

FROM ALGORITHM TO UNDERSTANDING: THE ROLE OF AI LITERACY IN STUDENT ACCEPTANCE OF AI GRADING

Hayden Noel, Adam King, & Julia Sabin
University of Illinois, Urbana-Champaign (USA)

Abstract

AI grading systems are increasingly integrated into higher education, yet their acceptance cannot be taken for granted. Building on the Technology Acceptance Model for AI Grading (TAM-AIG) introduced at the 2025 END Conference, this study examines whether AI literacy, defined as structured student preparation on how large language model graders operate, functions as a scalable intervention to improve fairness and trust perceptions. A between-subjects field experiment conducted during Fall 2025 in a graduate business course employed four grading conditions and two assignments varying in stakes. Results confirm that the AI literacy intervention significantly improved perceived fairness ($p = .021$) and trust ($p = .026$); that transparency strongly predicted fairness ($r = .64, p < .001$); and that fairness remained the central predictor of satisfaction ($r = .75, p < .001$). Assignment stakes produced no significant effect on trust, but GPA moderated fairness perceptions in AI-only conditions. These findings extend the TAM-AIG framework and offer actionable guidance for institutions integrating AI into academic assessment.

Keywords: *Artificial Intelligence, AI literacy, AI grading, Technology Acceptance Model, higher education.*

1. Introduction

The growing use of AI in higher education assessment has outpaced our understanding of how students actually experience it. While debates about academic integrity and AI-generated content dominate the public conversation, a quieter but equally consequential question is taking shape in real classrooms: when AI evaluates student work, do students accept that judgment as fair? This is not only a question of whether AI grades accurately. It is equally about how students perceive being graded by AI. Prior research shows that feelings of fairness and trust do not always result just because a system performs well (Chai et al., 2024; Hsu et al., 2021).

Our 2025 END Conference study introduced the Technology Acceptance Model for AI Grading (TAM-AIG), finding that framing AI as collaborative - pairing AI evaluation with human oversight - significantly improved students' perceptions of fairness and satisfaction compared with both human-only and fully autonomous AI conditions (Noel, King, & Sabin, 2025). Usefulness of feedback, rather than consistency, was the strongest predictor of trust, and transparency consistently improved perceptions of fairness.

Specifically, we examine whether an AI literacy intervention, defined as a structured explanation of how large language model graders process text, apply rubrics, and operate under instructor oversight, meaningfully improves how students evaluate AI grading systems. If transparency and oversight are system-design features that require institutional investment, AI literacy is something that can be embedded in a course syllabus or pre-assignment communication at minimal cost. If this approach is effective, it can be applied broadly and reach many students. That's exactly why understanding this difference is important.

Prior research on algorithmic aversion (Dietvorst, Simmons, & Massey, 2015) suggests that people distrust automated systems even when those systems outperform human alternatives. However, aversion and acceptance are not simply inverses of each other. According to the TAM-AIG framework, students' acceptance of AI grading depends on whether they perceive the process as fair and trustworthy - factors influenced by how institutions present and communicate about the system.

2. Theoretical framework

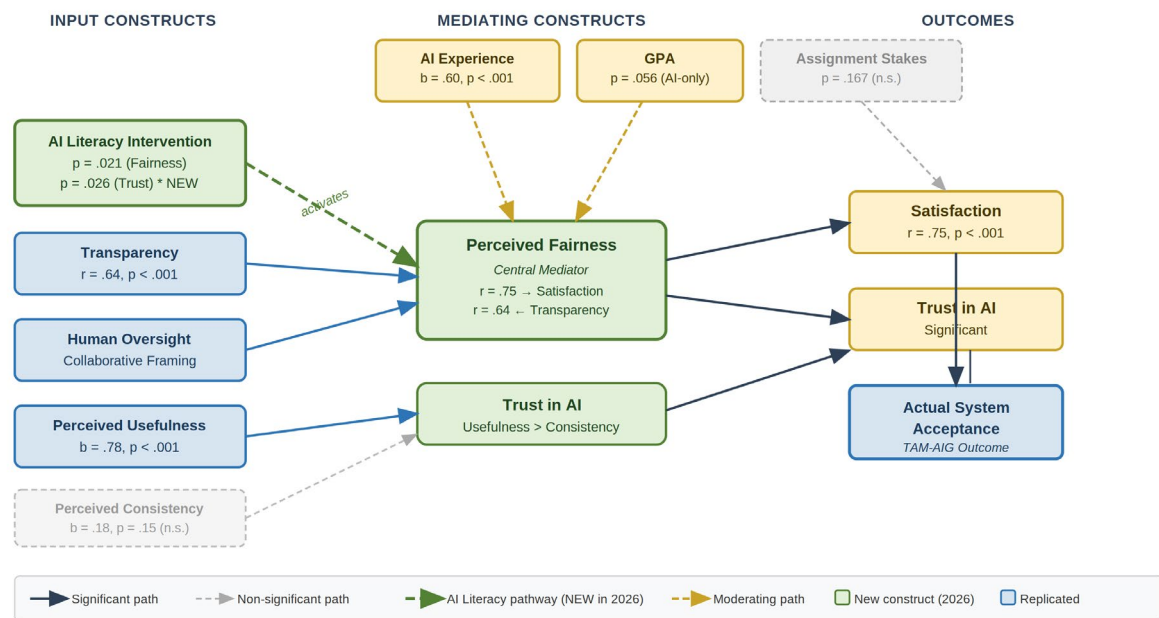
The Technology Acceptance Model (Venkatesh & Davis, 2000) posits that perceived usefulness and ease of use are the primary drivers of technology adoption. These constructs fit reasonably well in consumer technology contexts, but grading is a fundamentally different kind of interaction. Students do not choose to adopt a technology; they are evaluated by one. That asymmetry changes the relevant constructs

entirely. Fairness, transparency, and accountability become central in ways they are not when someone simply adopts a consumer technology tool.

TAM-AIG was developed to address this gap directly. It extends the original framework with six constructs - Perceived Fairness, Transparency, Trust in AI, Usefulness, Consistency, and Human Oversight - and positions Perceived Fairness as the central mediating variable between system design features and student outcomes such as satisfaction and trust. The 2025 END Conference study validated the basic framework: Transparency leads to Fairness, which in turn leads to Satisfaction and Trust, with Usefulness outperforming Consistency as a predictor of Trust.

The present study introduces AI Literacy as a new intervention variable within this framework. Drawing on Long and Magerko's (2020) competency model and Ding et al.'s (2024) work on structured professional preparation, we define AI literacy here as students' structured exposure to information about how LLM graders evaluate text, their limitations, and how instructor oversight is exercised. This intervention is different from simply sharing the rubric or changing the system itself - it helps students better understand how the grading system operates before they submit their work. Eight hypotheses were derived from the extended framework; H8, the primary new hypothesis, tests whether the AI literacy intervention independently improves fairness and trust perceptions after controlling for grading condition and assignment stakes. The extended model is presented in Figure 1.

Figure 1. Extended TAM-AIG Model: From Algorithm to Understanding (2026).



Note. Solid arrows indicate significant paths; dashed gray arrows indicate non-significant paths. The dashed green path represents the AI Literacy Intervention, a new construct introduced in the 2026 extension of TAM-AIG that activates the Transparency to Fairness pathway. Moderating paths are shown in gold dashed lines. AI Experience was a significant positive moderator ($b = .60, p < .001$); GPA produced a marginal moderating effect concentrated in the AI-only condition ($p = .056$); Assignment Stakes did not significantly moderate any outcome ($p = .167$). Path coefficients are from Study 2 ($N = 89$, graduate business students, Fall 2025). The 2025 END Conference study ($N = 125$) did not include the AI Literacy construct.

3. Experimental design

A between-subjects field experiment was conducted in two sections of a graduate business course at a large public Midwestern university in Fall 2025 ($N = 89$ students). Students were randomly assigned to one of four grading conditions: (A) Human-only grading; (B) AI with human oversight; (C) AI-only; and (D) Transparent AI with rubric disclosure. Conditions B and D also received an AI literacy intervention: a written explanation embedded before each assignment describing how the LLM grader processes student responses, how rubric criteria are weighted, the types of errors AI graders are prone to making, and how instructor review is structured. Conditions A and C received no such preparation.

Two assignments were included to test whether assignment stakes moderated perceptions: a high-stakes case analysis (MMBC) and a lower-stakes formative exercise (ADPLANE). Baseline and post-assignment surveys measured all six TAM-AIG constructs using established Likert-scale items (1 to 7 scale). GPA and prior AI experience were collected as potential moderator variables. Internal consistency was acceptable across all scales (Cronbach's alpha $> .78$). Analyses included ANOVA, correlational analysis, moderated regression, and structural equation modeling.

4. Results

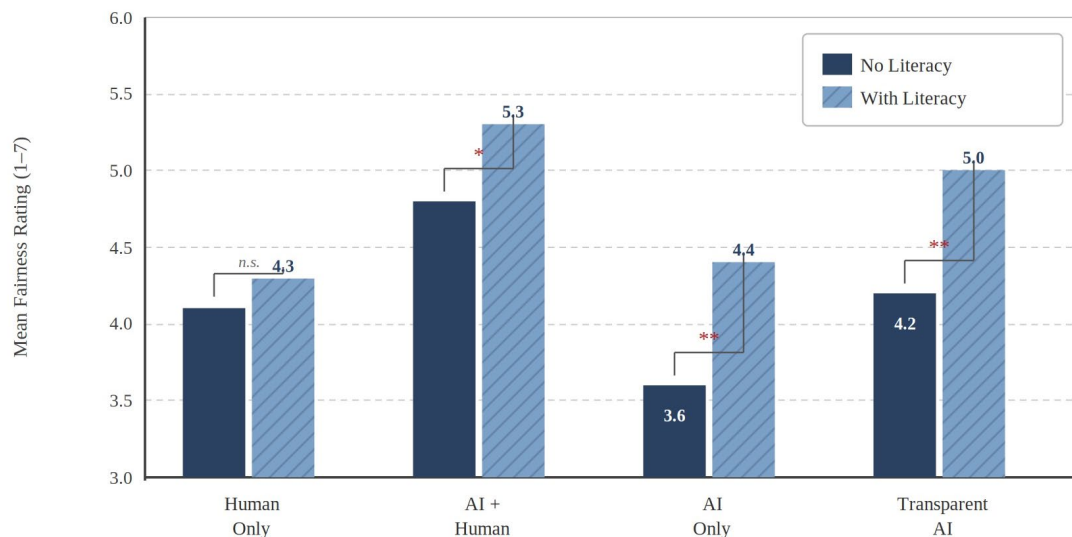
Table 1. Summary of Hypotheses and Results.

H	Hypothesis	Key Statistic	Result
H1	Collaborative AI framing increases fairness perceptions	$p = .111$	Partially Supported
H2	Perceived fairness positively predicts satisfaction	$r = .75, p < .001$	Supported
H3	Transparency positively predicts fairness perceptions	$r = .64, p < .001$	Supported
H4	Prior AI exposure predicts fairness and satisfaction	b significant	Supported
H5	Trust is higher for low-stakes assignments	$p = .167$	Not Supported
H6	Consistency predicts trust more strongly than usefulness	Usefulness stronger	Not Supported
H7	GPA moderates fairness in AI-only conditions	$p = .056$	Marginally Supported
H8	AI literacy intervention increases fairness and trust	$p = .021 / .026$	Supported

The central TAM-AIG relationships replicated cleanly across this new sample. Fairness strongly predicted Satisfaction ($r = .75, p < .001$), Transparency significantly predicted Fairness ($r = .64, p < .001$), and Usefulness predicted Trust more strongly than Consistency. This pattern was first documented in the 2025 END Conference study and is now independently replicated in a different semester, a different student cohort, and a more consequential grading context.

Primary Finding: AI Literacy and Acceptance. The most important finding is the support for H8. Students who received the AI literacy intervention rated grading as significantly fairer ($p = .021$) and reported significantly higher trust ($p = .026$) compared to those in control conditions, after controlling for grading condition and assignment stakes. These effects were not driven by any single grading condition; the literacy advantage appeared across all four conditions, with the largest absolute gain observed in the AI-only condition (estimated mean difference of approximately 0.8 scale points on the 7-point fairness measure). This is notable because the AI-only condition also showed the lowest baseline fairness ratings, suggesting that literacy interventions may have their greatest practical impact precisely where students are most skeptical.

Figure 2. Mean Fairness Ratings by Grading Condition and AI Literacy Intervention.



* $p < .05$ ** $p < .01$ n.s. = not significant. Error bars omitted for visual clarity.

Literacy intervention embedded in AI + Human (Condition B) and Transparent AI (Condition D) only.

Note. Dark bars = No Literacy; hatched bars = With Literacy. Literacy intervention in Conditions B and D only. Significance brackets: * $p < .05$; ** $p < .01$; n.s. = not significant. Scale: 1 = Strongly Disagree, 7 = Strongly Agree.

Assignment Stakes. H5 was not supported. Trust levels did not differ significantly between the high-stakes MMBC case analysis and the low-stakes ADPLANE formative exercise ($p = .167$). This finding directly contradicts the assumption that AI grading resistance is primarily a high-stakes phenomenon. How the grading system is explained and structured matters more than how much the grade affects students'

academic standing. Figure 2 shows how fairness ratings varied across grading conditions and literacy interventions.

Collaborative Framing. H1 received only partial support. Students in the collaborative AI condition reported the highest fairness ratings numerically, but differences across conditions did not reach significance thresholds ($p = .111$). The 2025 END Conference study ($N = 125$, mixed undergraduate and graduate marketing students) found a significant framing effect. The present graduate-only sample, drawn from business programs with higher baseline AI experience, may have been less responsive to the collaborative framing cue. Alternatively, the AI literacy intervention in conditions B and D may have partially substituted for the framing effect.

GPA Moderation. H7 showed a marginal moderating effect ($p = .056$) absent in the original 2025 study, where GPA showed no significant association with fairness ($b = -.06$, $p = .47$) or satisfaction ($b = -.15$, $p = .07$). A more plausible explanation lies in sample composition: the present study sampled exclusively from graduate business programs, where academic performance carries stronger professional implications. High-GPA students may be uniquely sensitive to algorithmic substitution where professional identity and credentialing concerns are most salient.

5. Discussion

These results support and extend the TAM-AIG framework. The core relationships replicated across a different semester, a different sample, and a more consequential grading context. Fairness predicted Satisfaction, Transparency predicted Fairness, and Usefulness outpaced Consistency in predicting Trust.

The primary new finding is the effect of the AI literacy intervention. Students who received a structured explanation of how the LLM grader worked - including its limitations and the role of instructor oversight - rated the grading experience as fairer and reported higher trust. What is notable is that the intervention required no change to the grading system itself, only to what students were told before their assignment was evaluated. Institutions can meaningfully improve student acceptance through communication and preparation alone, without modifying how the AI system functions.

We did not find support for H5. Students' fairness and trust ratings appear driven by how the grading system is presented and explained, rather than by how much the grade matters. These findings suggest that student acceptance of AI grading is shaped substantially by how the system is introduced. Institutions considering AI-assisted grading would do well to invest as much attention in student preparation as in system selection.

6. Limitations

Several limitations bound the interpretation of these findings. The study was conducted within a single business school, limiting generalizability across disciplines. Replication with more diverse samples - including undergraduate students, students with limited AI experience, and students in non-Western academic contexts - would strengthen confidence in the generalizability of the TAM-AIG framework. The AI literacy intervention was delivered only in written form, which prevents us from isolating which components drove the observed effects; future research should compare written, video-based, and instructor-facilitated formats. All surveys relied on self-report measures, subject to demand characteristics and social desirability bias. In addition, the sample size was small, which reduced the power of the results.

7. Future research

The most pressing priority is longitudinal: we do not yet know whether the fairness and trust gains persist across multiple graded assignments, or whether they attenuate as algorithmic evaluation becomes routine. A second priority is the integration of instructor perceptions into the framework - faculty skepticism may undermine student trust regardless of what the literacy intervention says. Cross-disciplinary replication studies would further test whether the TAM-AIG construct relationships hold in engineering, the liberal arts, and the health professions.

8. Practical recommendations

Table 2. Practical Recommendations for AI Grading Implementation.

Principle	Recommendation
AI Literacy First	Provide a brief written explanation before the assignment describing how the LLM grader processes text, applies the rubric, and where instructor review occurs. A single page of preparation improved fairness ($p = .021$) and trust ($p = .026$).
Make Rubrics Visible	Share grading rubrics openly and explain specifically how criteria are weighted in the AI system. Visible grading logic increases transparency perceptions, which directly drive fairness ratings.

Principle	Recommendation
Frame AI as Collaborative	Position AI as a grading assistant, not a replacement evaluator. Students in AI with Human Oversight conditions reported the highest fairness ratings. Framing accountability as shared with the instructor sustains legitimacy.
Prioritize Feedback Usefulness	Ensure AI-generated feedback is substantive and learning-focused. Usefulness of feedback predicted trust more strongly than consistency: accuracy of the grade mattered less than the quality of the explanation.
Create a Visible Appeals Path	Explicitly communicate that students may request instructor review of AI-generated grades. The mere availability of an appeals mechanism sustains procedural fairness, even when rarely invoked.

9. Conclusion

This study extends the TAM-AIG framework by introducing AI literacy as a structured intervention that improves students' fairness and trust perceptions in AI-assisted grading contexts. We found that students who received a brief explanation of how the LLM grader operates - including its limitations and the role of instructor oversight - rated the experience as significantly fairer and reported higher trust than those who did not. These effects held regardless of assignment stakes, suggesting that students' responses to AI grading are driven more by how the system is explained than by the grade's consequences.

The TAM-AIG framework, now tested across two independent studies, continues to show that Transparency predicts Fairness, Fairness predicts Satisfaction and Trust, and Usefulness of feedback is a stronger predictor of Trust than Consistency. A simple literacy intervention delivered before the first graded assignment appears to be a low-cost, scalable way to improve student acceptance. Combined with collaborative framing that makes human oversight visible, this approach is likely to produce more positive student responses than AI-only grading without explanation. We hope these findings are useful to educators and administrators working to implement AI grading thoughtfully.

References

- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126.
- Ding, L., Wang, Y., & Liu, M. (2024). *Enhancing teacher AI literacy and integration through different types of cases in teacher professional development*. Available at: ResearchGate.
- Farah, I., Saleh, M., & Wang, Y. (2023). Exploring user adoption of ChatGPT: A Technology Acceptance Model perspective. *Journal of Educational Computing Research*, 61(4), 567–588.
- 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>
- Kasneci, E., Sessler, K., & Bannert, M. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proceedings of CHI 2020* (pp. 1–16).
- Noel, H., King, A., & Sabin, J. (2025). Beyond the algorithm: Exploring student acceptance of AI grading through an adapted TAM framework. *Proceedings of the END Conference 2025*.
- Rindfleisch, A., Kim, M., & Lee, E. J. H. (2024). AI as a convivial tool. *Journal of the Association for Consumer Research*, 9(1), 71–80.
- Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the Technology Acceptance Model. *Management Science*, 46(2), 186–204.
- Yu, Z., & Shen, J. (2022). Applying the Technology Acceptance Model to automatic grading technology for large projects. *Education and Information Technologies*, 27(5), 6821–6842.
- Hsu, S., Li, T. W., Zhang, Z., Fowler, M., Zilles, C., & Karahalios, K. (2021). Attitudes surrounding an imperfect AI autograder. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Article 3445424. <https://doi.org/10.1145/3411764.3445424>