Calibration and Critical Thinking: Assessing Writing in the AI Era

Emily Bald, PhD, University Writing Program

PROBLEM OF PRACTICE

Many educators worry about how generative AI (GenAI) use will impact students' critical thinking skills [1,2]. A 2025 Microsoft and Carnegie Mellon study of 319 knowledge workers found that AI use can reduce critical thinking effort, particularly among users with low self-confidence and high confidence in AI's ability to perform a task [3].

In discipline-specific writing classes, students who lack field-specific expertise may be inclined to trust AI-generated products as models of genres that are new to them. **The present case study was motivated by the challenge of how to strengthen students' ability and confidence to critically evaluate--and in turn produce--discipline-specific writing in the age of AI.**

THE STRATEGY INVESTIGATED

Calibration, or 'norming,' is a collaborative and iterative process whereby a group of evaluators aims to reach consensus about how to score sample assignments using a shared rubric. A common training tool for improving interrater reliability [4], calibration may also be beneficial in the writing classroom: norming sample assignments with students could clarify genre conventions and empower them to evaluate written products more scrupulously. Yet there is **limited scholarship on the efficacy or best practices of rater calibration [e.g., 4-8]** and, to this author's knowledge, **no published research investigating calibration *with students* as a pedagogical strategy.** Anecdotal accounts from educators suggest in-class norming can clarify understandings of proficiency and foster students' assessment and self-assessment skills, which are essential to critically engaged AI use in professional writing contexts.

Resarch Questions

- How can the calibration process be adapted into an effective teaching practice for the writing classroom?
- Can calibration sessions using AI- and human-authored samples cultivate stronger critical assessment skills among nonexperts learning a new discipline-specific genre?

Goals

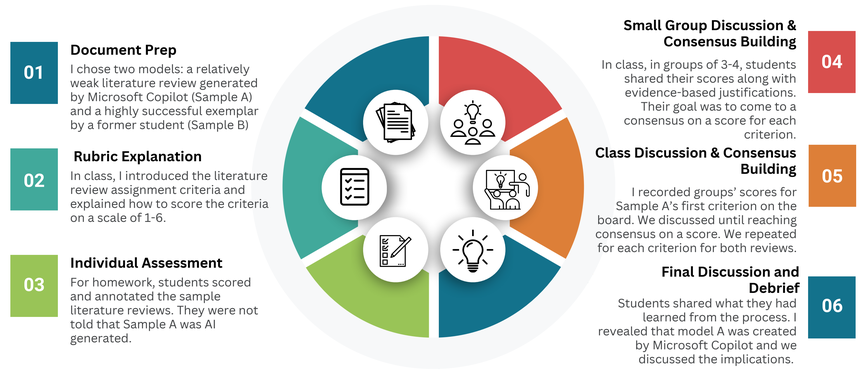
- To highlight the potential of an under-researched teaching practice for writing courses generally and for Writing in the Disciplines (WID) courses in the AI age more specifically
- To identify avenues for future research into the best practices and outcomes of in-class calibration with students

METHODS

This **case study** describes and qualitatively analyzes observations of a calibration activity carried out in two classrooms with a total of 28 third-year Writing in Interior Design students at the University of Florida. Students normed two sample literature reviews--a human-authored exemplar and a weaker review generated by Microsoft Copilot--using a shared rubric. Students were not told that one of the samples was AI-generated until the final discussion and debrief.

Calibration Process

Adapted from evidence-based calibration practices for interrater training [6-8], the process described below outlines the steps of each in-class norming session.

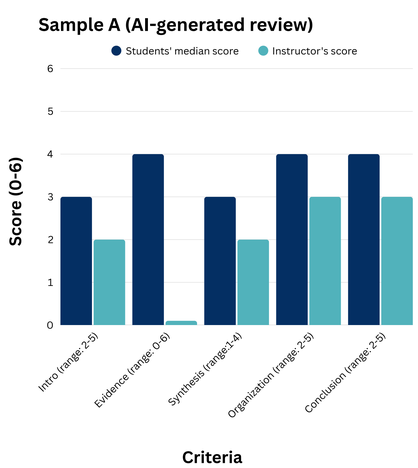


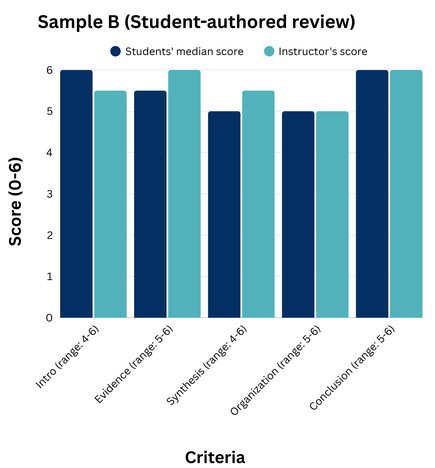
Criteria (Scored from 0-6)

Introduction	Evidence	Synthesis	Organization	Conclusion
Informs and persuades reader the review is timely and valuable; addresses topic, significance, research background, gap, and overview statement	Engages with at least 6 high-quality academic or trade sources	Demonstrates productive and clarifying synthesis of the literature	Holistic organization reflects thoughtful categorization of the literature; body is organized into topically cohesive sections with clear, descriptive headings; sections present central information before peripheral or unique information	Summarizes most significant findings, evaluates the implications of those findings, and makes recommendations for future design research and practice

Note: Assignment criteria pertaining to style, grammar, and formatting were omitted in the calibration session to encourage more focus on high-order skills.

RESULTS





DISCUSSION

There was **very little variation between groups' scores and my scores for the student-authored exemplar**. Variation by 1 point is considered acceptable in calibration sessions [6], so discussion served to reinforce with evidence, and in the language of the rubric, why the exemplar was a successful in each criterion.

There was **considerable variation among scores for the AI-generated review**. Students **overrated its success** and **overlooked key weaknesses, e.g.:**

- No group caught that 5 of the 6 sources were fabricated
- Few students noted that while the review seemed well organized, its section-level categorization of evidence was haphazard and illogical.

Consistent with literature on interrater calibration [5-8], discussing scores—and providing evidence-based justifications—was key to consensus building. As facilitator [8], I helped students learn actively, guiding the group toward consensus by identifying sources of agreement and teasing out conflicts.

Students struggled to identify weaknesses in the AI review. Some lacked the genre awareness to assess reliably, though we had spent weeks reading published reviews in the field. Before holding a calibration session, WID instructors should give students ample opportunities to engage with professional models.

In the final discussion, students shared that they had a better understanding of the review genre and my evaluative criteria, and would feel more confident using the criteria to self-assess during the drafting and revision process.

The debrief revealed weaknesses of my rubric and helped me identify how to clarify criteria (e.g., Evidence) to guide students' assessment of their own and others' (including AI-generated) writing. These outcomes align with evidence that rater calibration can support rubric and prompt revision [8].

This case study highlights several promising avenues for future research and practice:

- Measuring outcomes of in-class calibration (e.g., impacts on confidence, performance, peer review, self-assessment, and self-regulated learning)
- Developing evidence-based guidelines for calibration with students in the writing / WID classroom
- Designing calibration sessions among instructors with AI- and human-authored assignments to guide rubric revision and promote critical reflection on student learning outcomes, assignment design, and assessment practices in the age of AI

REFERENCES

1. Moran, A., & Wilkinson, B. (2025, March 2). Students' use of AI spells death knell for critical thinking. *The Guardian*. https://www.theguardian.com/technology/2025/mar/02/students-use-of-ai-spells-death-knell-for-critical-thinking
2. Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, *11*(1), 28. https://doi.org/10.1186/s40561-024-00316-7
3. Lee, H. P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025, May 26). *The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers*. CHI Conference on Human Factors in Computing Systems, Yokohama, Japan. https://doi.org/10. 1145/3706598.3713778
4. Stafford, L., Cousins, E., Bol, L., & Mize, M. (2023). Improving reliability in assessing integrative learning using rubrics: Does group norming help? *Research & Practice in Assessment*, *18*(1), 19–32.
5. Turbow, D. J., Werner, T. P., Lowe, E., & Vu, H. Q. (2016). Norming a written communication rubric in a graduate health science course. Journal of Allied Health, 45(3), e37-42. PMID: 27585624
6. Schoepp, K., Danaher, M., & Kranov, A. A. (2018). An effective rubric norming process. *Practical Assessment, Research, and Evaluation*, *23*(1), Article 1. https://doi.org/10.7275/z3gm-fp34
7. Crisp, E. A. (2017). Calibration: Are you seeing what I'm seeing? [Conference]. *Intersection*, *Winter 1*(3), 7-13.
8. Holmes, C. L., & Oakleaf, M. (2013). The official (and unofficial) rules for norming rubrics successfully. *The Journal of Academic Librarianship*, *39*(6), 599–602. https://doi.org/10.1016/j.acalib.2013.09.001