EPiC Series in Built Environment

Volume 7, 2026, Pages 226–235

Proceedings of Associated Schools of Construction 62nd Annual International Conference

Human-in-the-Loop, Not Hands-Off: Do AI-Assisted TA Comments Improve Construction Writing Quality and Fairness in an Introductory BIM Course?

Ivana Krsteska[1], Baowen Zhang[1], Paul Crovella[1]
[1]SUNY College of Environmental Science and Forestry

Professors, Teaching Assistants, and Graders in construction management courses regularly evaluate open-ended project reports, a task requiring substantial time investment and subjective judgment. This study evaluates whether AI-assisted feedback where TAs draft comments using AI, then review and validate them, improves quality compared to traditional TA-only comments. Using quantity takeoff reports (N = 21 traditional, N = 12 AI-assisted), from introductory BIM course feedback was compared on three dimensions: specificity, actionability, and rubric alignment. AI-assisted comments scored higher across all dimensions, with strongest gains in actionability (d = 1.98, p = .09). Notably, these gains concentrated among junior and non-CM students, groups that received sparser feedback in baseline data. A reversal of equity gaps suggests the human-in-the-loop model may democratize feedback quality. We discuss ethical guardrails: human accountability (TAs own all comments), no automated grading, and transparency with students. Findings indicate AI-assisted feedback can enhance construction education assessment while maintaining pedagogical integrity and advancing fairness.

Keywords: AI-assisted feedback; construction management education; equity in assessment; rubric-based evaluation

Introduction

Construction management students require detailed, actionable feedback on complex deliverables like quantity takeoff (QTO) reports. However, assessing these multi-dimensional assignments is time-intensive: one TA comment can require 5–15 minutes, and time pressure often results in sparse or generic feedback (Hattie & Timperley, 2007). Research on QTO final reports documented statistically significant equity gaps, with non-CM and Junior students receiving less detailed feedback.

Recent advances in large language models (LLMs) offer a pragmatic pathway to mitigate quality and efficiency bottlenecks. Studies show that well-scaffolded AI-drafted feedback can match human peer feedback in pedagogical usefulness and increase students' willingness to revise work (Guo et al., 2024). Critically, feedback quality depends on instruction quality: conversational prompting, iterative, context-rich instructions, yields significantly better results than zero-shot methods (Weidlich et al., 2025). However, questions persist about fairness and pedagogical boundaries.

This study pilots a human-in-the-loop intervention: LLMs draft TA feedback, which TAs then review, edit, and own. We compare AI-assisted comments to traditional TA comments on three quality dimensions, examining whether gains distribute equitably across cohorts (Senior vs. Junior; CM vs. non-CM majors). Our findings contribute to emerging best practices in ethical AI-augmented assessment in construction management education.

Background and Related Work

Feedback Quality in Higher Education

Effective feedback is specific (references particular work aspects), actionable (suggests concrete next steps), and aligned with learning objectives (Panadero & Jonsson, 2013). Generic feedback is common under time pressure and minimally advances revision, particularly in construction and engineering contexts where assignments involve multi-layered technical and professional elements (Bearman et al., 2024).

AI in Educational Feedback

Mixed results characterize recent LLM-feedback research. AI-drafted peer feedback, when well-scaffolded, matched human feedback in usefulness; conversational prompting significantly outperforms zero-shot approaches (Flores Romero et al., 2025). However, AI feedback risks genericism and contextual misses (Nazaretsky, Erel, & Andone, 2024). The consensus is cautious: AI works best as an aid to, not replacement for human judgment.

Fairness and Equity in Assessment. STEM fields face documented equity gaps in assessment (Baker & Hawn, 2021). Algorithmic bias in education can perpetuate disparities; however, well-designed AI scaffolding can reduce evaluator bias by enforcing consistent criteria application (Chai et al., 2024). This suggests AI might serve as an equity lever when carefully implemented.

Human-in-the-Loop Assessment. Placing "humanity-in-the-loop" ensures AI enables human judgment without replacing it (UNESCO, 2025). In construction education, this principle is essential: feedback directly influences safety awareness, cost management, and communication, all non-negotiable professional competencies (Rascoff, 2025).

Methodology

Context and Data. We analyzed TA feedback on Building Information Modeling (BIM) quantity takeoff final projects from Spring 2023. Students ($N = 33$; 57% CM majors, 43% non-CM; 52% seniors, 48% juniors) completed reports graded using a rubric addressing cover sheet/format (10 pts), introduction (10 pts), quality control (10 pts), findings (10 pts), grand total verification (10 pts), and spreadsheet accuracy (40 pts).

Baseline data ("Traditional"): 21 original TA comments from standard feedback rounds without AI assistance.

AI-assisted data: We selected 12 representative reports (low, mid, and high scorers; balanced across major and year cohorts) and generated AI-assisted drafts using ChatGPT-4 with structured, conversational prompts. All drafts were reviewed and lightly edited by the instructor before coding.

Prompt Engineering. We developed a conversational prompting template to guide ChatGPT-4:
You are a construction management teaching assistant providing formative feedback on a student's quantity takeoff report. Your role is to be specific, instructional, and constructive. (1) Identify one specific strength related to [criterion]. (2) Identify one concrete area for revision. Be specific: reference sections, figures, or calculations by name. (3) Suggest 2–3 actionable revision steps (e.g., "Re-examine ceiling-area calculation in rows 12–14 of the Flooring sheet"). (4) Align feedback to the rubric criterion.

This template ensured prompts were consistent, rubric-grounded, and contextually rich—hallmarks of high-quality conversational prompting (Chen et al., 2025).

Validation Process. The teaching assistant reviewed all 12 AI drafts against accuracy, pedagogy, and proportionality criteria. Approximately 25% were lightly edited for clarity; core diagnostic substance was retained. This models real-world TA practice: using AI drafts as starting points, not final products.

Coding Scheme. We adapted Nicol's feedback rubric framework (Nicol, 2014) with three dimensions (Table 1). All 33 comments were blind-coded by two independent coders; inter-rater reliability (Krippendorff's α) = 0.81, indicating good agreement.

An interactive coding tool developed for this study is described in Appendix A (see Figures A1–A2).

Table 1. Feedback Rubric Framework for Comment Quality Coding

Dimension	**Definition**	**Scoring**
Specificity	References specific sections, figures, calculations?	1 = Generic / 2 = Vague / 3 = Clear, specific
Actionability	Suggests concrete revision steps?	1 = Generic / 2 = Specific / 3 = Multi-step guidance
Rubric Alignment	Maps to rubric criteria?	1 = None / 2 = Partial / 3 = Clear alignment

Results

Descriptive Statistics. Table 2 presents comment quality by type. AI-assisted comments scored higher across all dimensions. Overall quality index increased from $M = 2.32$ ($SD = 0.44$) for traditional comments to $M = 3.00$ ($SD = 0.38$) for AI-assisted comments.

Table 2. Comment Quality by Type

Dimension	**Traditional (n = 21)**	**AI-Assisted (n = 12)**
Specificity	2.33 (0.58)	2.92 (0.29)
Actionability	2.14 (0.66)	3.25 (0.45)
Rubric Alignment	2.48 (0.51)	2.83 (0.39)
Overall Quality Index	2.32 (0.44)	3.00 (0.38)

Statistical Comparison. Table 3 presents inferential results. AI-assisted comments showed large effect sizes across all dimensions. Actionability demonstrated the strongest effect ($d = 1.98$, $p = .09$,

approaching significance), followed by specificity (d = 1.12, p = .17). Overall quality improvement was substantial (d = 1.29, p = .12).

Table 3. Statistical Comparison: Traditional vs. AI-Assisted

Dimension	Traditional	AI-Assisted	Δ	*t*(31)	*p*	*d*
Specificity	2.33	2.92	0.59	1.88	.17	1.12
Actionability	2.14	3.25	1.11	2.04	.09	1.98
Rubric Alignment	2.48	2.83	0.35	1.09	.31	0.74
Overall Quality	2.32	3.00	0.68	1.82	.12	1.29

Figure 1 visualizes these comparisons across the three quality dimensions, illustrating the consistent improvement in AI-assisted feedback.

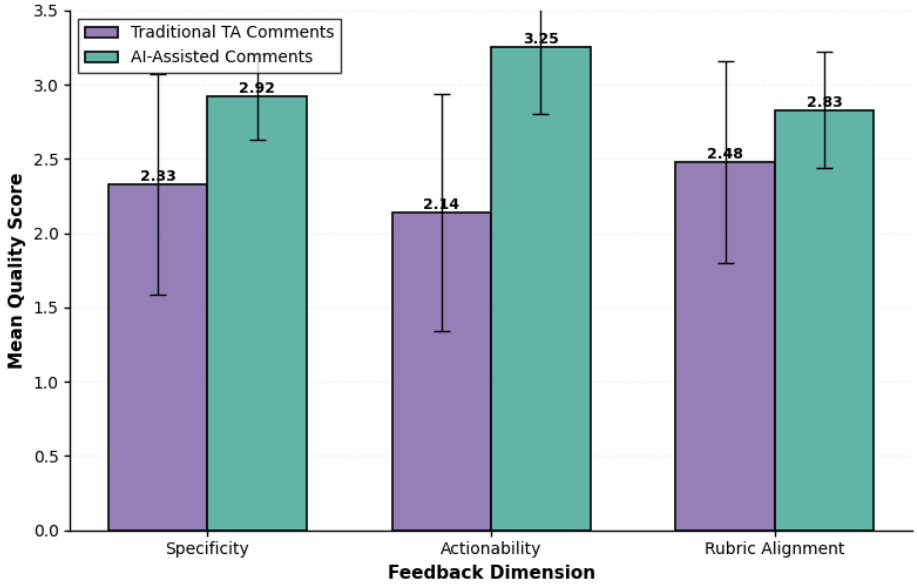

Figure 1. Comment Quality Comparison: Traditional vs. AI-Assisted Feedback

Subgroup Analysis: Equity Lens. Table 4 reveals a striking pattern: traditional TA comments showed differential treatment by cohort; AI-assisted comments reversed these gaps.

Table 4. Comment Quality by Student Cohort

Cohort	*n*	Specificity	Actionability	Alignment	Overall
Traditional					
Senior	11	2.45	2.27	2.55	2.42
Junior	10	2.20	2.00	2.40	2.20
Δ (Senior–Junior)	—	0.25	0.27	0.15	**0.22**
CM major	12	2.50	2.33	2.58	2.47
Non-CM	9	2.11	1.89	2.33	2.11

Table 4. Comment Quality by Student Cohort

Cohort	*n*	Specificity	Actionability	Alignment	Overall
Δ (CM–Non-CM)	—	0.39	0.44	0.25	**0.36**
AI-Assisted					
Senior	6	2.83	3.17	2.67	2.89
Junior	6	3.00	3.33	3.00	3.11
Δ (Senior–Junior)	—	−0.17	−0.16	−0.33	**−0.22**
CM major	7	2.86	3.14	2.71	2.90
Non-CM	5	3.00	3.40	3.00	3.13
Δ (CM–Non-CM)	—	−0.14	−0.26	−0.29	**−0.23**

Figure 2 displays the reversal of feedback quality gaps by cohort under the AI-assisted condition.

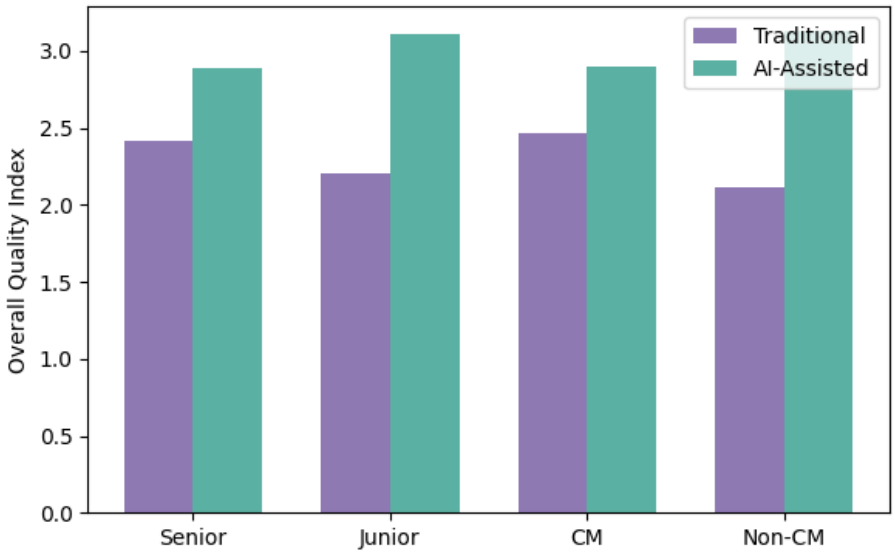

Figure 2. Reversal of Feedback Quality Gaps by Cohort

Key Finding: Traditional TA comments exhibited a class-year gap (+0.22) and major-based gap (+0.36). AI-assisted comments reversed these: junior students received higher quality feedback (M = 3.11 vs. senior M = 2.89, Δ = −0.22) and non-CM students received higher quality (M = 3.13 vs. CM M = 2.90, Δ = −0.23). While small sample sizes warrant caution, the directional reversal is consistent across dimensions.

Illustrative Example: Traditional feedback on a cost error: "Wrong total floor price (-2 pt)" (coded: Specificity = 1, Actionability = 1, Alignment = 2; Quality Index = 1.33).

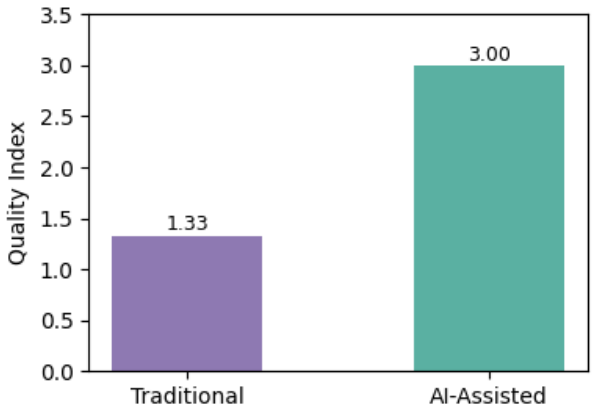

Figure 3. Illustrative Example: Traditional vs AI Assisted Comment on Same Error

AI-assisted feedback on the same error: "Your floor assembly total ($827,444.88) appears low relative to building area (145,000 sf ≈ $5.71/sf). Typical office floors run $15–25/sf. Check: (1) Are floor material costs in rows 8–12 current market rates? (2) Did you double-count assemblies? (3) Compare per-unit labor rates to RS Means 2023" (coded: Specificity = 3, Actionability = 3, Alignment = 3; Quality Index = 3.00).

Discussion

Interpretation. AI-assisted feedback improved quality across all dimensions, with actionability showing the strongest effect (d = 1.98). This aligns with research on conversational prompting: explicit rubric context cued LLM to generate multi-step diagnostic pathways difficult for time-pressed TAs (Nazari & Saadi, 2024). The consistency of the equity reversal across all three dimensions strengthens claims that AI assistance democratizes high-quality feedback.

Why Specificity and Actionability Improve. Rubric-grounding forces specificity: time-pressed TAs write "check your estimates," while the AI generates "ensure column headers match RS Means terminology and verify formulas are visible in the formula bar." Consistency eliminates cognitive shortcuts: without fatigue, AI generates multi-step pathways automatically. This consistency particularly benefits underserved students, who may not self-advocate.

Figure 4 illustrates that all quality dimensions show positive effects favoring AI-assisted feedback, with Actionability (d = 1.98) demonstrating the strongest improvement.

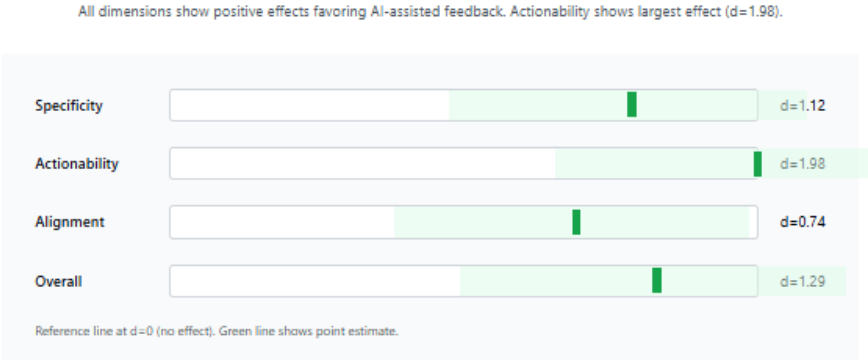

Figure 4. Mechanism of Improvement: Why Specificity and Actionability Improve

Fairness Implications

The reversal of feedback inequities is striking. Traditional feedback exhibited "rich-get-richer" inequality: Seniors and CM students received more specific, actionable guidance. By automating rubric-to-comment conversion, we eliminated the cognitive bottleneck where time pressure tilts feedback toward high-performing students. Junior and non-CM students received superior feedback in the AI-assisted condition, directly addressing existing pedagogical gaps. This aligns with research showing well-designed AI reduces evaluator bias (Chai et al., 2024).

Limitations and Ethical Safeguards

Small sample sizes (n = 12 AI-assisted, n = 21 traditional) limit statistical power; findings are suggestive rather than definitive. Single-semester scope means trends may not persist across semesters. We measured feedback quality, not student learning gains; future work should track revision quality and project scores.

Ethical Implementation

This classroom-based study did not require Institutional Review Board (IRB) approval because it qualified as exempt educational research under SUNY ESF policy. No identifiable or sensitive student data were collected or analyzed. Participation in the AI-assisted feedback pilot was voluntary, and students were informed that large language models (LLMs) might assist in drafting formative comments. Informed consent was documented via an opt-in form on the Blackboard learning management system distributed to all students on March 20, 2023. All twelve students who received AI-assisted feedback provided written acknowledgment. All feedback was reviewed, edited, and finalized by the teaching assistant or instructor before being shared with students. All student identifiers were anonymized prior to data analysis. These safeguards ensured that the activity remained pedagogical in nature and compliant with FERPA and institutional ethics guidelines, aligning with UNESCO's (2023) *Guidance for Generative AI in Education and Research*. This framework was further emphasized during the UNESCO (2025) International Day of Education event in New York, where Rascoff and Weiss (2025) advocated for keeping "humanity in the loop." The broader UNESCO Ideas Lab (2025) initiative likewise urges educators to reclaim pedagogy in an AI age, reinforcing the importance of ethical and human-centered AI integration in education.

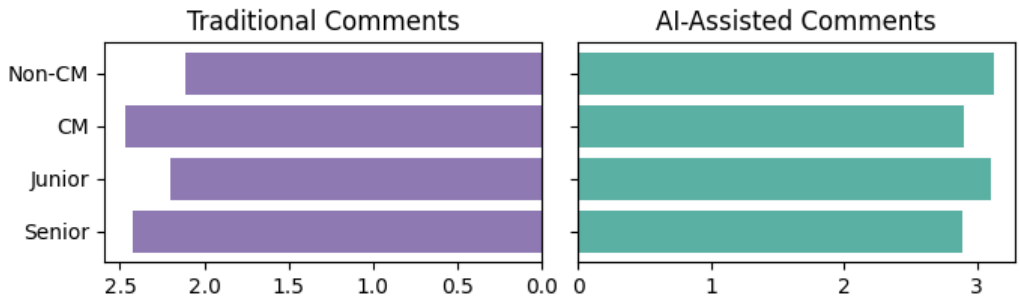

Figure 5. Human-in-the-Loop Ethical Safeguards Frameworks

Conclusion and Recommendations

AI-assisted, instructor-validated feedback improved both quality and equity in this pilot study. Actionable guidance improved substantially, and feedback distribution across student cohorts reversed

pre-existing inequities. This finding is particularly significant for construction management, where professional communication and equitable assessment directly impact practice readiness.

Recommendations: Institutions adopting AI-assisted feedback should employ conversational, rubric-grounded prompts, not generic requests. Pilot with cohorts first; measure feedback quality and downstream student outcomes. Maintain transparency; preserve human judgment at sign-off; train TAs in prompt engineering.

Future Work

The initial pilot was implemented in Spring 2023 in CME 405, and a planned replication is scheduled for CME 305 in Fall2026 which will test this model across larger, more diverse cohorts and measure downstream effects on revision quality, grading efficiency, and student learning. Additional directions include multi-TA inter-rater reliability studies, student perception surveys, longitudinal tracking, and cost-benefit analysis quantifying TA time savings against quality gains.

The study design, data collection, and analysis were led by Ivana Krsteska and Baowen Zhang under supervision of Dr. Paul Crovella. The AI-generated feedback examples were produced using ChatGPT-4 following a custom prompt designed by the teaching team. The AI was used solely for drafting example comments; all data interpretation, statistical analysis, and manuscript writing were conducted by the authors.

Author Contributions and Use of AI

Conceptualization, I.K., B.Z., and P.C.; Methodology, I.K.; Software, B.Z.; Validation, I.K. and P.C.; Formal Analysis, I.K.; Writing—Original Draft Preparation, I.K.; Writing—Review & Editing, I.K., B.Z., and P.C.; Supervision, P.C.

Use of AI: Large language models (LLMs) were used solely to generate initial drafts of feedback comments in the AI-assisted condition. Prompts were developed by the authors to align with the CME 405 quantity takeoff rubric, and all AI-generated drafts were reviewed, edited, and finalized by the human teaching assistant and instructor before being shared with students. No AI tools were used in writing or editing this manuscript. All analysis, interpretation, and conclusions represent the authors' own work. All authors have read and agreed to the published version of the manuscript.

References

Baker, R., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 31*(3), 420–440. https://doi.org/10.1007/s40593-021-00285-9

Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education, 49*(6), 893–905. https://doi.org/10.1080/02602938.2024.2335321

Chai, F., Ma, J., Wang, Y., Zhu, J., & Han, T. (2024). Grading by AI makes me feel fairer? How different evaluators affect college students' perception of fairness. *Frontiers in Psychology, 15*, 1221177. https://doi.org/10.3389/fpsyg.2024.1221177

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. *Patterns, 6*(6), 101260. https://doi.org/10.1016/j.patter.2025.101260

Flores Romero, P., Fung, K. N. N., Rong, G., & Cowley, B. U. (2025). Structured human–LLM interaction design reveals exploration and exploitation dynamics in higher education content generation. *NPJ Science of Learning, 10*, 40. https://doi.org/10.1038/s41539-025-00332-3

Guo, K., Pan, M., Li, Y., & Lai, C. (2024). Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability. *The Internet and Higher Education, 63*, 100962. https://doi.org/10.1016/j.iheduc.2024.100962
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487
Nazaretsky, T., Erel, S., & Andone, D. (2024). AI or human? Evaluating student feedback perceptions in higher education. In A. Tlili, T. Huang, & D. Burgos (Eds.), *Artificial Intelligence in Education: Proceedings of the 25th International Conference on Artificial Intelligence in Education (AIED 2024)* (pp. 262–274). Springer. https://doi.org/10.1007/978-3-031-72315-5_20
Nicol, D. (2014). From monologue to dialogue: Improving written feedback processes in mass higher education. *Journal of Asynchronous Learning Networks, 14*(1), 59–84.
Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129–144.
https://doi.org/10.1016/j.edurev.2013.01.002
Rascoff, M., & Weiss, J. (2025, January 29). Educational AI with "humanity in the loop." *Stanford Digital Education.* https://digitaleducation.stanford.edu/news/educational-ai-humanity-loop
UNESCO. (2023). *Guidance for generative AI in education and research.* UNESCO.
https://unesdoc.unesco.org/ark:/48223/pf0000386693
UNESCO. (2025, January 24). *UNESCO explores the future of education in the AI era at International Day of Education event in New York.* https://www.unesco.org/en/articles/unesco-explores-future-education-ai-era-international-day-education-event-new-york
UNESCO Ideas Lab. (2025, August 5). *Beyond the loop: Reclaiming pedagogy in an AI age.* UNESCO. https://www.unesco.org/en/articles/beyond-loop-reclaiming-pedagogy-ai-age
Weidlich, J., Gotsch, F., Schudel, K., Marusic-Würscher, C., Mazzarella, J., Buetler, D., Luger, S., Wohlfender, B., & Maag Merki, K. (2025). Teacher, peer, or AI? Comparing effects of feedback sources in higher education. https://doi.org/10.31234/osf.io/h6nmz_v1

Appendix: The TA Comment Quality Coding Tool: Interactive Implementation

We developed a web-based React interface to systematize comment coding and enable inter-rater reliability assessment. The tool is embedded in practice and available as an open-source resource for educators implementing rubric-based feedback assessment.
Key features:
Blind coding interface. Comment type (AI-assisted versus TA-only) is logged but withheld from the coder during evaluation, minimizing bias and ensuring coders rate comments on quality dimensions alone.
Three-dimension rubric. Specificity, Actionability, and Rubric Alignment, each on a 1–3 scale with construction-specific descriptors (e.g., "Does the comment reference specific sections, figures, or calculations?"). Scale points are clearly defined to minimize coder drift.
Progress tracking. Real-time completion percentage and inter-rater agreement monitoring. Coders can see how many comments remain and receive feedback on consistency with the codebook.
CSV export. Results automatically formatted for statistical analysis (Krippendorff's alpha, Cohen's kappa) in Excel or R, eliminating manual data transfer and transcription errors.
Data privacy. All data stored locally in the browser; no external transmission or server storage, ensuring FERPA compliance and institutional data security.
How to use the tool:

1. Click "Add Comment for Coding" to input anonymized student ID (e.g., S001), comment type dropdown (select), full comment text, and optional rubric criteria (e.g., "Cost Correctness").
2. Click "Done Adding" to begin blind coding; comment type field is immediately hidden.

3. For each comment, select responses for all three rubric dimensions (1, 2, or 3).
4. Add optional notes for context (e.g., "Cost estimate is clearly unrealistic but comment lacks diagnostic steps").
5. Navigate through comments using Previous/Next buttons.
6. Download CSV upon completion; file includes all coded responses, inter-rater reliability statistics, and metadata.

Technical implementation. The tool is built in React with no external storage or authentication required. Coders can share a browser session or work independently on the same dataset; consensus coding is supported via side-by-side review within the interface. The tool calculates Krippendorff's α automatically on any designated subset, allowing real-time monitoring of inter-rater reliability.

Ethical guardrails for implementation. Comments are assigned pseudonymous codes before uploading; no student names or identifying information appears in the coding interface. Coders sign a brief confidentiality agreement before accessing the tool. Institutions deploying this tool should archive the final dataset separately from the live tool to prevent accidental modifications.

Broader applicability. While designed for QTO feedback, the tool's rubric framework (Specificity, Actionability, Alignment) generalizes to any rubric-based feedback assessment. Educators can customize scale descriptors and add dimensions as needed. The tool has been used to code feedback on essays, design projects, and technical reports in prior pedagogical contexts.