Reimagining Equitable Assessment for AI-Generated Content: A Qualitative Case Study of Practitioner Judgement and Human-in-the-Loop Evaluation

Abstract
Higher-education institutions increasingly rely on generative AI to create instructional materials, yet lack validated, equitable ways to evaluate the quality of this content. This **qualitative case study** examines how ten practitioners used a prototype five-dimension rubric (Clarity, Relevance, Novelty, Tone, Variety) to evaluate AI-generated lesson materials across three domains. Data included rubric scores/comments, paired sample comparisons, and semi-structured interviews; analysis followed **reflexive thematic analysis**. Descriptive visuals (Figures 1–7) were used to situate practitioner interpretations. Findings reveal: (1) construct ambiguity, (2) missing educator-valued dimensions, (3) dependence on context metadata, (4) strong preferences for human-machine division of labor, and (5) desire for rubric flexibility. We propose a **context-aware, human-in-the-loop (HITL)** rubric aligned with values and practices of educators and informed by broader HITL scholarship (Fajardo-Ramos et al., 2025; ETS, 2025). The study offers actionable guidance for institutions seeking trustworthy evaluation frameworks for AI-generated educational content.

1. Introduction

Across higher education, generative AI now routinely contributes to lesson planning, writing prompts, and open educational resources. While institutions appreciate the **efficiency** and **scalability** of AI (Mpolomoka, 2025), they also face new concerns:

- *Is the AI's output instructionally sound?*
- *Does it align with the needs of specific learners?*
- *Can educators trust the quality and fairness of what AI produces?*

Recent scholarship stresses the importance of **human-in-the-loop (HITL)** approaches—where humans guide, supervise, and finalize judgments, while AI supports but never replaces expertise (Fajardo-Ramos et al., 2025). HITL practices

emphasize **explainability**, **traceability**, and an explicit division of responsibilities between humans and machines (ETS, 2025). However, the field lacks practitioner-tested tools for evaluating the very content AI generates.

This case study examines how real practitioners used a prototype rubric and identifies what an effective, equitable rubric would require. The study was motivated by a simple but powerful question posed in the narrative synthesis:

"What kind of rubric allows both humans and AI to align on quality—and which aspects require human judgment?"

2. Conceptual Framework

2.1 Case Study Methodology (Yin)

We follow Yin's case-study methodology, treating the evaluation process itself as a **bounded case** occurring within an authentic professional context (Yin, 2014). The case is bounded by:

- **Time**: November–December 2025
- **Participants**: 10 higher-ed practitioners
- **Artifacts**: 3 domains × ORIG/POOR versions [Ai Quality…Report v2 | Word]

This framework supports **multiple evidence sources**, **pattern matching**, and **contextual interpretation**, all central to understanding how practitioners actually engage with evaluative tools.

2.2 Reflexive Thematic Analysis (Braun & Clarke)

We use **reflexive thematic analysis** as articulated by Braun & Clarke (2006) and expanded in subsequent works emphasizing reflexivity and researcher interpretation. This approach is well-aligned with practitioner studies where meaning-making is contextual, situated, and co-constructed.

2.3 Human-Centered AI and HITL Assessment

Human-centered AI emphasizes:

- automation of *mechanical tasks*,
- preservation of *expert human judgment*,
- transparency and explainability (ETS, 2025).

These principles resonate with practitioners' comments about when AI supports them and when it cannot replace pedagogical insight.

3. Methods

3.1 Participants and Materials

Ten practitioners evaluated paired samples:

- **Persuasive Writing (G10–11)**
- **Leadership Styles (Adult/Professional)**
- **Plants OER (Grade 6)** [mdpi.com]

For each content type, participants received:

- **ORIG version** (higher quality)
- **POOR version** (intentionally degraded)

3.2 Rubric

Participants scored each sample on a **0–2 scale** across:

- Clarity
- Relevance
- Novelty
- Tone
- Variety [academia.edu]

3.3 Data Sources

- Interviews
- Rubric scores
- Qualitative comments
- Practitioner insights from the organizational synthesis [Ai Quality…Report v2 | Word]

3.4 Analysis

We conducted **reflexive thematic analysis** with triangulation across:

- scoring patterns
- participant explanations
- observed scorer uncertainty
- cross-sample comparisons (ORIG vs POOR)

Descriptive graphics (Figures 1–7) aided interpretation but do not imply generalization.

4. Descriptive Results (Contextual, Practitioner-Relevant)

4.1 ORIG vs POOR Patterns

Practitioners consistently recognized poor-quality samples.
Figure 1: ORIG outperformed POOR on Clarity (+0.44) and Relevance (+0.61).
Tone showed negligible difference—reflecting its **context-dependence**.

4.2 Divergent Judgments

Figure 2: Clarity and Novelty had the **widest dispersion**, reflecting conceptual ambiguity and inconsistent interpretation.

4.3 Artifact-Specific Profiles

Radar charts (Figures 5–7) show different "quality shapes" across content types, reinforcing practitioners' intuition that quality depends on **purpose** and **use case**.

5. Findings (Themes)

Theme 1 — Construct Ambiguity

Practitioners struggled most with **Novelty**, which lacked a clear definition and felt misaligned with classroom practice.

Theme 2 — Missing Dimensions

Participants identified these as essential to evaluate AI-generated instructional materials:

- Organization
- Flow
- Voice/Engagement
- Accuracy & Bias
- Media Diversity

Theme 3 — Context as Prerequisite

Raters frequently said they were "guessing" without details about:

- audience/level
- learning objectives
- prerequisites

- assignment role

Theme 4 — Human–Machine Complementarity

Participants wanted:

- AI → readability checks, link integrity, repetition
- Humans → nuance, ethics, pedagogy, coherence, voice

Theme 5 — Structural Flexibility

Requests included:

- Weighted dimensions
- Exemplars
- Optional expansion beyond 0–2 scale
- Ability to mark “N/A”

6. Discussion

Taken together, findings demonstrate that **evaluation of AI-generated content is not only a technical challenge but a pedagogical one**. Practitioners require:

- context,
- clarity,
- dimensions that reflect real teaching practice,
- and workflows that respect human expertise.

These needs map directly onto HITL and HCAI guidance (Fajardo-Ramos et al., 2025; ETS, 2025).

7. Recommendations

7.1 Add Context Metadata Upfront

Mandatory fields:

- Audience
- Level
- Learning objectives
- Prerequisites
- Assignment role

7.2 Expanded Rubric Dimensions

New dimensions:

- Readability & Coherence
- Organization
- Flow
- Accuracy & Bias
- Media Diversity
- Voice/Engagement
- Fit-for-Purpose

7.3 HITL Workflow

1. AI performs mechanical checks.
2. Human evaluates nuance.
3. System provides rationales and tracks versioning.

8. Conclusion

AI-generated content can support teaching—but only if educators have reliable, context-aware, equitable tools for evaluation. This study shows how practitioners interpret quality criteria and what supports they need to judge effectiveness confidently.

References

Adamakis, M., & Rachiotis, T. (2025). **Artificial intelligence in higher education: A state-of-the-art overview of pedagogical integrity, AI literacy, and policy integration**. *Encyclopedia*, 5(4), 180. https://doi.org/10.3390/encyclopedia5040180 [frontiersin.org]

Archer, E., Young, K. A., Grover, R., & Khalil, M. (2025). **Editorial: Educational evaluation in the age of artificial intelligence: Challenges and innovations**. *Frontiers in Education*, 10, 1612274. https://doi.org/10.3389/feduc.2025.1612274

Braun, V., & Clarke, V. (2006). **Using thematic analysis in psychology**. *Qualitative Research in Psychology*, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Byrne, D. (2021). **A worked example of Braun and Clarke's approach to reflexive thematic analysis**. *Quality & Quantity*, 56, 1391–1412. https://doi.org/10.1007/s11135-021-01182-y [simplypsychology.org]

Ding, S., & Magerko, B. (2025). **Rethinking AI evaluation in education: The TEACH-AI framework and benchmark for generative AI assistants**. arXiv. https://arxiv.org/pdf/2512.04107 [link.springer.com]

Educational Testing Service (ETS). (2025a, February). **Human in the loop: Human-centered AI accelerates discovery of knowledge from digital-based large-scale educational assessments** (Highlights). https://www.ets.org/research/human-in-the-loop.html [journals.sagepub.com]

Educational Testing Service (ETS). (2025b). **Human in the loop—Highlights PDF**. https://www.in.ets.org/content/dam/ets-org/Rebrand/pdf/ETS-Research-Institute-Human-in-the-loop-highlight.pdf [jstor.org]

Fajardo-Ramos, D. C., Chiappe, A., & Mella-Norambuena, J. (2025). **Human-in-the-loop assessment with AI: Implications for teacher education in Ibero-American universities**. *Frontiers in Education*, 10, 1710992. https://doi.org/10.3389/feduc.2025.1710992 [people.bath.ac.uk]

Huang, Q., Lv, C., Lu, L., & Tu, S. (2025). **Evaluating the quality of AI-generated digital educational resources for university teaching and learning**. *Systems*, 13(3), 174. https://doi.org/10.3390/systems13030174 [mdpi.com]

Mpolomoka, D. L. (2025). **Utilizing artificial intelligence for assessment in higher education**. *Pedagogical Research*, 10(3), em0243. https://doi.org/10.29333/pr/16677 [searchwork…anford.edu]

Yin, R. K. (2014). **Case study research: Design and methods** (5th ed.). SAGE. (See also 6th ed., 2017: *Case Study Research and Applications*). [nature.com], [link.springer.com]

Armfield, D., Chen, E., Omonkulov, A., Tang, X., Lin, J., Thiessen, E., & Koedinger, K. (2025). **Avalon: A human-in-the-loop LLM grading system with instructor calibration and student self-assessment**. *AIED 2025* (CCIS 2592, pp. 111–118). Springer. https://link.springer.com/chapter/10.1007/978-3-031-99267-4_14 [files.eric.ed.gov]

Appendix — Data Sources & Provenance (Enterprise)

- *Reimagining Equitable Assessment for AI-Generated Content.docx* (annotated draft with comments). [academia.edu]
- *Narrative – AI Quality Rubric Research.docx* (figure callouts and storyline used in this v3). [mdpi.com]
- *AI Quality Project – Findings & Rubric Recommendations.docx* (compiled interview synthesis). [Ai Quality…Report v2 | Word]
- Participant rubric files: *Par131*, *Par1411*, *Par141*, *Par13*, *Par13111*, and the blank instrument file. [Read Me Fi…arch Study | Word], [Par1411 AI…Recovered) | Word], [Par141 AI…*ic*Ratings | Word], [Par13 AI*C…-completed | Word],* *[Par13111 A…ic*Ratings | Word], [AI*Content…ic*Ratings | Word]