Journal of Data Science 24 (1), 264–266
January 2026

DOI: 10.6339/25-JDS1208D
Education in Data Science

Discussion of "Addressing the Challenges of AI-Generated Assignment Submissions in Education: Insights and Strategies"☆

Alyssa Columbus[1]

[1]*Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA 21205*

Keywords *academic assessment; academic integrity; AI-resilient assessment; data science education; generative AI in education*

From Integrity Enforcement to Epistemic Validity in Assessment

Wang et al. (2025) offer a timely and well-reasoned contribution to the growing literature on generative AI and academic assessment, particularly within data science and technically oriented disciplines. Their central move, shifting attention away from unreliable AI detection tools and toward assessment redesign grounded in pedagogical intent, is both necessary and overdue. In this discussion, I extend their argument by situating AI-generated assignment submissions within a broader epistemic challenge that predates generative AI but has now become unavoidable: the difficulty of drawing valid inferences about student understanding from conventional assessment artifacts.

At stake is not merely academic integrity in the narrow sense of misconduct prevention, but the *epistemic validity* of assessment itself. Generative AI forces a reconsideration of a foundational question: What, precisely, is an assignment evidence of?

Generative AI as an Epistemic Stress Test

A key insight implicit in Wang et al.'s analysis deserves explicit articulation: generative AI has not fundamentally disrupted assessment so much as it has exposed long-standing weaknesses in how learning outcomes are operationalized. When a generative model can satisfy grading criteria without engaging the intended learning process, the problem lies less with student behavior than with construct under-specification.

Many traditional assignments, particularly written reports and take-home programming tasks, exhibit high surface validity but low construct validity Darling-Hammond et al. (2019). They appear to measure understanding, synthesis, or reasoning, yet in practice they reward fluency, stylistic competence, or conformity to recognizable templates. Generative models excel precisely at these surface features Russo (2024). The resulting concern over AI-generated submissions thus reveals a mismatch between *what educators claim to assess* and *what their instruments can reliably detect.*

From this perspective, AI functions as an epistemic stress test. It reveals which assessments were already fragile proxies for learning and which remain robust when authorship is no longer a reliable signal. This reframing productively shifts the conversation away from policing toward the more difficult work of assessment repair.

Why Detection Is Conceptually Misguided

Wang et al.'s critique of AI detection tools is well-supported on technical and ethical grounds. However, the deeper problem with detection regimes is conceptual rather than computational

☆Main article: https://doi.org/10.6339/25-JDS1208.
* Email: acolumb1@jhu.edu.

Weber-Wulff et al. (2023). Detection presumes that the variable of interest is *origin* (human versus machine), whereas the educational variable of interest is *understanding*.

Even a hypothetically perfect detector would not resolve this tension. A student may produce work independently without understanding it, while another may engage deeply with AI-generated material through critique, revision, and contextualization. Origin is neither necessary nor sufficient evidence of learning. Treating it as such risks reinforcing a folk theory of authorship that is poorly aligned with modern knowledge production, especially in data science and computational fields where tool-mediated cognition is normative rather than exceptional.

The collapse of detection tools, therefore, should not be viewed as a temporary enforcement failure but as the end of an assessment paradigm that conflated authorship with epistemic agency.

Process Documentation as Epistemic Evidence

Among the strategies proposed by Wang et al., process documentation is the most conceptually consequential. Its importance lies not merely in transparency, but in rendering otherwise latent cognitive processes observable and assessable Michelene and Wylie (2014). What becomes visible is not simply *tool use*, but *judgment*: how students evaluate outputs, recognize limitations, reconcile conflicting information, and make decisions under uncertainty.

To realize this potential, process documentation must be framed explicitly as epistemic evidence rather than as an audit mechanism. In data science education in particular, competencies such as diagnostic reasoning, assumption checking, sensitivity to modeling trade-offs, and reflective calibration of trust in automated outputs are closer to professional expertise than unaided content generation. Rubrics must therefore evolve to reward these forms of reasoning explicitly. Without this alignment, process documentation risks devolving into performative compliance rather than genuine insight into student understanding.

Oral Assessment and Complementary Validity

The authors' endorsement of oral examinations and presentations is well-grounded, particularly as a means of probing conceptual coherence, transfer, and adaptive reasoning. Oral assessment can reveal whether students understand why an approach works, not merely how it was produced.

At the same time, oral formats should be understood as complementary rather than universal solutions. They introduce their own validity threats, including differences in communication style, language proficiency, and performance under time pressure. Their greatest value lies in interrogating critical decision points rather than reproducing entire analytical workflows. When paired with documented processes, oral assessment enables instructors to test the stability of understanding across contexts, which is among the strongest indicators of genuine learning in technical domains.

The Institutional Risk of "Counter-Magic"

Wang et al.'s notion of "counter-magic," in which AI is used both to generate and to evaluate student work, points to a deeper institutional risk than is often acknowledged. This is not merely an efficiency problem but an abdication of epistemic responsibility.

Assessment is inherently normative. It encodes judgments about what counts as knowledge, rigor, and competence. Delegating evaluative judgment to automated systems risks obscuring these values while simultaneously hardening them into opaque procedures. In data-driven disciplines already prone to equating automation with objectivity, this move is particularly dangerous. The authors' insistence that AI remain assistive rather than substitutive is therefore not only pedagogically sound but ethically necessary.

Toward AI-Resilient Assessment Design

Taken together, the strategies proposed by Wang et al. gesture toward a shift from AI-resistant to AI-resilient assessment. AI-resilient assessments are those in which the presence of generative tools does not invalidate the inference from student performance to student learning.

Such assessments share several characteristics: explicit articulation of epistemic targets (e.g., critique, synthesis, modeling judgment); visibility of reasoning and iteration; contextual specificity that resists generic solutions; and evaluation criteria centered on judgment rather than production alone. In this framework, AI becomes neither an illicit shortcut nor a neutral aid, but a test of whether assessment is aligned with the forms of expertise educators genuinely seek to cultivate.

Concluding Remarks

Wang et al. correctly argue that generative AI is a permanent feature of the educational landscape. The deeper challenge, however, is not technological adaptation but epistemic clarity. Educators must decide what kinds of knowing matter, how those forms of knowing manifest in observable behavior, and which assessment designs support valid inference under conditions of ubiquitous tool use.

The contribution of this paper lies in demonstrating that safeguarding academic integrity does not require retreating from AI, but rather confronting long-standing ambiguities about what assessment is for. If taken seriously, the proposed strategies do more than mitigate misconduct; they offer a path toward assessments that remain meaningful precisely because generative AI exists. In this sense, AI may ultimately strengthen education by forcing a long-overdue reckoning with what we truly value as evidence of learning.

References

Darling-Hammond L, Flook L, Cook-Harvey C, Barron B, Osher D (2019). Implications for educational practice of the science of learning and development. *Applied Developmental Science*, 24(2): 97–140. https://doi.org/10.1080/10888691.2018.1537791

Michelene T, Wylie R (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4): 219–243. https://doi.org/10.1080/00461520.2014.965823

Russo D (2024). Navigating the complexity of generative AI adoption in software engineering. *ACM Transactions on Software Engineering and Methodology*, 33(5): 1–50. https://doi.org/10.1145/3652154

Wang S, Xu L, Liu J, Zhai Y (2025). Addressing the challenges of AI-generated assignment submissions in education: Insights and strategies. *Journal of Data Science*, 24(1): 1–7. https://doi.org/10.6339/25-JDS1208

Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, Foltýnek T, Guerrero-Dib J, Popoola O, et al. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(26). https://doi.org/10.1007/s40979-023-00146-z