FROM ANSWER GENERATION TO CRITICAL VALIDATION: AN AI-MEDIATED ASSESSMENT DESIGN IN ENGINEERING MATHEMATICS

Deolinda M. L. Dias Rasteiro[1,*], Corália Pimenta[1], Maria Anciones Polo[2], Araceli Queiruga-Dios[2]

[1] *Polytechnic University of Coimbra, Coimbra, Portugal*
[2] *Universidad de Salamanca, Salamanca, Spain*

ABSTRACT

Generative artificial intelligence has made fluent solutions to mathematical problems instantly available to students, shifting the central question in engineering mathematics from whether AI is used to how assessment can make disciplinary judgement visible. This paper reports a classroom-based study of an assessment design in which AI output is deliberately externalised and converted into an object of mathematical validation. The design was implemented in two courses at the same institution — an undergraduate Statistical Methods course in Mechanical Engineering (74 students) and a master's-level Applied Mathematics to Engineering course (34 students) — through three tasks per cycle. In each task, students first produced their own solution, then consulted a generative system, documented prompts and outputs, compared the two lines of reasoning, and justified a final judgement. Work was assessed through a common rubric with three dimensions: Rigour, AI Analysis, and Clarity. Because the tasks contributed to summative marks, associations with achievement were computed against a task-independent component of the final grade. Within the master's cohort, all three dimensions improved significantly across the task sequence (Friedman, $p < .005$; for AI Analysis, T1 to T3 Cliff's $\delta = 0.56$). The undergraduate cohort showed no comparable progression: Rigour remained stable, AI Analysis fell in the second task before partially recovering, and Clarity declined monotonically with a large effect ($\delta = -0.74$, $p < .001$). Undergraduate participation also fell from 74 students in the first task to 35 in the third, with no evidence that those who withdrew were the weaker performers. The master's cohort scored higher throughout, and the gap widened across tasks. Read through the RBC+C model of abstraction in context (Hershkowitz, Schwarz, & Dreyfus, 2001), the divergence is one of consolidation, which is theorised to require repeated opportunities to revisit and reorganise prior constructs. Cohort maturity and prior selection plausibly account for the difference in level, but not for the divergence in direction: one cohort consolidated while the other deteriorated. The evidence is more consistent with cumulative workload in a large undergraduate class, an interpretation supported by time management being the lowest-rated item in both cycles and by the observed attrition. The paper argues that the educational value of generative AI lies in assessment designs that convert AI output into an object of validation and critique, and that the same design does not transfer across cycles without attention to task load, cohort size, and the number of documented AI interactions students can sustain.

***K*eywords** generative AI · engineering mathematics · assessment redesign · evaluative judgement · AI literacy · mathematics education.

References

[1] Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research? *Educational Researcher*, 41(1), 16–25. https://doi.org/10.3102/

[*] *Corresponding Author's E-mail: dml@isec.pt*

0013189X11428813

[2] Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology*, 54(5), 1160–1173. https://doi.org/10.1111/bjet.13337

[3] 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

[4] Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education*, 49(7), 1005–1016. https://doi.org/10.1080/02602938.2024.2386662

[5] Dreyfus, T., & Tsamir, P. (2004). Ben's consolidation of knowledge structures about infinite sets. *The Journal of Mathematical Behavior*, 23(3), 271–300. https://doi.org/10.1016/j.jmathb.2004.06.002

[6] Hershkowitz, R., Schwarz, B. B., & Dreyfus, T. (2001). Abstraction in context: Epistemic actions. *Journal for Research in Mathematics Education*, 32(2), 195–222. https://doi.org/10.2307/749673

[7] Korkmaz Guler, N., Dertli, Z. G., Boran, E., & Yildiz, B. (2024). An artificial intelligence application in mathematics education: Evaluating ChatGPT's academic achievement in a mathematics exam. *Pedagogical Research*, 9(2), em0188. https://doi.org/10.29333/pr/14145

[8] Luckin, R. (2025). Nurturing human intelligence in the age of AI: Rethinking education for the future. *Development and Learning in Organizations*, 39(1), 1–4. https://doi.org/10.1108/DLO-04-2024-0108

[9] Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice*, 21(6). https://doi.org/10.53761/q3azde36