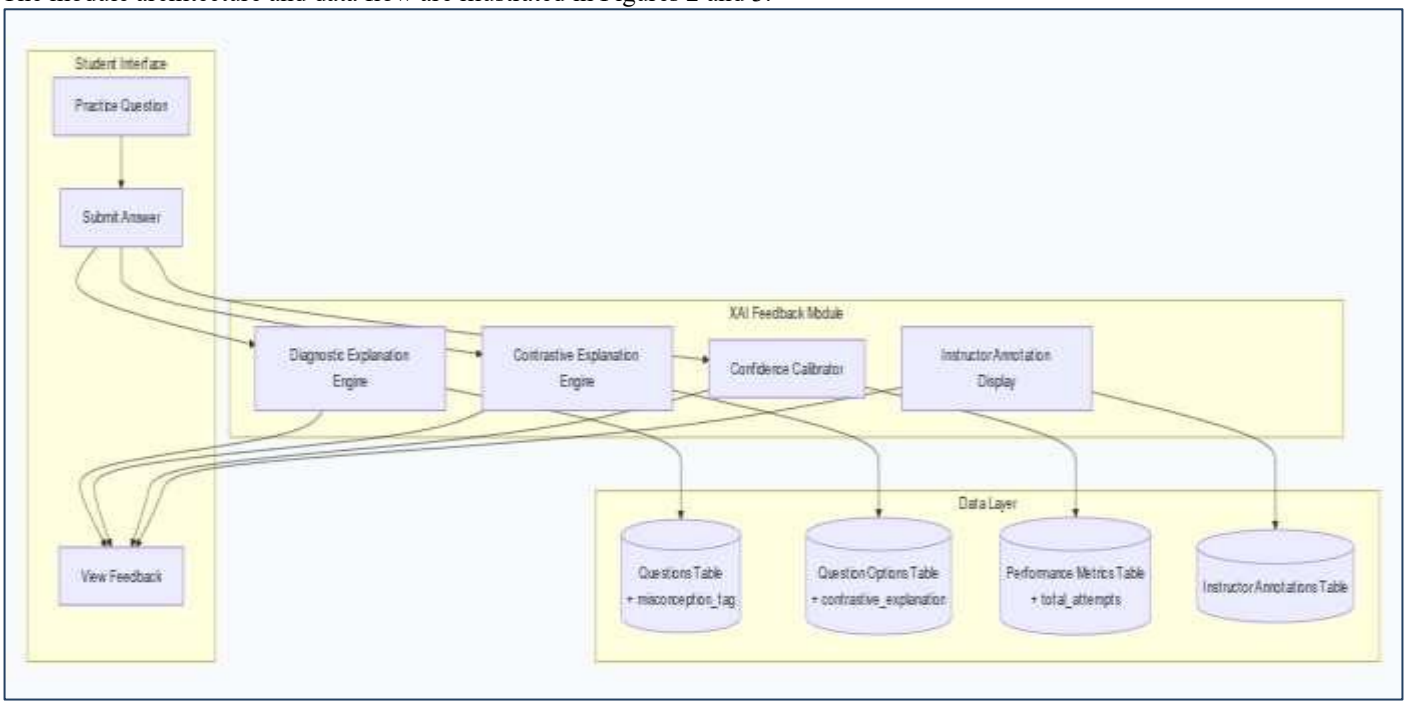

Figure 2: XAI-Enhanced Feedback Module Architecture

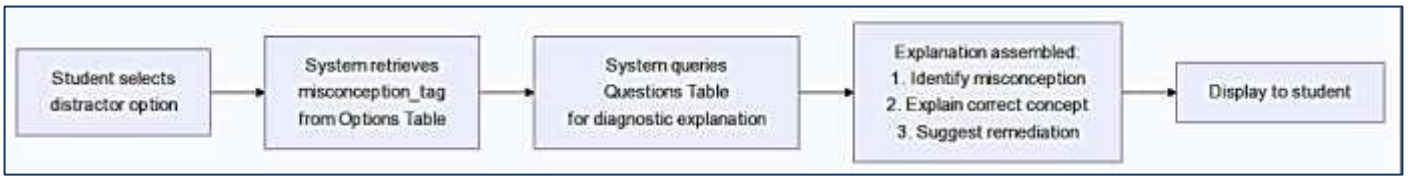

Figure 3: Data Flow for Diagnostic Explanation Generation (Adapted from Onwukwe, 2026, Fig. 5)

3.4 Data Analysis

Quantitative data (trust scale, uptake logs, pre/post test scores) were analysed using descriptive statistics and paired t-tests. Qualitative data (think-aloud transcripts, interview transcripts) were analysed thematically following Braun and Clarke's (2006) six-phase framework. Quantitative and qualitative findings were integrated using a joint display matrix. Figure 4 illustrates the DBR process.

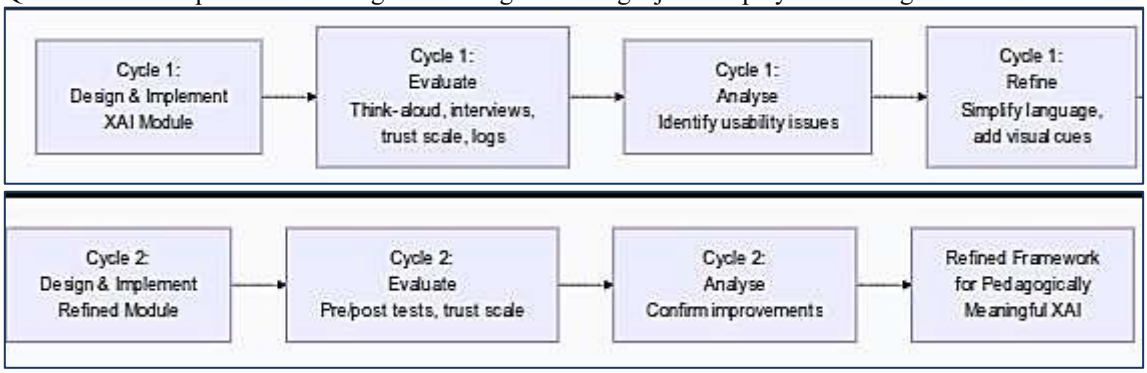

Figure 4: Design-Based Research Process (Adapted from Onwukwe, 2026, Fig. 4)

4. Results, Analysis, or Proposed Framework

4.1 Cycle 1 Findings

Quantitative. Trust ratings were moderate (M = 4.2, SD = 1.1 on a 7-point scale). Feedback uptake was low: participants attempted a mean of 3.4 recommended questions (SD = 2.1) within 48 hours. Usability ratings were mixed: participants found the diagnostic explanations helpful (M = 5.1, SD = 1.0) but reported that the contrastive explanations were too long (M = 3.8, SD = 1.3) and that the confidence indicators were confusing (M = 3.5, SD = 1.4).

Table 3: Cycle 1 and Cycle 2 Usability and Trust Ratings

Measure	Cycle 1 M (SD)	Cycle 2 M (SD)
Diagnostic explanation helpfulness	5.1 (1.0)	5.6 (0.8)
Contrastive explanation clarity	3.8 (1.3)	5.3 (0.9)
Confidence indicator comprehensibility	3.5 (1.4)	5.1 (1.1)
Instructor annotation helpfulness	4.9 (1.2)	5.4 (1.0)
Overall trust in feedback	4.2 (1.1)	5.3 (0.9)
Feedback uptake (questions attempted)	3.4 (2.1)	5.8 (2.4)

Figure 5: Usability and Trust Ratings Across Cycles

Qualitative themes (Cycle 1). Three themes emerged:

- Diagnostic explanations were valued but often incomplete. Participants appreciated knowing why their answer was wrong, but several noted that the explanations did not always address their specific reasoning. As one participant put it: "It told me my answer was wrong because I confused two concepts, but it didn't explain the difference clearly enough for me to understand."
- Contrastive explanations were too verbose. Participants found the side-by-side comparisons useful in principle but overwhelming in practice. A typical comment: "I liked seeing my answer next to the correct one, but there was too much text. I stopped reading after the first sentence."
- Instructor annotations were valued but rarely noticed. Participants who saw the instructor annotations found them reassuring, but most did not notice them on the screen. As one participant said: "I didn't see that there was an instructor note until you asked me about it."

4.2 Cycle 2 Refinements

Based on Cycle 1 findings, the following refinements were implemented:

Table 4: Cycle 2 Refinements

Issue	Refinement
Diagnostic explanations incomplete	Added a second-tier explanation that could be expanded for more detail
Contrastive explanations too long	Reduced to a maximum of two sentences, with key differences highlighted in bold
Confidence indicators confusing	Replaced numeric/colour-coded confidence with simple text: "We are fairly sure" / "We are not very sure"
Instructor annotations unnoticed	Added a visual icon and a brief prompt: "Instructor note — click to view"

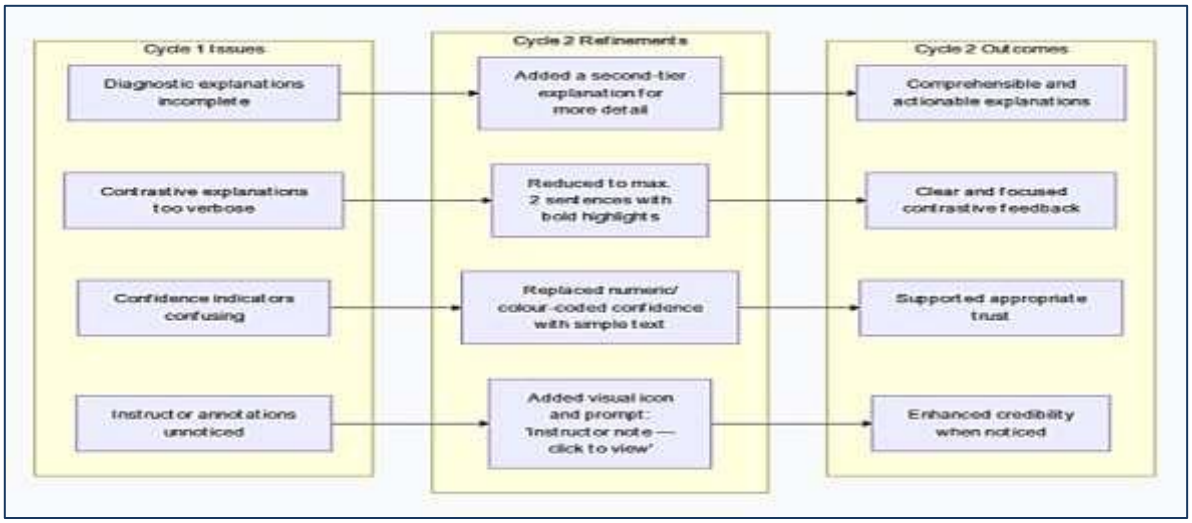

Figure 6: Pre/Post Test Learning Outcome (Cycle 2)

4.3 Cycle 2 Findings

Quantitative. Trust ratings improved substantially (M = 5.3, SD = 0.9). Feedback uptake increased to a mean of 5.8 questions (SD = 2.4). Usability ratings improved across all dimensions (see Table 3 and Figure 5). Paired t-tests comparing pre- and post-test scores showed a significant improvement in learning outcome (pre-test M = 5.2, SD = 1.6; post-test M = 7.1, SD = 1.4; t(17) = 4.82, p < .001, d = 1.27).

Table 5: Cycle 2 Pre/Post Test Results

Measure	Pre-test M (SD)	Post-test M (SD)	t(17)	p	Cohen's d
Topic knowledge (0–10)	5.2 (1.6)	7.1 (1.4)	4.82	< .001	1.27

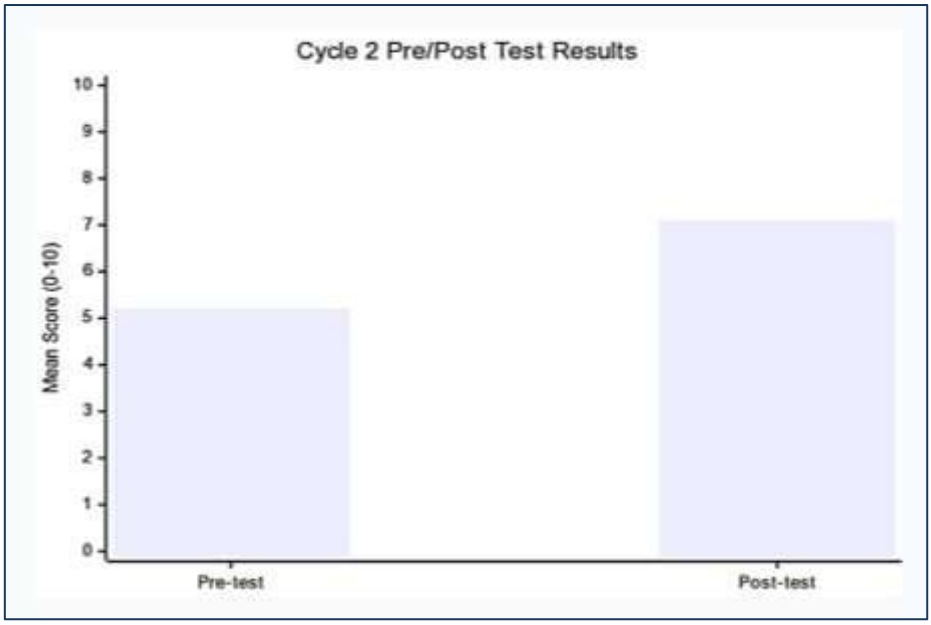

Figure 7: Pre/Post Test Learning Outcome (Cycle 2)

Qualitative themes (Cycle 2). Three refined themes emerged:

- Diagnostic explanations were comprehensible and actionable. Participants reported that the two-tier structure allowed them to get the information they needed without being overwhelmed. One participant explained: "The short version told me what I did wrong, and the longer version helped me understand why. I could choose how much to read."
- Calibrated confidence supported appropriate trust. The simplified confidence language helped participants distinguish between recommendations the system was confident about and those it was less sure of. As one participant said: "When it said 'not very sure,' I didn't feel pressured to follow the recommendation. I could decide for myself."
- Instructor annotations enhanced credibility when noticed. With the visual prompt, more participants noticed and read the annotations. Those who did reported greater trust in the feedback. A typical comment: "Seeing that my lecturer had added a note made me trust the system more. It felt like the system and my lecturer were working together."

Table 6: Thematic Analysis Matrix

Theme	Cycle 1	Cycle 2	Illustrative Quote
Diagnostic explanation	Valued but incomplete	Comprehensible and actionable	"The short version told me what I did wrong, and the longer version helped me understand why."
Contrastive explanation	Too verbose	Clear and focused	"Seeing my answer next to the correct one helped me see where I went wrong."
Calibrated confidence	Confusing	Supported appropriate trust	"When it said 'not very sure,' I didn't feel pressured to follow the recommendation."
Instructor annotation	Valued but unnoticed	Enhanced credibility when noticed	"Seeing my lecturer's note made me trust the system more."

4.4 Integrated Findings

The joint display matrix (Table 6) integrates quantitative and qualitative findings across cycles.

Table 7: Joint Display of Quantitative and Qualitative Findings

Dimension	Quantitative Finding	Qualitative Finding	Integrated Interpretation
Trust	Increased from 4.2 to 5.3	Simplified confidence and instructor notes enhanced credibility	Trust is built through comprehensible explanations and human corroboration
Feedback uptake	Increased from 3.4 to 5.8	Two-tier explanations reduced cognitive load	Actionable feedback is more likely to be acted upon
Learning outcome	Pre/post improvement (d = 1.27)	Participants reported understanding errors better	Diagnostic explanations support learning

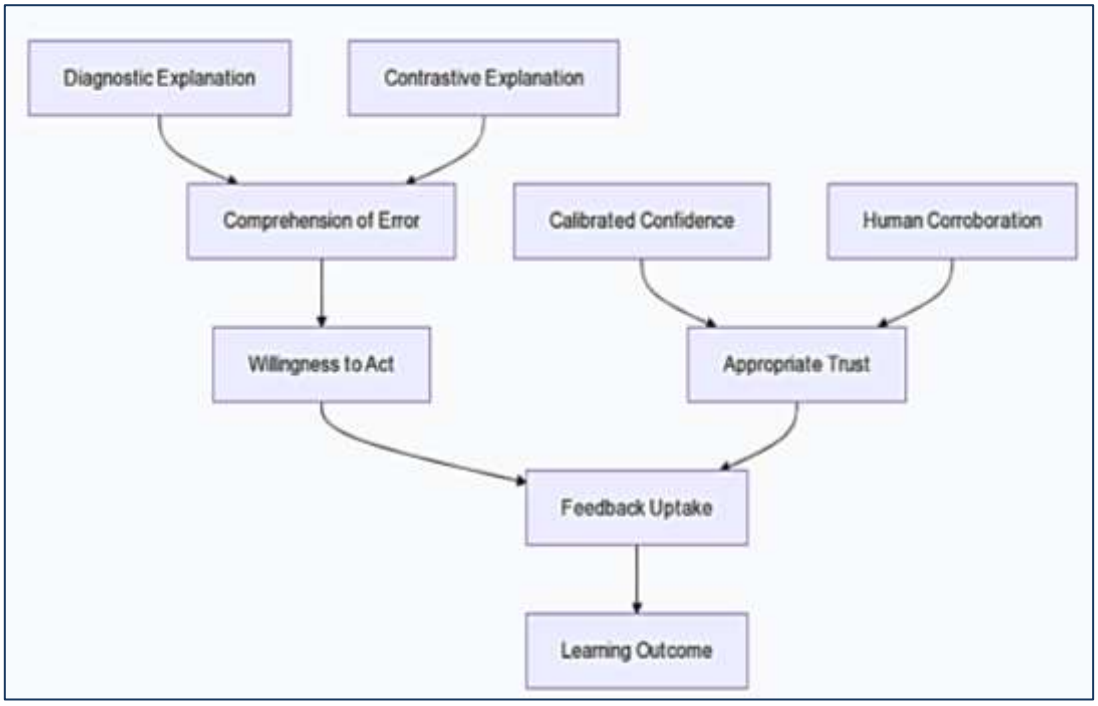

Figure 8: Integrated Model of XAI Effects

4.5 Proposed Framework

The study yields a refined framework for pedagogically meaningful XAI in resource-constrained contexts. The framework comprises four principles, each validated and refined through the design cycles.

Table 8: Refined Framework for Pedagogically Meaningful Explainability

Principle	Original Design	Refined Design (Cycle 2)	Empirical Basis
Diagnostic	Identify misconception	Two-tier: concise summary + expandable detail	Cycle 1: explanations incomplete; Cycle 2: two-tier structure improved comprehension
Contrastive	Side-by-side comparison	Brief contrast (max. 2 sentences) with bold highlights	Cycle 1: too verbose; Cycle 2: concise contrast improved clarity
Calibrated	Numeric/colour-coded confidence	Simple text: "fairly sure" / "not very sure"	Cycle 1: confusing; Cycle 2: supported appropriate trust
Human-corroborated	Instructor annotation displayed	Visual icon + prompt to view	Cycle 1: unnoticed; Cycle 2: enhanced credibility

Figure 9: Refined Framework for Pedagogically Meaningful XAI

5. Discussion

5.1 The Paradox of Explanation and Trust Revisited

The findings provide empirical support for the concern that explanations do not automatically enhance trust. In Cycle 1, trust ratings were moderate despite the presence of diagnostic and contrastive explanations. The qualitative data revealed that this was because the explanations were not usable: they were too long, too technical, or too easily overlooked. The Cycle 2 refinements, which simplified language, reduced text length, and added visual cues, produced substantial improvements in both trust and feedback uptake. This suggests that the paradox identified in the literature—that explanations may increase acceptance without improving understanding—can be addressed through user-centred design. The problem is not explanation per se, but explanation that is not designed for its audience.

5.2 The Role of Human Corroboration

The finding that instructor annotations enhanced credibility, but were initially overlooked, has implications for the design of XAI in assessment. Human oversight cannot be assumed to work simply because it is available; it must be made visible and integrated into the feedback workflow. The visual prompt added in Cycle 2 was a simple but effective design intervention. This finding aligns with Bhagavatula's (2025) observation that students value "human in the loop" validation, and extends it by showing that the presentation of human oversight matters as much as its presence.

5.3 Contextual Constraints and Design Adaptations

The study's design was shaped by the constraints of the Nigerian polytechnic context: rule-based rather than machine-learning AI, limited connectivity, and variable digital literacy. These constraints are not merely limitations; they are the conditions under which many educational technologies in sub-Saharan Africa operate. By designing for these constraints—using two-tier explanations that can be accessed with limited bandwidth, simple confidence language that does not assume technical literacy, and instructor annotations that can be added asynchronously—the study demonstrates that pedagogically meaningful XAI is achievable without sophisticated infrastructure. This is a significant contribution to a literature that has largely assumed well-resourced contexts (Ayanwale, 2025; Tınmaz, 2025).

5.4 Implications for Practice

The refined framework offers practical guidance for institutions developing or adopting AI-driven assessment systems. First, explanation should be treated as a pedagogical feature, not a technical afterthought. Institutions should require that feedback includes diagnostic content, contrastive examples, and uncertainty communication. Second, instructor involvement in the feedback loop should be designed in from the outset, with attention to visibility and accessibility. Third, explanation design should be iteratively tested with actual students; what works in theory may not work in practice, and what works for one group may not work for another. The study demonstrates that XAI can be implemented effectively in resource-constrained contexts through careful design. The refined framework—diagnostic, contrastive, calibrated, and human-corroborated—provides a transferable model for other institutions. The finding that simplification and visual cues significantly improved trust and uptake suggests that the primary barrier to effective XAI is not technical sophistication but user-centred design.

5.5 Limitations

The study has several limitations. First, the sample size is small (n = 36 across two cycles), limiting statistical power and generalisability. Second, the intervention period was short (single sessions), so long-term effects on learning and trust were not assessed. Third, the study relied on a simulated interface layer rather than full integration with the production system, which may limit ecological validity. Fourth, the participants were drawn from two departments only; students in other departments may have different digital literacy profiles and trust dynamics. Fifth, the study did not include a control group, so causal claims about the effect of the XAI features must be made cautiously. The pre/post test in Cycle 2 provides some evidence of learning gains, but without a control group, these gains cannot be attributed definitively to the XAI features.

5.6 Future Research Directions

Four priorities for future research emerge from this study. First, larger-scale studies with control groups are needed to establish causal effects of different explanation types on trust and learning outcomes. Second, longitudinal studies should examine whether trust in XAI feedback persists over time and how it evolves with repeated use. Third, research should explore how students' prior experiences with technology and their perceptions of institutional authority shape their responses to XAI. Fourth, the framework should be tested in other resource-constrained contexts, including other Nigerian polytechnics and institutions in other sub-Saharan African countries, to assess its transferability.

6. Conclusion

This design-based empirical study developed and iteratively evaluated an XAI-enhanced feedback module for an examination preparation system at Akanu Ibiam Federal Polytechnic, Unwana. Across two design cycles, we found that diagnostic and contrastive explanations improved students' comprehension of errors and their willingness to act on feedback when designed for usability. Calibrated confidence displays supported appropriate trust, and instructor annotations enhanced credibility when made visible. The refined framework—diagnostic, contrastive, calibrated, and human-corroborated—offers a practical model for pedagogically meaningful XAI in resource-constrained higher education contexts. The study demonstrates that the primary barrier to effective XAI is not technical sophistication but user-centred design, and that design-based research is a valuable methodology for developing context-sensitive AI assessment tools.

References

[1] Anderson, J. R. (2010). Cognitive psychology and its implications (7th ed.). Worth Publishers.
[2] Ayanwale, M. A. (2025). Building trust in AI-powered assessment through explainable machine learning models. Proceedings of the ACM Global Computing Education Conference 2025, 2, 403–404. https://doi.org/10.1145/3736251.3747313
[3] Bhagavatula, S. (2025). Trust through human oversight: Students' perceptions of AI explanations in education [Master's thesis, Eindhoven University of Technology]. TU/e Research Portal.
[4] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101.
[5] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340.
[6] Dawodu, O. A., Iposu, N. O., & Abimbola, O. G. (2025). Integrating explainable artificial intelligence into adaptive personalized learning method. Journal of Science and Information Technology, 19(1), 160–170.
[7] Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. Educational Researcher, 32(1), 5–8.
[8] Eze, C. N., & Nwafor, J. C. (2021). Student reliance on past examination questions as a study resource: Implications for learning and assessment practices in Nigerian tertiary institutions. Journal of Education and Practice, 12(15), 78–89.
[9] Kumar, R., Singh, P., & Patel, A. (2020). Artificial intelligence-driven examination preparation systems: Mechanisms, effectiveness, and future directions. IEEE Transactions on Education, 63(2), 89–101.
[10] Oladipo, O. A., Adekunle, B. O., & Oladejo, A. T. (2022). Digital transformation of examination preparation systems in Nigerian tertiary institutions: Challenges, opportunities, and strategic implementation frameworks. African Journal of Educational Technology and Digital Literacy, 4(3), 156–174.
[11] Onuora, A. C., Ana, P., & Otiko, A. O. (2024). Machine learning architecture for combating hate speech in Igbo language. 3rd International Conference, Academic Staff Union of Polytechnic (ASUP), Akanu Ibiam Federal Polytechnic Unwana Chapter. https://doi.org/ZENODO.20812203
[12] Onuora, A. C., Maidoh, N. E., Otiko, A. O., Ikem, M. E., Umeh, I. O., & Udo, I. N. (2026). Designing a secure and scalable architecture for a cloud-ready, AI-powered music recommendation system. International Conference 2026, March, 24, 27. https://doi.org/10.5281/ZENODO.20745687
[13] Onuora, A. C., Oyeniran, T. A., & Yesuraj, R. (2026). Quantum-safe (post-quantum) cryptography for secure network infrastructures. In Mathematics for Analysis, Algorithms, and Modelling (pp. 146–186). Deep Science Publishing.

[14] Onwukwe, P. O. (2026). Design and implementation of an AI-based examination preparation and practice question system: A case study of Akanu Ibiam Federal Polytechnic Unwana, Afikpo Ebonyi State [Unpublished National Diploma project]. Department of Computer Science, Akanu Ibiam Federal Polytechnic, Unwana.
[15] Piaget, J. (1970). The science of education and the psychology of the child. Orion Press.
[16] Skinner, B. F. (1958). The technology of teaching. Appleton-Century-Crofts.
[17] Tinmaz, H. (2025). From transparency to trust: A literature review on explainable AI in educational systems. Mersin Üniversitesi Eğitim Fakültesi Dergisi, 21(3), 999–1018. https://doi.org/10.17860/mersinefd.1739347
[18] Umeh, B. C., & Nwite, V. O. (2023). AI-enabled performance analytics platforms in Nigerian university contexts: Predictive modeling for early identification of examination risk and enhanced examination readiness. African Journal of Artificial Intelligence Research, 6(2), 112–137.
[19] Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.