A Turing Test for Academic Evaluation: Detection, Grades, and Student Satisfaction in a Field Experiment

Scott Shane[*] Roman Sheremeta[*] Sining Wang[*] Heyu Xiong[*]

Weatherhead School of Management, Case Western Reserve University

March, 2026

[*] All authors are at Weatherhead School of Management, Case Western Reserve University. Authors are listed alphabetically

Abstract

We study whether AI evaluation can pass a behavioral Turing test in a consequential academic setting. In a semester-long field experiment, undergraduates in two economics courses complete multiple essay assignments, each randomly graded by one of three AI systems or one of three human teaching assistants, with students blinded to grader identity. After receiving their grade and written feedback, students report satisfaction and guess whether their grader was human or AI. We find that students cannot identify their grader type above chance (52.1% accuracy; $p = 0.229$ using a two-sided binomial test), and detection failure is uninformative: neither the grade received nor any observable feature of the evaluation predicts students' guesses. A naive comparison shows that human-graded assignments produce higher satisfaction, but this gap decomposes entirely into two channels: the grade received and students' beliefs about who graded them. Actual grader identity has no independent effect on satisfaction once these channels are controlled, and the grade–satisfaction slope is identical across grader types. These results establish behavioral substitutability under source uncertainty: AI evaluation is not merely undetected—it is functionally equivalent to human evaluation in the eyes of those being evaluated. The apparent satisfaction cost of AI grading reflects grade differences and attribution, not grader identity itself.

Keywords: AI grading; evaluative authority; Turing test; grade inflation; student satisfaction

JEL Code: I23, C93, D91, O33

1. Introduction

Artificial intelligence systems increasingly exercise evaluative authority in consequential domains, including hiring, performance reviews, credit underwriting, criminal justice, and education. The policy and scholarly debate surrounding these systems has focused heavily on predictive accuracy, bias, fairness metrics, and explainability (Bearman et al., 2022; Das et al., 2023; Gardner et al., 2021; Hartmann & Wenzelburger, 2021; Holm, 2023; Idowu, 2024; Lavanchy et al., 2023; Lepri et al., 2017; Sargeant, 2022; Starke et al., 2021). Yet institutional adoption depends not only on how algorithms perform, but also on how their decisions are perceived by those subject to them. A growing literature documents that individuals respond differently to algorithmic versus human judgments, even when outcomes are identical. In some contexts, individuals display "algorithm aversion," discounting or rejecting algorithmic decisions after observing error (Berger et al., 2020; Castelo et al., 2019; Filiz et al., 2021; Han & Ko, 2025; Maasland & Weißmüller, 2022; Renier et al., 2021). In other contexts, perceptions of fairness and legitimacy depend critically on the source of the evaluation and the availability of explanations (Chai et al., 2024; Kern et al., 2022; Marcinkowski et al., 2020). These findings suggest that the social acceptance of algorithmic authority cannot be inferred from predictive performance alone.

Despite this progress, a central empirical gap remains. Most existing research on reactions to algorithmic decisions compares explicitly labeled conditions—subjects are told "this decision was made by an AI" or "this decision was made by a human"—and asked to evaluate the outcome (Bankins et al., 2022; Baron et al., 2025; Formosa et al., 2022; Lee, 2018; Shulner-Tal et al., 2024). Source salience is baked into the experimental design; inference is not required of the subject. Yet in many real-world settings, recipients do not know, and are not told, who or what evaluated them. A student receiving a graded essay, an applicant receiving a rejection letter, or an employee

receiving a performance score may have no reliable way to determine whether the evaluation was produced by a human or a machine. We know very little about whether people can correctly infer the source of evaluative authority when it is not disclosed—or what happens to their reactions when they cannot.

The classic Turing framework offers a useful organizing logic for this question. Rather than asking whether machines are more intelligent than humans, Turing (1950) proposed a behavioral criterion: a machine exhibits intelligence if a human observer cannot reliably distinguish it from a human in interaction (Turing, 1950). Translated to evaluative authority, the relevant question becomes whether recipients can distinguish AI-generated evaluations from human-generated ones—and when they cannot, what drives their satisfaction.

We examine this question in the context of undergraduate essay grading. Grading is an especially informative domain for studying algorithmic authority under source uncertainty. First, it involves repeated, personally consequential evaluations that directly affect students—not hypothetical vignettes or one-shot scenarios. Second, students experience the actual output of the evaluation (their grade and written feedback) rather than an abstract description of a decision process. Third, grading is a standardized evaluative task with clear rubrics, allowing systematic comparison of distributional properties across evaluator types. These features permit simultaneous analysis of distributional differences between AI and human graders, students' ability to detect grader identity, and downstream satisfaction conditional on outcomes received.

Within the context of grading, the Turing framework yields two related research questions. First, in a high-stakes, repeated setting, can students correctly infer the source of their evaluation under blinded conditions? Second, when detection fails, what actually determines satisfaction—the evaluator's actual identity, the grade received, or the identity the student believes they received?

The detection and satisfaction questions are separable: the first establishes whether AI evaluation is perceptually indistinguishable from human evaluation; the second establishes whether it is behaviorally substitutable in consequential response.

Our empirical design proceeds as follows. Undergraduates across two economics courses complete 24 essay assignments over the course of a semester. Each assignment is randomly graded by either one of three AI systems or one of three human teaching assistants, under blinded conditions: students are not told which grader they received. After receiving their grades and feedback, students complete a survey measuring satisfaction with the grading and are asked to guess whether their assignment was graded by AI or a human. We first test whether students can correctly identify their grader type at rates exceeding chance. We then estimate the drivers of satisfaction using a progressive regression sequence: a benchmark model regressing satisfaction on assigned grader identity alone; a second model adding the grade received; and a third model further adding students' stated beliefs about who graded them. This sequence isolates the independent contributions of actual grader identity, grade, and perceived identity to satisfaction.

Our empirical findings show that students cannot identify grader type at rates statistically distinguishable from chance (52.1% accuracy; $p = 0.229$). Neither the grade received nor any observable feature of the evaluation predicts whether students correctly identify the grader or what they believe the grader to be. Furthermore, a naive comparison of satisfaction across grader types shows that human graders produce meaningfully higher satisfaction—but this difference is confounded by the fact that human graders also assign substantially higher grades. Once grade is controlled, the assigned grader effect falls sharply and becomes marginally significant. Once students' stated beliefs about grader identity are additionally controlled, the assigned grader effect disappears entirely. What remains is clear: satisfaction is driven by the grade received and by the

identity students believe they had—not by the evaluator's actual identity. The grade–satisfaction slope is identical across AI and human graders, and the belief effect persists regardless of whether students guessed correctly.

These findings contribute to the literature in three ways. First, we provide a clean empirical test of behavioral indistinguishability in a high-stakes evaluative setting, extending prior work on algorithm aversion and fairness perceptions beyond labeled comparisons (Burton et al., 2019; Jussupow et al., 2024; Mok et al., 2023). Second, we decompose satisfaction into three components—actual evaluator identity, grade received, and perceived evaluator identity—and show that the naive grader effect in unlabeled comparisons is accounted for entirely by grade differences and attribution, not by any intrinsic property of machine evaluators. This decomposition clarifies that institutional resistance to AI grading, if it arises, operates through outcomes and beliefs rather than through identity itself. Third, by situating grading within the broader literature on algorithmic decision-making in institutions (Dowding & Taylor, 2024; Grimmelikhuijsen & Meijer, 2022), we show that the core policy question is not simply whether algorithms are accurate, but whether they can function as legitimate substitutes for human evaluators under realistic conditions of source uncertainty—and that, under blinded conditions, they already can.

In a moment when institutions across sectors are considering whether to delegate evaluative authority to AI systems, understanding whether such systems can pass a behavioral "Turing test" of legitimacy is of first-order importance. Our evidence suggests that, under blinded conditions, AI evaluators may already be behaviorally substitutable for humans in consequential educational assessments.

2. Conceptual Framework

This section develops four testable propositions that operationalize the framework introduced above. Each proposition is derived from existing theoretical and empirical work on algorithmic decision-making, procedural justice, and source attribution. Together, they specify the conditions under which AI evaluation can be considered behaviorally substitutable for human evaluation.

2.1. Detection Under Source Uncertainty

The algorithm aversion literature has established that individuals often discount or reject algorithmic judgments relative to equivalent human judgments (Yalcin et al., 2021). However, most existing studies explicitly disclosed the source of the decisions to the subject. Participants are told that a recommendation, prediction, or evaluation was produced by an algorithm or by a human, and their reactions are measured conditional on this knowledge (Wesche et al., 2022; You et al., 2022).

This design choice leaves open a fundamental question: can individuals identify the source of an evaluation when it is not disclosed? The Turing framework provides a sharp criterion for answering this question. If AI-generated evaluations are behaviorally indistinguishable from human-generated evaluations—in the sense that recipients cannot reliably determine which type of evaluator produced the output—then the AI system passes the Turing test for that evaluative task.

In the context of grading, several features of the task environment may promote indistinguishability. Rubric-based evaluation constrains the format and content of feedback, potentially causing AI and human outputs to converge in structure. Both evaluator types assign

numerical grades on the same scale and provide written comments organized around the same rubric dimensions. If the surface characteristics of AI and human feedback are sufficiently similar, students may lack the informational basis to distinguish between them.

Proposition 1 (Detection). *Under blinded conditions, students should not identify grader type at rates exceeding chance.*

2.2. Outcome-Driven Satisfaction

A separate literature on procedural justice and outcome satisfaction suggests that individuals' reactions to evaluative decisions are driven primarily by the outcomes they receive and the perceived fairness of the process, rather than by the identity of the decision-maker per se (Magalhães, 2016; Tyler, 2000; Tyler & Caine, 1981). In contexts where the decision-maker's identity is known, source effects can contaminate outcome effects: a low grade from an AI may trigger different reactions than an identical low grade from a human, not because the grade differs but because the source is salient (Bonezzi & Ostinelli, 2021; Choi & Chao, 2024; Choung et al., 2023; Chugunova & Luhan, 2024; Lee, 2018; Wesche et al., 2022; Yalcin et al., 2021).

Under conditions of source uncertainty, however, the evaluator's identity is not available as a cue for interpreting the outcome. If students do not know—and cannot reliably infer—whether their assignment was graded by AI or a human, their satisfaction should be determined by what they directly experience: the grade received and the quality of feedback. The evaluator's hidden identity should have no independent effect on satisfaction once the outcome is conditioned upon.

Proposition 2 (Outcome Dominance). *Conditional on the grade received, actual evaluator identity will not independently drive student satisfaction.*

2.3. Perceived Identity as a Satisfaction Driver

Proposition 2 concerns the effect of actual evaluator identity on satisfaction, conditional on grade. A distinct and complementary question is whether perceived evaluator identity—the student's belief about who graded them—independently shapes satisfaction.

Even under blinded conditions, students form beliefs about who evaluated them. These beliefs may be based on stylistic features of the feedback, prior expectations about AI grading, or intuition. If satisfaction responds to the identity the student believes they received rather than the identity they actually received, then the mechanism driving reactions to evaluation is attribution-based rather than identity-based. People respond to their mental model of the evaluator, not to the evaluator's true nature.

This distinction matters both theoretically and empirically. A blinded design is uniquely capable of separating actual from perceived identity, because both are independently observed: actual identity is determined by the randomization, and perceived identity is elicited through the post-evaluation survey. In labeled designs, the two are conflated by construction. Controlling for both simultaneously allows us to determine which is the active ingredient in satisfaction: the source of the evaluation, or the student's attribution of that source.

The relevance of perceived identity is supported by the algorithm aversion literature, which has established that reactions to evaluative decisions are sensitive to how the decision is framed and attributed (Burton et al., 2019; Hou & Jung, 2021; Jussupow et al., 2020; Morewedge, 2022). When a decision is attributed to an algorithm, satisfaction and legitimacy perceptions decline even if the outcome is identical to a human decision (Yalcin et al., 2021). The literature has documented

these effects in labeled settings where the source is disclosed. Our setting raises a complementary question: does the attribution effect operate even when the label is absent and the belief is self-generated? If so, it would suggest that the mechanism is not disclosure per se but the mental model the recipient holds, regardless of how it was formed.

Proposition 3 (Perceived Identity). *Conditional on the grade received and actual evaluator identity, students' belief about grader type will independently predict satisfaction. Students who believe they received feedback from a human grader will report higher satisfaction than those who believe they received AI feedback, regardless of who actually graded them.*

2.4. Invariance of the Grade–Satisfaction Mapping

Propositions 2 and 3 address whether the level of satisfaction differs across grader types—actual and perceived—for a given grade. A further test of behavioral substitutability concerns whether the relationship between grade and satisfaction is invariant across actual evaluator types. If actual evaluator identity is truly non-salient under blinded conditions, then the marginal effect of a higher or lower grade on satisfaction should be the same regardless of whether the grade was assigned by an AI or a human.

If students harbor latent distrust of algorithmic evaluation—even without explicit knowledge of the source—one might expect asymmetric reactions to grades: a steeper satisfaction penalty for low grades from AI, or a weaker reward for high grades from AI, relative to identical grades from humans. Such an interaction would indicate that evaluator identity modulates the grade–satisfaction mapping even under conditions of source uncertainty. Behavioral substitutability, by

contrast, requires that no such interaction exist: the grade–satisfaction slope should be statistically identical across grader types.

Proposition 4 (No Interaction). *The relationship between grade and satisfaction will not differ across AI and human graders. There will be no interaction between actual evaluator type and grade in predicting satisfaction.*

Taken together, these four propositions specify a comprehensive test of behavioral substitutability under source uncertainty. Proposition 1 tests whether AI evaluation is perceptually indistinguishable from human evaluation. Propositions 2 through 4 test whether, and through what mechanism, satisfaction responds to evaluator identity—distinguishing the roles of actual identity, grade received, perceived identity, and their interactions.

3. Experimental Design and Data

3.1. Setting and Randomization

The study was conducted during the Fall 2025 semester in two undergraduate courses taught by the same instructor at a large private university in the United States. In total, 37 students participated in the experiment: 19 students in one course and 18 in the other. Students complete a sequence of 24 written assignments that are graded on a common five-point numerical scale. Each assignment is evaluated along five rubric dimensions: clarity, depth, accuracy, use of evidence, and grammar.

For each student submission, grading is performed independently by three AI graders and three human graders. Students observe only one of these six grades, randomly selected to be either from

the AI or human pool. The three AI graders are implemented using distinct foundation models configured with identical parameters and system prompts (see the Appendix for technical specifications). The three human graders are teaching assistants trained on the same grading rubric. All evaluators assess each submission independently and without knowledge of the other evaluations.

Randomization occurs at the submission level rather than the student level, so that the same student receives feedback from human graders on some assignments and from AI graders on others. Across the semester, each student observes 12 grades from human graders and 12 grades from AI graders. Students are not informed of grader identity when grades and feedback are returned. As a result, conditional on assignment timing and student characteristics, the grader type whose evaluation is observed by the student is orthogonal to potential outcomes by design.

This randomized assignment of observed grader identity allows differences in satisfaction and detectability to be causally attributed to whether feedback was delivered from a human or an AI evaluation. Because all submissions receive independent evaluations from both grader types, differences in assigned grades can be attributed directly to grader type by comparing evaluations of identical work.

3.2. Measures

The primary outcomes are assignment grades and student satisfaction. Grades are recorded on the standard course grading scale used by the instructor and are directly observed by students.

Student satisfaction is measured using a composite index constructed as the sum of three survey items drawn from the assessment and feedback dimension of the Students' Evaluation of

Educational Quality (SEEQ) instrument (Marsh, 1982). These items capture the perceived usefulness of feedback, the fairness and appropriateness of evaluation methods, and the alignment of graded materials with course content emphasized by the instructor. The SEEQ instrument has been extensively validated over multiple decades, with studies documenting stable factor structure, high internal consistency, and convergent validity across disciplines, course levels, and institutional contexts. Reported reliability coefficients for SEEQ subscales typically range from approximately 0.65 to above 0.90, depending on the sample and level of aggregation (Abrami & Marsh, 1987; Marsh, 1982).

The three items are measured on a common scale and aggregated without weighting. The resulting satisfaction index exhibits high internal consistency (Cronbach's $\alpha = 0.92$), indicating that the items load strongly on a common underlying construct and that measurement error is limited. Secondary outcomes include students' guesses about whether their assignment was graded by a human or by AI. Students do not receive any feedback on the accuracy of their guesses, so beliefs are formed without reinforcement or correction over the semester.

The study generates two linked datasets. The primary dataset records, for each assignment observation, the grader randomly selected for the student, the grade and feedback that student received, and the student's post-grade survey responses (satisfaction items and grader-type guess). This dataset has a panel structure with 846 assignment-level observations nested within 37 students across the two courses, and is the basis for all detection and satisfaction analyses.

The secondary dataset contains the full grading record: every submission is independently evaluated by all six graders (three AI, three human), yielding $846 \times 6 = 5{,}076$ grading records in total. Students never observe this full record—they see only the single grade randomly assigned

to them. This dataset is used exclusively for inter-rater reliability analysis and for characterizing the distributional properties of each grader pool on identical work.

4. Empirical Strategy

The empirical analysis first tests whether students can identify the grader type under blinded conditions, and then decomposes the determinants of satisfaction using a progressive regression sequence.

We define the following indicator variables used throughout. $HumanGrader_{it} = 1$ if student i's assignment t was randomly assigned to a human grader (the treatment), 0 if assigned to AI. $Grader_{it}$ is the numerical grade received. $BelievedHuman_{it} = 1$ if the student believes their grader was human (their stated guess), 0 if they believe it was AI. $GuessCorrect_{it} = 1$ if the student's belief matches their actual assigned grader. $GuessAI_{it} = 1 - BelievedHuman_{it}$ equals 1 if the student guesses AI, 0 if human.

4.1. Detection

We first use a two-sided binomial test to examine whether the overall proportion of correct guesses differs from 0.50. Under the null of behavioral indistinguishability, students should guess correctly at chance. We conduct this test on the full sample and separately by course section as a robustness check.

We supplement the binomial test with a set of linear probability models that ask whether any observable information predicts students' guesses. We estimate three specifications with $GuessAI_{it}$ as the dependent variable:

$$GuessAI_{it} = \alpha_i + \beta_1 HumanGrader_{it} + \varepsilon_{it} \quad (1)$$

$$GuessAI_{it} = \alpha_i + \beta_1 Grade_{it} + \varepsilon_{it} \quad (2)$$

$$GuessAI_{it} = \alpha_i + \beta_1 HumanGrader_{it} + \beta_2 Grade_{it} + \varepsilon_{it} \quad (3)$$

Equation (1) tests whether the actual grader type—the treatment—leaks into student perception. Equation (2) tests whether the grade received informs the guess. Equation (3) tests both jointly. If detection failure is uninformative noise consistent with P1, coefficients in all three specifications should be near zero and statistically insignificant.

4.2. Satisfaction Decomposition

The estimation framework here is OLS with student fixed effects α_i and a section fixed effect δ_j, where j indexes the two courses. Student fixed effects absorb all time-invariant student characteristics; the section fixed effect absorbs baseline differences in satisfaction levels across the two courses. Standard errors are clustered at the student level throughout. The primary dependent variable S_{it} is the overall satisfaction index, the mean of three SEEQ items. We estimate a sequence of six models, each adding one component to isolate its independent contribution to satisfaction.

$$S_{it} = \alpha_i + \delta_j + \beta_1 HumanGrader_{it} + \varepsilon_{it} \quad (4)$$

$$S_{it} = \alpha_i + \delta_j + \beta_1 HumanGrader_{it} + \beta_2 Grade_{it} + \varepsilon_{it} \quad (5)$$

$$S_{it} = \alpha_i + \delta_j + \beta_1 HumanGrader_{it} + \beta_2 Grade_{it} + \beta_3 BelievedHuman_{it} + \varepsilon_{it} \quad (6)$$

Model (4) establishes the naive grader effect. Model (5) tests P2 by asking how much of the $HumanGrader_{it}$ effect survives once grade is controlled. Model (6) tests P2 fully and P3 by further adding perceived identity: if β_1 becomes insignificant while β_3 is significant, actual identity does not drive satisfaction—perceived identity does. Three additional specifications address the grade–satisfaction interaction (P4) and robustness:

$$S_{it} = \alpha_i + \delta_j + \beta_1 HumanGrader_{it} + \beta_2 Grade_{it} + \beta_3 HumanGrader_{it} \times Grade_{it} + \beta_4 BelievedHuman_{it} + \varepsilon_{it} \quad (7)$$

$$S_{it} = \alpha_i + \delta_j + \beta_1 HumanGrader_{it} + \beta_2 Grade_{it} + \beta_3 BelievedHuman_{it} + \beta_4 BelievedHuman_{it} \times Grade_{it} + \varepsilon_{it} \quad (8)$$

$$S_{it} = \alpha_i + \delta_j + \beta_1 HumanGrader_{it} + \beta_2 Grade_{it} + \beta_3 BelievedHuman_{it} + \beta_4 GuessCorrect + \varepsilon_{it} \quad (9)$$

Model (7) tests P4 via the interaction $HumanGrader_{it} \times Grade_{it}$; a null β_3 indicates the grade–satisfaction slope is invariant to actual grader type. Model (8) introduces the analogous interaction with perceived identity: $BelievedHuman_{it} \times Grade_{it}$. A null interaction here means the grade–satisfaction slope does not vary by what students believe about their grader, complementing P4 on the perception side. Model (9) adds $GuessCorrect_{it}$ alongside $BelievedHuman_{it}$, distinguishing whether it is the belief itself or the accuracy of the belief that drives the perceived identity effect.

For robustness check, we estimate Model (6) three additional times substituting each individual SEEQ sub-item as the dependent variable.

Since $HumanGrader_{it}$ is randomized at the submission level, so β_1 in all models has a causal interpretation. $BelievedHuman_{it}$ is a post-treatment belief—not randomly assigned—so it captures the conditional association between perceived identity and satisfaction, holding actual identity and grade fixed. The progression from Model (4) to Model (6) decomposes the naive grader effect into two channels: the grade channel ($HumanGrader_{it} \rightarrow Grade_{it} \rightarrow S_{it}$) and the perceived identity channel ($HumanGrader_{it} \rightarrow BelievedHuman_{it} \rightarrow S_{it}$, if any).

5. Results

5.1. Descriptive Statistics

The analysis sample consists of 846 graded assignment observations drawn from two under-graduate courses over the Fall 2025 semester. By design, grader type is randomly assigned at the submission level, so that each student receives approximately half of their assignments graded by

AI and half by human TAs across the semester. The sample is exactly balanced: 423 observations are assigned to AI graders and 423 to human graders.

Table 1 reports grade distributions by grader type. Human graders assign substantially higher grades on average ($4.82, SD = 0.44$) compared to AI graders ($4.14, SD = 0.71$), a gap of 0.68 points on the five-point scale. The difference in dispersion is equally striking. As shown in Figure 1, human graders cluster almost entirely in the 4–5 range—consistent with ceiling compression—while AI graders use the full scale more evenly. Figure 2 confirms this pattern holds within each individual grader: all human graders have narrow interquartile ranges near the top of the scale, while the three AI graders spread more broadly, with AI_2 somewhat harsher ($mean = 3.89$) than AI_1 (4.24) and AI_3 (4.29). AI_2 is notably harsher than both AI1 and AI_3 (mean differences of 0.35 and 0.40 points respectively, both $p < 0.001$), while AI_1 and AI_3 produce statistically indistinguishable means ($p = 0.062$).

The gap is not driven by any single rubric dimension. Figure 3 shows that human graders outscore AI graders by $0.56 - 0.77$ points across all five dimensions (clarity, depth, accuracy, evidence, and grammar). The full dimension-level means are reported in Table 1.

Importantly, the grade gap is a calibration difference, not a quality-signal disagreement. Figure 4 plots assignment-level means for AI and human graders across all 24 homework assignments. The two series track each other closely ($r = 0.87$), confirming that both grader types respond to the same underlying variation in submission quality. The gap between them is persistent throughout the semester with no sign of convergence, ruling out learning effects or drift as explanations. Two assignments stand out as exceptions: homework 11 and homework 21 show notably larger gaps (0.87 and 0.93 points respectively), driven by lower AI means (3.99 and 4.01)

rather than higher human means. Both fall immediately before scheduled academic breaks (midterm break and Thanksgiving break). The pattern is consistent with AI graders being more responsive to variation in submission quality than human graders, whose scores remain near the ceiling regardless.

Table 2 formally establishes the robustness of this gap across three specifications. Column (1) is a base OLS model with student and assignment fixed effects; the $HumanGrader$ coefficient is 0.680 ($p < 0.001$), confirming the descriptive gap. Column (2) adds a continuous assignment-sequence variable and its interaction with $HumanGrader$ to test for time trends, and it shows that the gap neither widens nor narrows over the semester. Column (3) estimates a Tobit model with upper censoring at 5.0 to account for ceiling compression. The latent gap under Tobit is 1.174 ($p < 0.001$), substantially larger than the OLS estimate, indicating that OLS is conservative: the true difference in grading standards is understated because human grades are compressed against the scale maximum. Together, these three specifications confirm that the AI–human grading gap is statistically robust, stable over time, and not an artifact of ceiling truncation.

Table 3 reports intraclass correlation coefficients within each grader pool. Human graders achieve $ICC(2,1) = 0.869$, consistent with good reliability, with a mean absolute pairwise difference of just 0.16 points. AI graders achieve $ICC(2,1) = 0.643$—a moderate level of agreement—with a mean absolute difference of 0.55 points. The lower AI reliability is partly a scale artifact: when raters are compressed near a ceiling, as human graders are, agreement is mechanically higher. The AI models disagree more in absolute terms, but each grades systematically rather than erratically, as evidenced by the stable individual-grader means in Figure 2.

Taken together, the descriptive and regression evidence establishes a clear and robust pattern: AI and human graders differ substantially in their absolute grading standards and scale usage, but agree on the relative ordering of submission quality. The grading gap is real, uniform across rubric dimensions, persistent across the semester, and robust to ceiling correction. This sets up the central empirical question of whether students can detect this difference under blinded conditions—and whether it translates into lower satisfaction.

5.2. Can Students Detect the Grader?

We begin with the detection question: given only the grade and written feedback they receive, can students correctly identify whether their assignment was graded by a human or an AI? We first test overall guess accuracy against the 50% chance benchmark using a two-sided binomial test. Students correctly identify their grader type in 52.1% of cases ($p = 0.229$), which is statistically indistinguishable from chance. Accuracy is 55.6% in one class and 48.1% in the other, neither course displaying a consistent directional signal. Pooled across the semester, students cannot identify their grader type above chance. The confusion matrix (Figure 5) reveals the directional structure of these errors: students assigned to AI graders guessed "AI" only 39.7% of the time, with the remaining 60.3% wrongly attributed to human graders; students assigned to human graders guessed "Human" 64.5% of the time, with 35.5% wrongly attributed to AI. Students systematically lean toward guessing "Human" regardless of actual grader identity.

To characterize the structure of detection failure more precisely, we estimate three linear models with $GuessAI_{it}$ as the dependent variable, including student fixed effects, section fixed effects, and clustered standard errors throughout (Table 4, columns 1–3). If detection failure is merely

uninformative noise, then neither actual grader identity nor the grade received should predict the direction of the guess.

The evidence is consistent with this interpretation. Actual grader type does not predict guess direction (column 1: $\hat{\beta} = -0.04, p = 0.276$): students assigned to AI graders are no more likely to guess "AI" than students assigned to human graders. Grade received also does not predict the guess (column 2: $\hat{\beta} = 0.012, p = 0.114$): harsher grades do not push students toward guessing AI, nor do more generous grades push them toward guessing human. Estimated jointly (column 3), neither predictor is significant (assigned grader $p = 0.147$; grade $p = 0.742$). Finally, no individual grader is more detectable than any other. Among AI-graded assignments, neither AI_2 nor AI_3 differs significantly from AI_1 in guess accuracy (Table 5, all $p > 0.15$). Among human-graded assignments, neither $Human_2$ nor $Human_3$ differs from $Human_1$ (Table 6, all $p > 0.67$). Detectability is uniform across all graders.

Observations from the above suggest that students cannot identify grader type above chance under blinded conditions. Figure 6 confirms this holds across the entire semester: guess accuracy fluctuates around the 50% chance level assignment-by-assignment with no upward trend, ruling out learning effects. Further, neither the behavioral signal (who graded them) nor the outcome signal (what grade they received) informs the guess. This satisfies the behavioral indistinguishability condition at the heart of the Turing test framing, and it establishes the foundation for interpreting the satisfaction results that follow—any satisfaction differences cannot be attributed to students having detected, even partially, who their grader actually was.

5.3. Does the Grader's Identity Drive Satisfaction?

Table 7 reports satisfaction statistics by grader type. Overall satisfaction is higher for human-graded assignments (3.70 vs. 3.36), a raw gap of 0.34 points. The gap is consistent across all three SEEQ sub-items, with the largest difference on fair methods (0.47) and the smallest on valuable feedback (0.28). At face value, grader identity appears to matter. However, human graders also assign substantially higher grades—a gap of roughly 0.68 points per assignment—so the raw satisfaction difference is confounded: it may reflect nothing more than students being happier when they receive higher grades. The regression sequence below progressively decomposes this gap into its constituent channels.

Table 8 reports the six-column decomposition. In all models the dependent variable is S_{it} (i.e., the three-item SEEQ satisfaction index), with student fixed effects, section fixed effect, and standard errors clustered at the student level throughout.[1]

Column (1) shows that students assigned to human graders report 0.31 points higher satisfaction on average ($p < 0.001$). This is the raw grader-identity effect, unadjusted for anything the grader does differently—and it appears to validate the concern that AI grading depresses student satisfaction.

[1] The pooling of individual graders into two groups (AI and Human) is validated by an F-test of within-pool homogeneity: we re-estimate Model (6) replacing the pooled $HumanGrader_{it}$ indicator with six individual grader dummies (ref. = AI_1) and test whether the within-AI dummies (AI_2, AI_3) and within-Human dummies ($Human_2$, Human3) are jointly zero. The test does not reject ($F\ (4, 35) = 0.776, p = 0.548$), and the unrestricted model produces a lower adjusted R^2 (0.562 vs. 0.563), confirming that individual grader identity adds no explanatory power beyond pool membership.

Column (2) adds the grade received. The Human Grader coefficient falls from 0.31 to 0.08 ($p = 0.056$), a reduction of roughly 73%. Grade alone enters at $\hat{\beta} = 0.358$ ($p < 0.001$) and accounts for nearly all of the raw gap. The grade students receive, not who issued it, is the primary determinant of their satisfaction. A small residual identity effect remains marginally significant, suggesting something beyond grade still operates—but this residual is itself likely confounded by students' beliefs about the source of their grade.

Column (3) adds $BelievedHuman_{it}$—an indicator equal to one if the student guessed that their grader was human. This single variable eliminates the residual identity effect: $HumanGrader_{it}$ falls to 0.07 ($p = 0.101$), while $BelievedHuman_{it}$ enters at 0.34 ($p = 0.005$). Students who believe they received human feedback report substantially higher satisfaction regardless of who actually graded them; students who believe they received AI feedback report lower satisfaction regardless of their actual grader. The satisfaction premium attributed to human graders in the naive comparison does not reflect anything intrinsic to human evaluation—it reflects the attribution students make about its source.

Column (4) tests whether the grade–satisfaction relationship itself varies by grader type. The $HumanGrader_{it} \times Grade_{it}$ interaction enters at 0.08 ($p = 0.582$). The slope of grade on satisfaction is statistically identical for AI and human graders. There is no differential penalty for a harsh grade from AI, nor any amplified reward for a generous grade from a human. The grade channel operates with the same force regardless of who issued it.

Column (5) tests whether the grade–satisfaction slope itself varies by perceived grader identity, adding the $BelievedHuman_{it} \times Grade_{it}$ interaction to the Column (3) specification. The interaction coefficient is −0.164 ($p = 0.249$), statistically indistinguishable from zero. Students

who believe they received human grading do not derive systematically more satisfaction from a higher grade than students who believe they received AI grading, nor do they suffer a steeper satisfaction penalty from a lower grade. The grade channel operates with the same slope regardless of what students believe about their evaluator. This mirrors the finding in Column (4) for actual grader identity and extends it to perceived identity: the grade–satisfaction mapping is invariant to grader type whether defined by true assignment or by student attribution.

Column (6) adds $GuessCorrect_{it}$ —whether the student correctly identified their grader type— to the Column (3) specification. The coefficient on $GuessCorrect_{it}$ is −0.0004 ($p = 0.996$): accuracy of belief has no bearing on satisfaction. The $BelievedHuman_{it}$ coefficient is unchanged at 0.34 ($p = 0.005$). What drives the perceived-identity effect is the belief itself, not whether the belief happens to be correct. This rules out the interpretation that students are subconsciously detecting actual grader identity and that this detection, rather than the stated belief, is the operative mechanism.

Table 9 confirms that the pattern is not an artifact of index construction. Repeating the Model (6) specification with each of the three SEEQ sub-items as the dependent variable yields the same structure across all outcomes. $HumanGrader_{it}$ is never significant (valuable feedback: p = 0.281; fair methods: p = 0.384; emphasized content: $p = 0.060$). $BelievedHuman_{it}$ is always significant ($p = 0.006, 0.007, 0.018$, respectively). The result holds for perceived feedback value, for perceived fairness, and for perceived content emphasis equally—the decomposition does not hinge on any single item or any particular facet of satisfaction.

Taken together, the four propositions are all confirmed. Students cannot detect their grader above chance (P1). When they cannot detect, grade—not identity—drives satisfaction(P2).

Whatever residual satisfaction difference remains after controlling for grade traces entirely to what students believe about their grader, not who actually graded them (P3). And the grade–satisfaction mapping is grader-invariant: AI and human grades carry identical hedonic weight (P4). The implication is that the raw dissatisfaction gap observed in naive comparisons of AI and human grading is entirely a product of two separable channels—grade and attribution—neither of which is intrinsic to AI evaluation itself.

6. Discussion

6.1. Institutional Implications

The satisfaction gap between AI and human grading in our study traces largely to a grade gap. Human graders award 0.68 more points per assignment on average, with a strong ceiling concentration: the large majority of human-graded assignments cluster at or near the maximum score. AI graders apply the rubric more evenly across the scale. Our design documents this gap but does not identify its cause—the difference could reflect social pressure, effort-based sympathy, uncertainty about standards, norm-following among teaching assistants, or other mechanisms we cannot distinguish. What the design does establish is that the gap exists, is uniform across rubric dimensions, and is stable across the semester. The satisfaction consequences of that gap are what this paper identifies and decomposes.

Even after grade is fully controlled, students' beliefs about who graded them independently predict satisfaction (= 0.34, p = 0.005), consistently across all three SEEQ subitems. This is a distinct channel from the grade effect. Students who believe they received human feedback report higher satisfaction regardless of who actually graded them, and the effect does not depend on

whether the belief happens to be accurate. This finding has a practical implication: under blinded conditions, satisfaction responds not to who actually evaluates students but to who they believe evaluated them. Institutions deploying AI grading systems should therefore expect that satisfaction outcomes will depend in part on what students believe about the source of their evaluation—a consideration that operates independently of the grades those systems produce.

The detection result carries its own practical implication, independent of the satisfaction decomposition. Students in our study cannot identify their grader type above chance even after 24 repeated exposures across a full semester. AI grading systems producing structured, rubric-aligned feedback have, in this domain, reached a level of output indistinguishability from human teaching assistant grading. This is both a capability finding and a design observation: rubric structure may be the feature that makes AI and human outputs converge in ways recipients cannot distinguish. Institutions should not assume students will naturally detect or resist AI grading; the behavioral evidence across an entire semester suggests they will not.

6.2. Limitations

The satisfaction decomposition we report depends on the detection result being valid. If students could reliably identify their grader, $BelievedHuman_{it}$ would not function as an independent variable—it would collapse into a proxy for actual grader identity—and the empirical separation between the grade channel (P2) and the belief channel (P3) would not be recoverable. Our design is constructed to support this separation: random assignment of grader type, no disclosure of grader identity, and repeated exposure over a full semester under consistently blinded conditions. The results are valid within that structure. In settings where AI grading is labeled on feedback, disclosed

proactively, or stylistically detectable, the satisfaction mechanism would operate differently. The decomposition we report is a product of the blinded design and should not be generalized to transparent deployment contexts without further evidence. A natural extension—and a direct test of whether the satisfaction results survive disclosure—would be a follow-up experiment with a treatment arm in which grader identity is announced. We leave this for future work.

Several features of our setting may also limit the scope of generalizability. Our participants are undergraduates in two economics courses graded on structured five-dimension rubrics, a context where grades are the primary evaluative signal and written feedback follows a relatively standardized format. Students in higher-stakes or more qualitative evaluation settings—graduate seminars, professional performance reviews, clinical assessments—may scrutinize feedback content more carefully and be more sensitive to the identity of the evaluator. The grading gap we document is also a property of the specific prompt configuration, parameters, and rubric instructions used in this experiment—not a fixed characteristic of AI grading systems in general. A prompt designed to reward effort generously could produce grades indistinguishable from or higher than human graders, and the indistinguishability result itself depends on AI feedback that is structured and rubric aligned. AI feedback that is more generic, formulaic, or stylistically distinctive might be detectable in ways ours was not, which would tighten the link between detection and the belief channel.

The randomization of grader type affects not only the grade students receive but also the content and quality of the written feedback. We observe satisfaction responses to the full grading package—grade plus feedback—and cannot separately identify whether satisfaction differences, however small, trace to the grade, the feedback content, or their interaction. A richer analysis would require independent evaluation of the written comments produced by each grader type.

Finally, we measure student satisfaction using a validated instrument, and satisfaction is a legitimate and widely monitored indicator of evaluative legitimacy—but it is a partial welfare indicator, not a complete one. We do not claim that equal satisfaction implies equal welfare. Students who are behaviorally indifferent to grader identity under source uncertainty may nonetheless prefer human grading when given an explicit choice, and outcomes beyond satisfaction—grade appeals, learning effort, longer-run perceptions of course fairness, and actual performance—may respond differently to AI grading than immediate satisfaction does. The mechanism driving the grading gap itself also remains undetermined by our design. Human graders may award higher grades due to social pressure, effort-based sympathy, time constraints on commenting, uncertainty about standards, or norm-following—we cannot distinguish among these, and the paper makes no claim to have identified which is operative. The behavioral substitutability we document is specific to satisfaction under blinded conditions, and whether it extends to other outcomes, other welfare dimensions, and other informational environments remains an open question.

7. Conclusion

Whether AI can substitute for human judgment in consequential evaluations is one of the central institutional questions of the current moment. Most research on reactions to algorithmic decisions studies explicitly labeled conditions—subjects are told an AI or a human made the decision—and asks how they respond. We study what happens when recipients do not know, and must infer, whether their evaluation came from a human or a machine.

In a semester-long field experiment, undergraduates complete 24 essay assignments graded by either an AI system or a human teaching assistant under blinded conditions. We find that students cannot identify their grader type at rates distinguishable from chance: overall accuracy is 52.1% ($p = 0.229$) and does not improve across the semester. Detection failure is not patterned—neither the grade received nor any observable feature of the evaluation predicts what students believe about who graded them or whether they guess correctly. Under blinded conditions, AI evaluation is perceptually indistinguishable from human evaluation.

A naive comparison of satisfaction shows that human-graded assignments produce 0.34 points higher satisfaction—a gap that would appear to validate concerns about AI grading. But this gap decomposes. Controlling for the grade received reduces the identity effect by 73%; controlling additionally for students' beliefs about who graded them eliminates it entirely. What predicts satisfaction is (a) the grade the student receives and (b) what the student believes about the source—not who actually graded the assignment. The grade–satisfaction slope is identical across grader types, and whether the belief is accurate has no bearing on satisfaction. Actual grader identity, net of grade and belief, is irrelevant.

These findings reframe the policy conversation around AI grading. The satisfaction cost of AI grading is not intrinsic to AI grading itself; it flows from a grade gap that is itself better understood as evidence of human grade inflation than of AI severity. As AI grading systems are deployed more widely—in higher education, professional certification, and other evaluative contexts—the relevant question is not whether AI grades as accurately as humans but whether it achieves behavioral legitimacy in the eyes of those being evaluated. Under realistic conditions of source uncertainty, our evidence suggests that threshold has already been crossed.

Reference

Abrami, P., & Marsh, H. (1987). Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research. *Educational Researcher*. https://doi.org/10.1016/0883-0355(87)90001-2

Bankins, S., Formosa, P., Griep, Y., & Richards, D. (2022). AI Decision Making with Dignity? Contrasting Workers' Justice Perceptions of Human and AI Decision Making in a Human Resource Management Context. *Information Systems Frontiers*, *24*, 857–875. https://doi.org/10.1007/s10796-021-10223-8

Baron, S., Latham, A. J., & Varga, S. (2025). Explainable AI and stakes in medicine: A user study. *Artif. Intell.*, *340*. https://doi.org/10.1016/j.artint.2025.104282

Bearman, M., Ryan, J., & Ajjawi, R. (2022). Discourses of artificial intelligence in higher education: A critical literature review. *Higher Education*, *86*, 369–385. https://doi. org/10.1007/s10734-022-00937-2

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2020). Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn. *Business & Information Systems Engineering*, *63*, 55–68. https://doi.org/10.1007/s12599-020-00678-5

Bonezzi, A., & Ostinelli, M. (2021). Can algorithms legitimize discrimination? *Journal of experimental psychology. Applied*. https://doi.org/10.1037/xap0000294

Burton, J., Stein, M.-K., & Jensen, T. (2019). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*. https://doi. org/10.1002/bdm.2155

Castelo, N., Bos, M., & Lehmann, D. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, *56*, 809–825. https://doi.org/10.1177/0022243719851788

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*. https://doi.org/10.3389/fpsyg.2024.1221177

Choi, J., & Chao, M. M. (2024). For Me or Against Me? Reactions to AI (vs. Human) Decisions That Are Favorable or Unfavorable to the Self and the Role of Fairness Perception. *Personality & Social Psychology Bulletin*, *52*, 671–691. https://doi.org/10.1177/01461672241288338

Choung, H., Seberger, J. S., & David, P. (2023). When AI is Perceived to Be Fairer than a Human: Understanding Perceptions of Algorithmic Decisions in a Job Application Context. *International Journal of Human–Computer Interaction*, *40*, 7451–7468. https://doi.org/10.1080/10447318.2023.2266244

Chugunova, M., & Luhan, W. J. (2024). Ruled by robots: Preference for algorithmic decision makers and perceptions of their choices. *Public Choice*, *202*, 1–24. https://doi.org/10.1007/s11127-024-01178-w

Das, S., Stanton, R., & Wallace, N. (2023). Algorithmic Fairness. *Annual Review of Financial Economics*. https://doi.org/10.1257/pandp.20181018

Dowding, K., & Taylor, B. R. (2024). Algorithmic Decision-Making, Agency Costs, and Institution-Based Trust. *Philosophy & Technology*, *37*. https://doi. org /10.1007 / s13347-024-00757-5

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2021). Reducing algorithm aversion through experience. *Journal of Behavioral and Experimental Finance*, *31*. https://doi.org/10.1016/j.jbef.2021.100524

Formosa, P., Rogers, W., Griep, Y., Bankins, S., & Richards, D. (2022). Medical AI and

human dignity: Contrasting perceptions of human and artificially intelligent (AI) de-cision making in diagnostic and medical resource allocation contexts. *Comput. Hum. Behav.*, *133*. https://doi.org/10.1016/j.chb.2022.107296

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?' *Journal of Computer Assisted Learning*, *37* (5), 1207–1216. https://doi.org/10.1111/jcal.12577

Grimmelikhuijsen, S., & Meijer, A. (2022). Legitimacy of algorithmic decision-making: Six threats and the need for a calibrated institutional response. *Perspectives on Public Management and Governance*, *5* (3), 232–242. Retrieved February 23, 2026

Han, J., & Ko, D. (2025). Trust Formation, Error Impact, and Repair in Human–AI Financial Advisory: A Dynamic Behavioral Analysis. *Behavioral Sciences*, *15*. https://doi.org/10.3390/bs15101370

Hartmann, K., & Wenzelburger, G. (2021). Uncertainty, risk and the use of algorithms in policy decisions: A case study on criminal justice in the USA. *Policy Sciences*, *54*, 269–287. https://doi.org/10.1007/s11077-020-09414-y

Holm, S. (2023). Algorithmic legitimacy in clinical decision-making. *Ethics and Information Technology*, *25*, 1–10. https://doi.org/10.1007/s10676-023-09709-7

Hou, Y. T.-Y., & Jung, M. F. (2021). Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, *5*, 1–25. https://doi.org/10.1145/3479864

Idowu, J. (2024). Debiasing Education Algorithms. *International Journal of Artificial Intelligence in Education*, 1–31. https://doi.org/10.1007/s40593-023-00389-4

Johnson, V. (2004). Grade Inflation: A Crisis in College Education. https://doi.org/10.2307/40252602

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards Algorithms? A comprehensive literature Review on Algorithm aversion. Retrieved February 23, 2026, from https://consensus.app/papers/why-are-we-averse-towards-algorithms-a-comprehensive-jussupow-benbasat/0c3c7b0aed725b06bd3c438106995a3b/

Jussupow, E., Benbasat, I., & Heinzl, A. (2024). An Integrative Perspective on Algorithm Aversion and Appreciation in Decision-Making. *MIS Q.*, *48*, 1575–1590. https://doi. org/10.25300/misq/2024/18512

Kern, C., Gerdon, F., Bach, R. L., Keusch, F., & Kreuter, F. (2022). Humans versus machines: Who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making. *Patterns*, *3*. https://doi. org /10 . 1016 /j. patter . 2022 . 100591

Lavanchy, M., Reichert, P., Narayanan, J., & Savani, K. (2023). Applicants' Fairness Perceptions of Algorithm-Driven Hiring Procedures. *Journal of Business Ethics*, *188*, 125–150. https://doi.org/10.1007/s10551-022-05320-w

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*. https://doi. org/10.1177/2053951718756684

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2017). Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology*, *31*, 611–627. https://doi.org/10.1007/s13347-017-0279-x

Maasland, C., & Weißmüller, K. S. (2022). Blame the Machine? Insights From an Experiment on Algorithm Aversion and Blame Avoidance in Computer-Aided Human Resource Management. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.779028

Magalhães, P. C. (2016). Economic Evaluations, Procedural Fairness, and Satisfaction with Democracy. *Political Research Quarterly*, *69*, 522–534. https://doi.org/10.1177/1065912916652238

Marcinkowski, F., Kieslich, K., Starke, C., & Lünich, M. (2020). Implications of AI (un-)fairness in higher education admissions: The effects of perceived AI (un-)fairness on exit, voice and organizational reputation. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3351095.3372867

Marsh, H. (1982). SEEQ: A Reliable, Valid, and Useful Instrument for Collecting Students' Evaluations of University Teaching. *British Journal of Educational Psychology*, *52*, 77–95. https://doi.org/10.1111/j.2044-8279.1982.tb02505.x

Mok, L., Nanda, S., & Anderson, A. (2023). People Perceive Algorithmic Assessments as Less Fair and Trustworthy Than Identical Human Assessments. *Proceedings of*

the ACM on Human-Computer Interaction, *7*, 1–26. https://doi.org/10.1145/3610100

Morewedge, C. K. (2022). Preference for human, not algorithm aversion. *Trends in Cognitive Sciences*, *26*, 824–826. https://doi.org/10.1016/j.tics.2022.07.007

Renier, L., Mast, M. S., & Bekbergenova, A. (2021). To err is human, not algorithmic - Robust reactions to erring algorithms. *Comput. Hum. Behav.*, *124*. https://doi.org/ 10.1016/j.chb.2021.106879

Rojstaczer, S., & Healy, C. A. (2012). Where a is Ordinary: The Evolution of American College and University Grading, 1940-2009. *Teachers College Record: The Voice of Scholarship in Education*, *114*, 1–23. https://doi.org/10.1177/016146811211400707

Sargeant, H. (2022). Algorithmic decision-making in financial services: Economic and normative outcomes in consumer credit. *AI and Ethics*, 1–17. https://doi.org/10.1007/ s43681-022-00236-7

Shulner-Tal, A., Kuflik, T., Kliger, D., & Mancini, A. (2024). Who Made That Decision and Why? Users' Perceptions of Human Versus AI Decision-Making and the Power of Explainable-AI. *International Journal of Human–Computer Interaction*, *41*, 4230–4247. https://doi.org/10.1080/10447318.2024.2348843

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2021). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, *9*. https://doi.org/10.1177/20539517221115189

Turing, A. M. (1950). Computing machinery and intelligence (1950). *Mind*, *59*(236), 33–60. Retrieved February 22, 2026,

Tyler, T. (2000). Social Justice: Outcome and Procedure. *International Journal of Psychology*, *35*, 117–125. https://doi.org/10.1080/002075900399411

Tyler, T., & Caine, A. (1981). The influence of outcomes and procedures on satisfaction with formal leaders. *Journal of Personality and Social Psychology*, *41*, 642–655. https://doi.org/10.1037/0022-3514.41.4.642

Wesche, J. S., Hennig, F., Kollhed, C. S., Quade, J., Kluge, S., & Sonderegger, A. (2022). People's reactions to decisions by human vs. algorithmic decision-makers: The role of explanations and type of selection tests. *European Journal of Work and Organizational Psychology*, *33*, 146–157. https://doi.org/10.1080/1359432x.2022.2132940

Yalcin, G., Lim, S., Puntoni, S., & van Osselaer, Stijn M. J. (2021). Thumbs Up or Down: Consumer Reactions to Decisions by Algorithms Versus Humans. *Journal of Marketing Research*, *59*, 696–717. https://doi.org/10.1177/00222437211070016

You, S., Yang, C., & Li, X. (2022). Algorithmic versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation? *Journal of Management Information Systems*, *39*, 336–365. https://doi.org/10.1080/07421222.2022.2063553

Table 1: Grade Distributions by Grader Type

	AI Graders	**Human Graders**	**Difference**
	Mean (SD)	Mean (SD)	(AI −Human)
Overall and by rubric dimension			
Overall grade	4.138 (0.707)	4.818 (0.438)	−0.680
Clarity	4.122 (0.738)	4.893 (0.412)	−0.771
Depth	4.079 (0.762)	4.740 (0.619)	−0.661
Accuracy	4.130 (0.808)	4.825 (0.567)	−0.695
Evidence	4.225 (0.827)	4.782 (0.597)	−0.557
Grammar	4.135 (0.732)	4.907 (0.423)	−0.772
By individual grader			
AI_1	4.236 (0.607)	–	–
AI_2	3.891 (0.871)	–	–
AI_3	4.288 (0.530)	–	–
$Human_1$	–	4.851 (0.427)	–
$Human_2$	–	4.812 (0.408)	–
$Human_3$	–	4.790 (0.476)	–
Observations	423	423	

Notes: Grades are measured on a five-point scale. Overall grade is the mean across all five rubric dimensions. The difference column reports AI mean minus human mean; all differences are negative, indicating human graders assign higher scores on every dimension. The correlation between assignment-level AI and human grade means is $r = 0.87$, indicating both grader types respond to the same underlying quality signal despite differing calibration.

Table 2: Grading Regressions: The AI–Human Grade Gap

Dependent Variable: Grade	OLS	OLS + Trend	Tobit
$HumanGrader_{it}$	0.680***	0.607***	1.174***
	(0.054)	(0.066)	(0.028)
$AssignmentNumber_{it}$		0.013***	
		(0.002)	
$HumanGrader_{it} \times AssignmentNumber_{it}$		0.006	
		(0.003)	
Assignment FE	Yes	No	Yes
Student FE	Yes	Yes	No
Section FE	Yes	Yes	Yes
Robust S.D.	Yes	Yes	Yes
Observations	5,127	5,127	5,127
Adj. R^2	0.480	0.424	–

Notes: Column (1) is OLS with student and assignment fixed effects; standard errors clustered at the student level. Column (2) replaces assignment fixed effects with a continuous assignment-sequence variable and its interaction with Human Grader, testing whether the gap changes over the semester; standard errors clustered at the student level. Column (3) is a Tobit model with upper censoring at 5.0 (the scale maximum), accounting for ceiling compression; standard errors based on the numerical Hessian of the log-likelihood. The Tobit coefficient on Human Grader reflects the latent gap net of censoring and exceeds the OLS estimate, indicating that OLS is conservative due to ceiling compression. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.10$.

Table 3: Inter-Rater Reliability by Grader Pool

	AI Graders	**Human Graders**
ICC(2,1)	0.643	0.869
Mean absolute difference	0.55	0.16
ICC(2,1) by rubric dimension		
Clarity	0.634	0.835
Depth	0.691	0.863
Accuracy	0.458	0.780
Evidence	0.518	0.851
Grammar	0.681	0.701
Observations (rater-pairs)	2,566	2,560

Notes: ICC(2,1) is the intraclass correlation coefficient from a two-way random effects model, appropriate for assessing agreement among a fixed set of raters on a common set of targets. Mean absolute difference is the average pairwise absolute difference in scores between raters within each pool. Each submission receives three independent evaluations from each grader pool.

Table 4: Detection Regressions: Predictors of Grader-Type Guess

Dependent Variable: **GuessAI$_{it}$**	(1)	(2)	(3)
HumanGrader$_{it}$	- 0.004		- 0.050
	(0.040)		(0.035)
Grade$_{it}$		-0.012	0.011
		(0.037)	(0.033)
Student FE	Yes	Yes	Yes
Section FE	Yes	Yes	Yes
Robust S.D.	Yes	Yes	Yes
Observations	846	846	846
Adj. R^2	0.359	0.357	0.359

Notes: OLS linear probability models. Dependent variable equals 1 if the student guesses their grader was AI, 0 if human. HumanGrader$_{it}$ = 1 if the assignment was randomly assigned to a human grader. Standard errors (in parentheses) are clustered at the student level. Intercepts absorbed by student fixed effects and not reported. p-values: (1) HumanGrader: $p = 0.276$; (2) Grade: $p = 0.114$; (3) HumanGrader: $p = 0.147$, Grade: $p = 0.742$.

Table 5: Individual AI Model Detectability (AI-Graded Assignments Only)

Dependent Variable:	**GuessRight$_{it}$**	**GuessRight$_{it}$**
AI_2 (ref: AI_1)	*0.061*	*0.011*
	(0.057)	*(0.058)*
AI_3 (ref: AI_1)	0.079	0.084
	(0.057)	(0.059)
$Grade_{it}$		- 0.023
		(0.049)
Student FE	Yes	Yes
Section FE	Yes	Yes
Robust S.D.	Yes	Yes
Observations	423	423
Adj. R^2	0.350	0.349

Notes: Sample restricted to AI-graded assignments. OLS with student fixed effects and standard errors clustered at the student level (in parentheses). Dependent variable equals 1 if the student correctly identified the grader type. AI_1 is the omitted category. *p*-values: (1) AI_2: $p = 0.788$, AI_3: $p = 0.164$; (2) AI_2: $p = 0.848$, AI_3: $p = 0.156$, Grade: $p = 0.642$.

Table 6: Individual Human TA Detectability (Human-Graded Assignments Only)

Dependent Variable:	**GuessRight$_{it}$**	**GuessRight$_{it}$**
Human$_2$ (ref: Human$_1$)	0.013	0.011
	(0.050)	(0.050)
Human$_3$ (ref: Human$_1$)	-0.019	-0.035
	(0.040)	(0.040)
Grade$_{it}$		- 0.112
		(0.043)
Student FE	Yes	Yes
Section FE	Yes	Yes
Robust S.D.	Yes	Yes
Observations	423	423
Adj. R^2	0.364	0.372

Notes: Sample restricted to human-graded assignments. OLS with student fixed effects and standard errors clustered at the student level (in parentheses). Dependent variable equals 1 if the student correctly identified the grader type. Human$_1$ is the omitted category. p-values: (1) Human$_2$: $p = 0.741$, Human$_3$: $p = 0.676$; (2) Human$_2$: $p = 0.781$, Human$_3$: $p = 0.460$, Grade: $p = 0.01$.

Table 7: Satisfaction Statistics by Grader Type

	AI Graders	**Human Graders**	**Pooled**	**Difference**
	Mean (SD)	Mean (SD)	Mean (SD)	(AI − Human)
Overall satisfaction index	3.36 (1.18)	3.70 (1.08)	3.53 (1.15)	−0.34
Valuable feedback	3.19 (1.24)	3.47 (1.16)	3.33 (1.21)	−0.28
Fair methods	3.31 (1.31)	3.78 (1.18)	3.54 (1.27)	−0.47
Emphasized content	3.57 (1.26)	3.86 (1.17)	3.72 (1.22)	−0.29
Observations	423	423	846	

Notes: Satisfaction is measured using three items from the SEEQ assessment and feedback subscale (Cronbach's $\alpha = 0.92$), each on a five-point scale. The overall satisfaction index is the mean of the three sub-items. The difference column reports AI mean minus human mean. The raw satisfaction gap is confounded by grade differences; see the regression results in Section 5.3 for the decomposition.

Table 8: Satisfaction Decomposition: Actual Identity, Grade, and Perceived Identity

Dependent variable **Satisfaction**	(4)	(5)	(6)	(7)	(8)	(9)
$HumanGrader_{it}$	0.311***	0.084*	0.067	−0.301	0.056	0.067
	(0.069)	(0.044)	(0.041)	(0.663)	(0.042)	(0.045)
$Grade_{it}$		0.358***	0.362***	0.339***	0.473***	0.362***
		(0.071)	(0.071)	(0.093)	(0.126)	(0.071)
$BelievedHuman_{it}$			0.340**	0.344**	1.072	0.340**
			(0.122)	(0.122)	(0.694)	(0.121)
$HumanGrader_{it} \times Grade_{it}$				0.079		
				(0.144)		
$Grade_{it} \times BelievedHuman_{it}$					−0.164	
					(0.142)	
$GuessCorrect_{it}$						−0.000
						(0.071)
Section FE	Yes	Yes	Yes	Yes	Yes	Yes
Student FE	Yes	Yes	Yes	Yes	Yes	Yes
Robust S.D.	Yes	Yes	Yes	Yes	Yes	Yes
Observations	846	846	846	846	846	846
Adj. R^2	0.523	0.550	0.563	0.563	0.565	0.563

Notes: OLS panel regressions. Dependent variable is the overall satisfaction index (mean of three SEEQ sub-items, scale 1–5). $HumanGrader_{it}$ equals 1 if the assigned grader is human. $BelievedHuman_{it}$ equals 1 if the student guessed their grader was human. $GuessCorrect_{it}$ equals 1 if the guess was correct. All models include student fixed effect and section fixed effect. Standard errors clustered at the student level in parentheses.

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.10$.

Table 9: Sub-Item Robustness: Model (6) Repeated for Each SEEQ Item

	Valuable Feedback	Fair Methods	Emphasized Content
$HumanGrader_{it}$	0.062	0.050	0.089*
	(0.058)	(0.057)	(0.047)
$Grade_{it}$	0.286***	0.571***	0.230**
	(0.069)	(0.099)	(0.080)
$BelievedHuman_{it}$	0.341**	0.363**	0.316*
	(0.124)	(0.134)	(0.134)
Section FE	Yes	Yes	Yes
Student FE	Yes	Yes	Yes
Robust S.D.	Yes	Yes	Yes
Observations	846	846	846
Adj. R^2	0.472	0.528	0.574

Notes: OLS panel regressions repeating the specification in Table 8 with each SEEQ sub-item as the dependent variable (scale 1–5). All models include student and section fixed effects. Standard errors clustered at the student level in parentheses. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.10$.

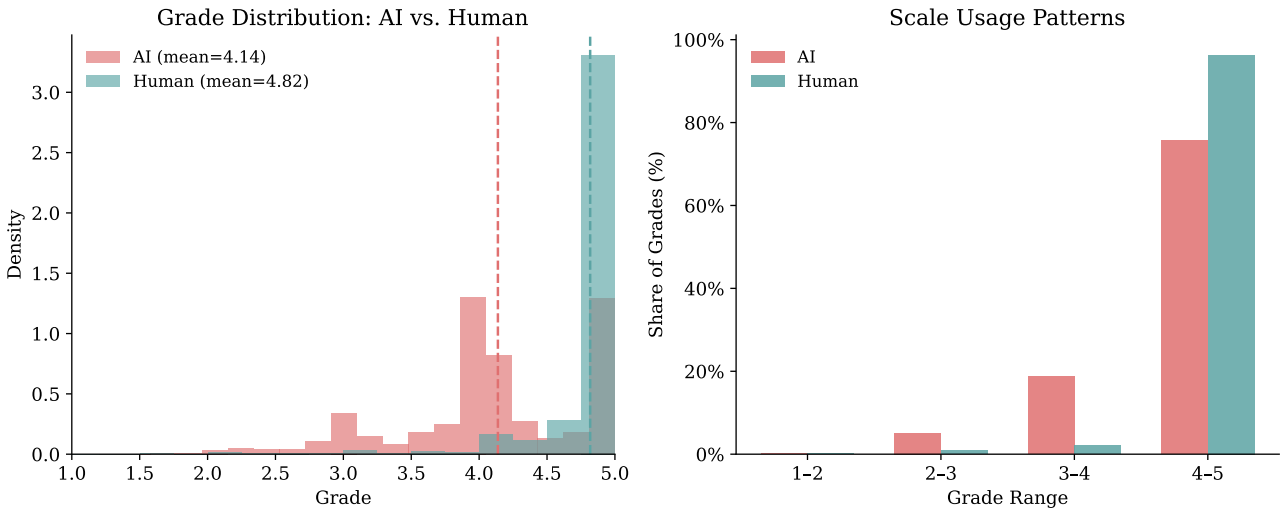


Figure 1: Scale Use Patterns: AI vs. Human Graders. *Left*: Density distributions of assignment grades. Human graders cluster heavily at the ceiling (mean = 4.82) relative to AI graders (mean = 4.14). *Right*: Share of grades falling in each one-point range. Human graders assign nearly all grades in the 4–5 range; AI graders make greater use of the lower portions of the scale.

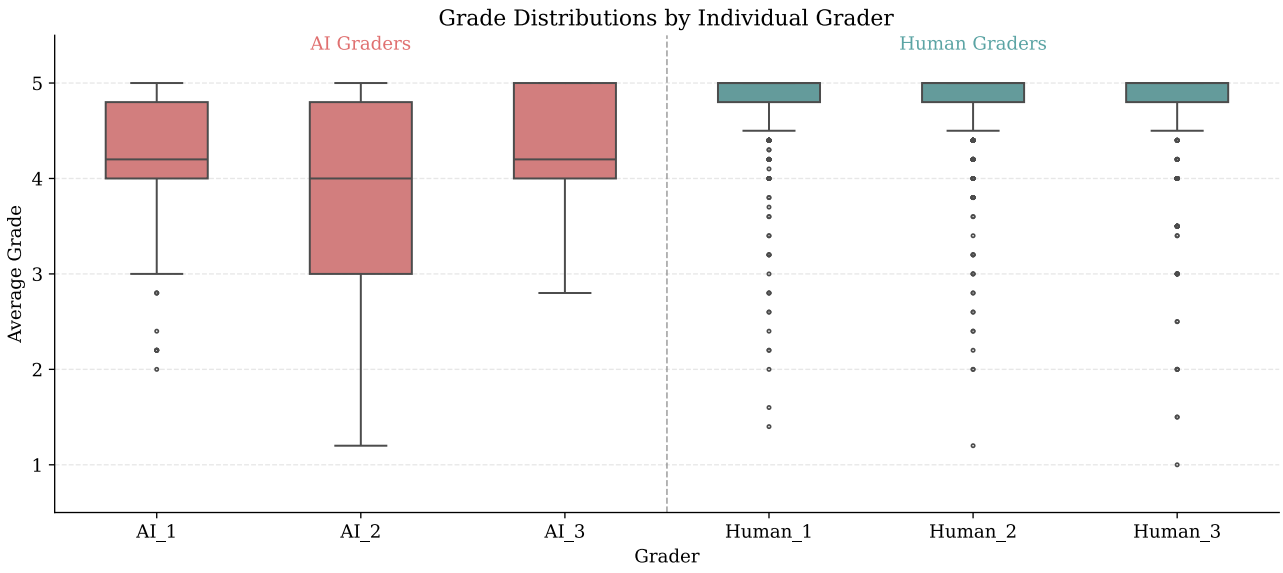


Figure 2: Grade Distributions by Individual Grader. Box plots show the distribution of average grades for each of the six graders (AI_1, AI_2, AI_3; $Human_1$, $Human_2$, $Human_3$). Human graders (right group) cluster at the ceiling with narrow interquartile ranges. AI graders (left group) use the scale more broadly; AI_2 is somewhat harsher than AI_1 and AI_3 but all three remain well below human grader levels.

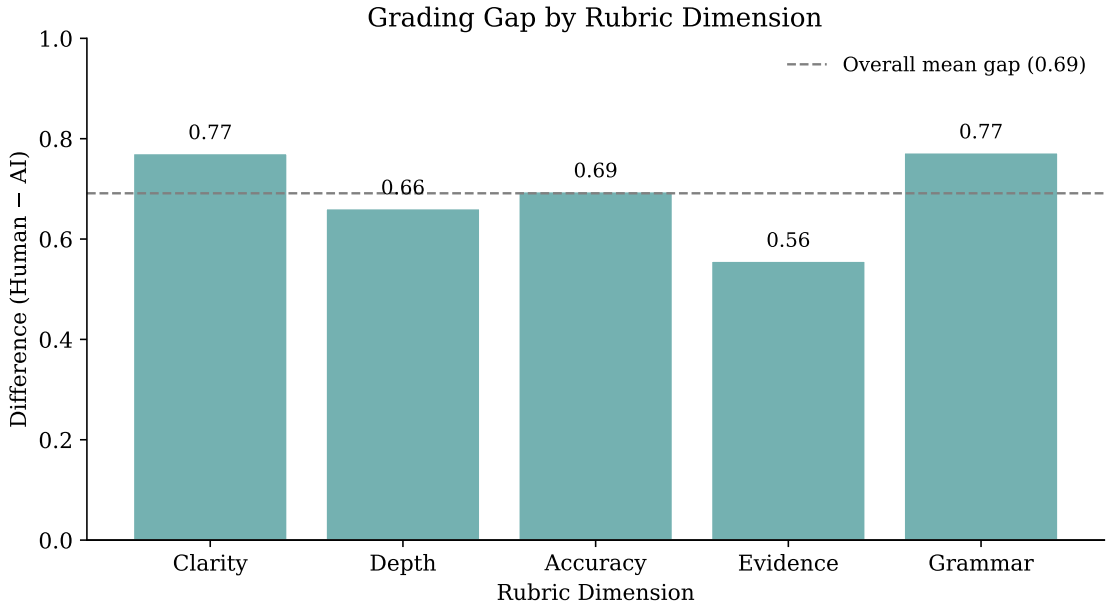


Figure 3: Grading Gap (Human − AI) by Rubric Dimension. The human-minus-AI difference is positive and similar in magnitude across all five rubric dimensions (Clarity, Depth, Accuracy, Evidence, Grammar). The dashed line marks the overall mean gap (0.68 points). The uniform pattern indicates the gap reflects a global calibration difference, not a domain-specific judgment.

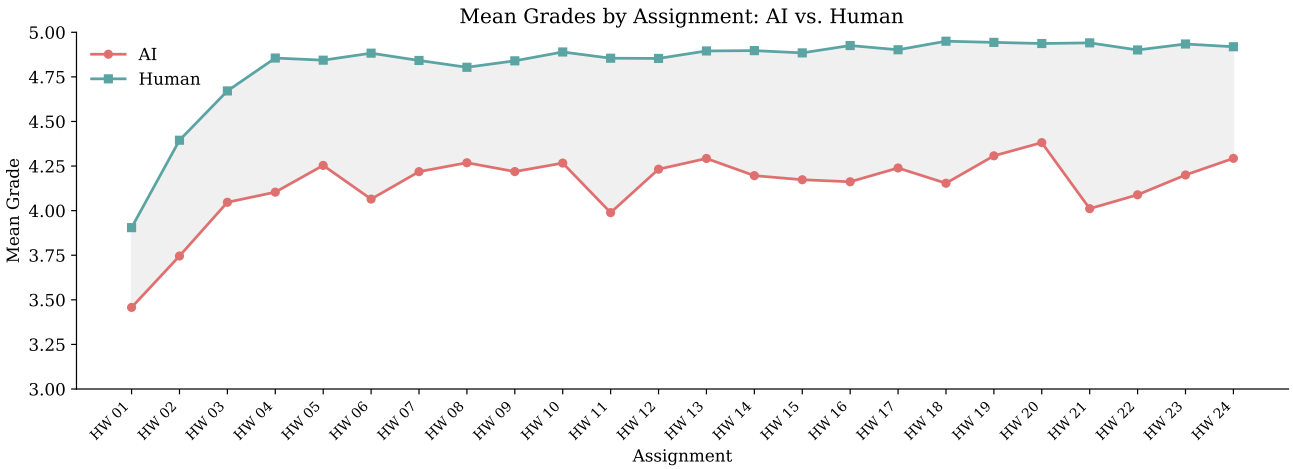


Figure 4: Mean Grades by Assignment Across the Semester. Assignment-level mean grades for AI and human grader pools (hw_01 through hw_24). The shaded region marks the gap between the two pools. The gap is persistent throughout the semester with no sign of convergence, ruling out learning or calibration drift as explanations.

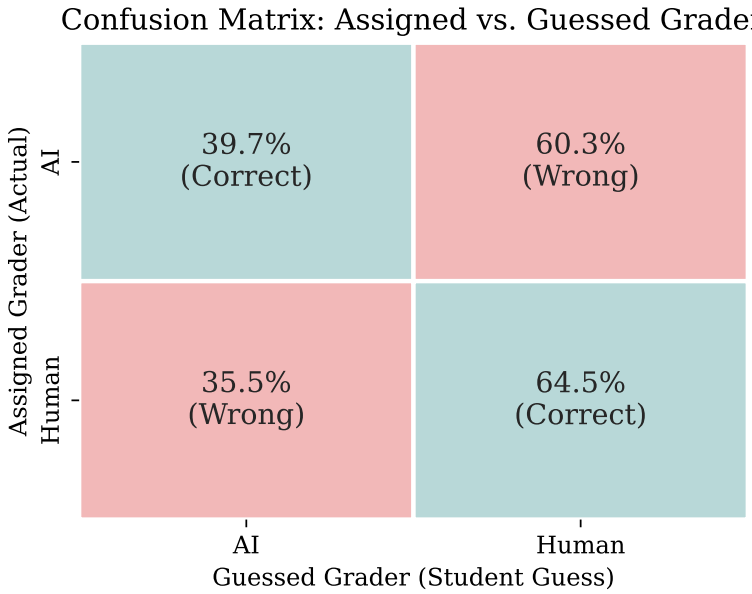


Figure 5: Confusion Matrix: Assigned Grader vs. Student Guess. Cells show the row-wise percentage of guesses within each assigned-grader category. Correct guesses (diagonal) are shaded teal; incorrect guesses are shaded red. Students assigned to AI graders guessed "AI" only 39.7% of the time (60.3% wrongly guessed "Human"), while students assigned to human graders guessed "Human" 64.5% of the time (35.5% wrongly guessed "AI"). The pattern reveals a systematic lean toward "Human" rather than uninformative random guessing

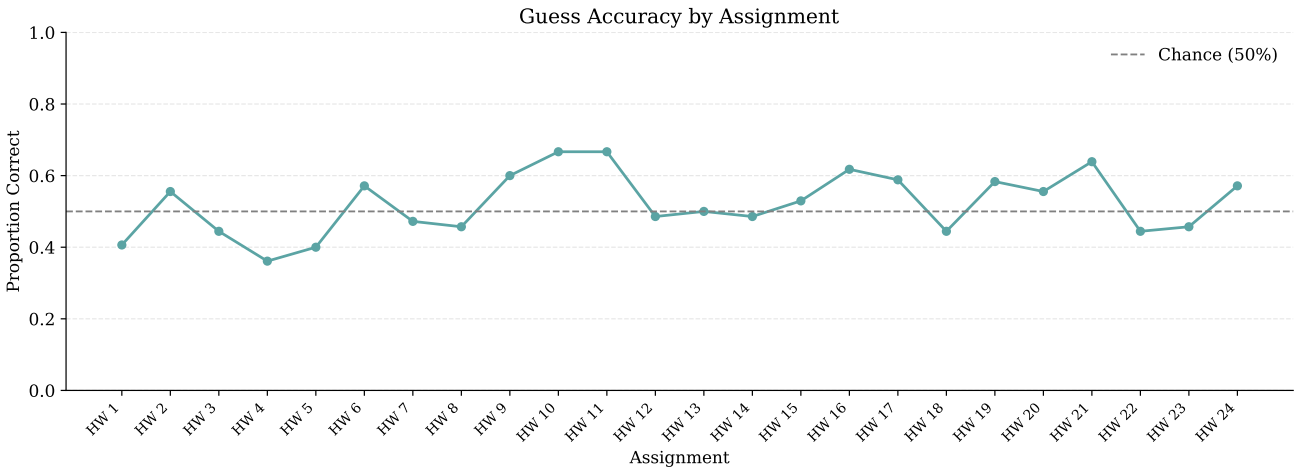


Figure 6: Guess Accuracy by Assignment Across the Semester. Each point shows the pro-portion of students who correctly identified their grader type for a given assignment. The dashed line marks the 50% chance level. Accuracy fluctuates around chance throughout the semester with no upward trend, ruling out learning effects as an explanation for the pooled detection result.

Appendix: AI Grading System Specification

This appendix documents the technical implementation of the AI grading system used in the experiment. Transparency in AI system design is essential for replicability and for understanding the scope of the findings. To assess whether grading patterns are robust across different AI architectures, the three AI graders were implemented using distinct foundation models while maintaining identical parameter configurations, system prompts, and access to course materials. The same system prompt was used identically across all three AI graders.

Model Selection and Configuration

The three AI graders were configured using the following foundation models:

- AI Grader 1: OpenAI GPT-4o
- AI Grader 2: Anthropic Claude Sonnet-4-20250514
- AI Grader 3: Meta Llama-4-Maverick

All three models were configured with identical parameter settings:

- Temperature: 0.5
- Top-p: 0.95
- Frequency penalty: 0.0

Temperature was set to 0.5 to balance grading consistency with sufficient stochasticity to allow meaningful variation in evaluation. This moderate temperature prioritizes reliable application of the grading rubric while avoiding fully deterministic outputs that could artificially inflate inter-rater agreement. Top-p sampling at 0.95 ensures diverse but coherent response generation. Frequency penalty was set to zero to avoid penalizing legitimate repetition of rubric-specific terminology.

Grading Protocol

Submissions were graded in batch rather than sequentially, mirroring the workflow of human teaching assistants who typically evaluate all submissions for a given assignment at once. Each batch consisted of all student submissions for a single assignment. To prevent order effects from influencing grading outcomes, the order in which submissions were presented to each AI model was randomized independently for every grading session. Each submission was evaluated in isolation: the AI model had access only to the individual submission and the RAG knowledge base, with no access to other students' submissions or prior grading outputs within the same batch.

Retrieval-Augmented Generation Architecture

Each AI grader was implemented with a retrieval-augmented generation (RAG) system providing access to the course's knowledge base. The knowledge base included assigned readings and instructor-provided reference materials. When evaluating each submission, the AI system

retrieved relevant context from the knowledge base to ground assessments in course-specific content and standards. This architecture ensures that AI grading reflects the specific learning objectives and material emphasized in the course rather than relying solely on the model's pre-training.