Revision-Loop Behavior and Learning Outcomes under Voluntary AI Formative Feedback in an Undergraduate Statistics Course

Lifeng Han[1]

[1]Department of Mathematics, Tulane University, New Orleans, LA 70118, USA

Abstract

We report a single-section pilot (n=63) of an instructor-built AI grader-and-tutor that lets students upload draft worksheets and receive itemized rubric feedback before submission to a teaching assistant. Across one semester, the system logged 913 substantive conversations, roughly half of the student-initiated ones continued past the first AI response. Cross-sectional regressions of exam outcomes on usage measures yield no statistically significant associations after controlling for prior performance — consistent with strong selection into voluntary opt-in. A within-student panel on per-week recitation worksheet scores tells a different story: in weeks a student opts in, their recitation score rises by 0.90 points out of 10 relative to their own typical week, and each additional revision-loop conversation that week adds a further 0.68 points. Clustering reveals three behavioral phenotypes — light, skimmer, and iterator — not captured by usage counts alone: high-volume skimmers gain little, whereas iterators, who revise most despite a lower baseline, are the only group to outperform what their baseline predicts. We argue that pedagogical evaluation of automated feedback should look past usage volume to behavioral telemetry of revision — telemetry passively collected by any platform that supports re-submission.

Keywords

automated feedback, formative assessment, large language models, statistics education, behavioral telemetry, evaluation methodology

1. Introduction

Large-language-model (LLM) feedback tools are increasingly deployed in higher education [1], and a common framing contrasts "students who use the tool" with "students who do not" [2]. In voluntary deployments with high adoption, this contrast collapses: in our pilot, 97% of enrolled students logged in at least once, leaving little between-student variation in any binary use indicator. A more pedagogically meaningful question is *how* a student uses the tool — whether the conversation closes after a first AI response, or whether the student revises their draft and re-uploads, treating the system as a formative-feedback loop rather than a one-shot oracle.

This paper analyzes one semester of usage logs from an instructor-built AI grader-and-tutor deployed in a 63-student undergraduate statistics course. The platform records timestamps, message roles, and PDF re-uploads at the granularity of every conversation. We use these logs to construct per-student usage *patterns* — volume, follow-up depth, revision-loop share — and ask whether they predict course outcomes (recitation worksheets, midterm and final exams).

Our contribution is methodological rather than an estimate of treatment effect: working retrospectively from observational usage logs, we show that within-student week-to-week variation in tool use — which the platform records passively — yields a sharper signal than between-student volume comparisons, which are dominated by selection.

2. System and Dataset

Tool. The grader is a custom web application built and hosted by the instructor, deployed at the start of the spring semester of 2026 and used continuously for thirteen weeks. A student uploads a PDF draft of

PEAF 2026: Workshop on Pedagogical Evaluation of Automated Feedback

lhan1@tulane.edu (L. Han)

a course worksheet; the application calls Anthropic's `claude-opus-4-5` with an instructor-authored system prompt and returns an itemized rubric report; the student may continue the conversation as chat, most often by re-uploading a revised draft. Three properties distinguish this setup from a student pasting a worksheet into a public LLM chat product. *(i) Course-specific scaffolding.* The instructor-curated system prompt encodes the course's rubric, terminology, and solution conventions. *(ii) Persistent scaffold.* The same prompt is held constant on every turn, so follow-up chats stay anchored to course-appropriate feedback rather than drifting into generic tutoring. *(iii) First-party logs.* All conversations are persisted on the instructor's server, which supplies the telemetry analyzed below.

Cohort and gradebook. The cohort is 63 enrolled students in a single section of an undergraduate one-semester course in probability and statistics. Outcome data is the Canvas gradebook: thirteen weekly recitation worksheets (problem sets completed in small discussion sections led by a teaching assistant (TA)), two midterms, and a final exam. The recitation worksheet is the assignment students opt to pre-grade with the AI before submission to the TA for grading — the natural target for a per-week panel design. This work was determined to be exempt from full IRB review by the Tulane University Human Research Protection Office under 45 CFR §46.104(d)(1) (study ID 2026-724).

Conversation log. The production data snapshot contains 913 substantive conversations covering two distinct assignment types. *Student-initiated (opt-in)* conversations — 473 in total — target the weekly *recitation worksheet*: a student uploads a draft on their own initiative to receive formative AI feedback before submitting the worksheet to the TA for grading, may re-upload revisions, and is never required to respond. *Instructor-initiated (batch)* conversations — the remaining 440 — target the *in-class worksheet*, a separate assignment completed during the lecture session; after a lecture session the instructor runs a batch script that uploads the scanned in-class worksheets and obtains AI grading on the student's behalf, without the student present. The 1.20% follow-up rate quoted below is the share of those 440 batch conversations in which the student later sent any message, used here as a near-zero baseline for what tool use without student initiative looks like.

Of the 63 enrolled students, 42 (67%) opted in at least once; the mean per-student opt-in count is 7.5 (median 8, range 0–28). 49.05% of opt-in conversations continued past the first AI response, versus 1.20% of instructor-batch runs — consistent with opt-in being the formative-feedback channel — and 30.02% include at least two student PDF uploads, a canonical revision loop.

3. Method

Usage metrics. For each enrolled student we compute, from opt-in conversation logs only: opt-in count (n_{optin}), follow-up rate (share of conversations with ≥ 1 message after the initial AI response), revision-loop share (proportion of conversations with ≥ 2 student PDF uploads), and mean messages per conversation.

Outcomes. Per-week recitation worksheet score, Midterm 2, Final Exam, and a composite exam average (mean percent across Midterm 1, Midterm 2, and Final). Midterm 1 is reserved as a baseline-ability control, since it is given before the cohort has accumulated enough usage to differentiate.

Three complementary analyses. (1) Cross-sectional ordinary least squares (OLS) regression on the analytic sample of 42 students who opted in at least once, regressing each outcome on a usage metric and Midterm 1, with heteroscedasticity-consistent (HC3) standard errors. (2) Within-student panel: for each (student, week) cell we pair the recitation grade with that week's opt-in usage — multiple opt-ins in a week are aggregated to that single cell, so estimates do not double-count students who use the tool repeatedly in one week — and regress on student-demeaned predictors, equivalent to OLS with student fixed effects; standard errors are HC3-robust on the demeaned regression. (3) Behavioral phenotypes: k-means clustering (k=3) on n_{optin}, follow-up rate, revision-loop share, and mean messages

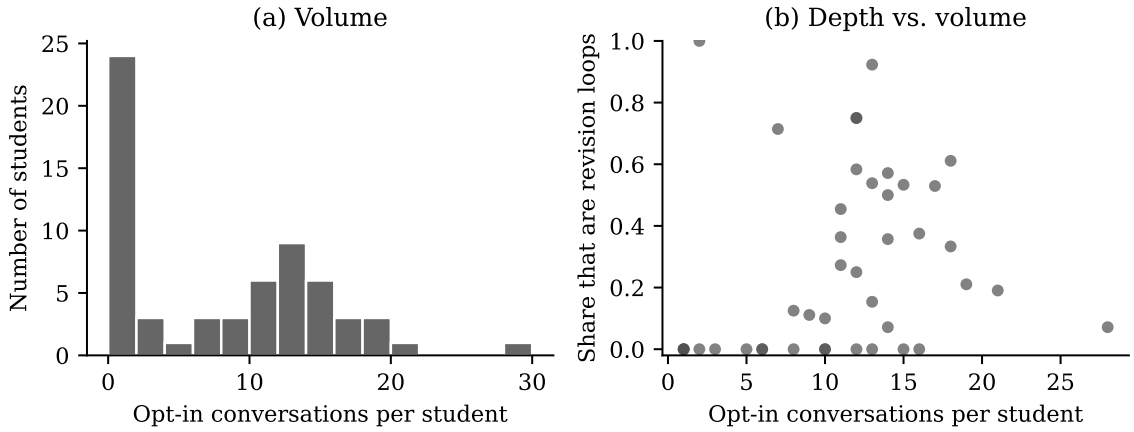

Figure 1: Heterogeneity in voluntary AI use across the cohort (n=63). **(a)** Histogram of opt-in conversations per student. **(b)** Among opt-in users, revision-loop share is not collinear with volume — two students with comparable opt-in counts can have radically different depth of engagement.

per conversation (all z-scored before clustering); cluster names are assigned post hoc by inspection of the centroid profile and introduced in the Results.

Identification. The cross-sectional design cannot identify a causal effect because Midterm 1 only partially absorbs selection on unobserved conscientiousness. The within-student panel absorbs that selection via fixed effects but still requires that within-student week-to-week usage be uncorrelated with within-student weekly performance shocks.

4. Results

Usage is heterogeneous. Figure 1 shows the distribution of opt-in count and the joint distribution of opt-in volume and revision-loop share. Volume is approximately bimodal — a cluster of low-volume users and a long upper tail — and depth is only loosely coupled to volume: there are high-volume students with a near-zero revision-loop share, and lower-volume students whose every conversation is a revision loop.

Opt-in students start out stronger. Consistent with strong selection into voluntary opt-in, the 42 opt-in students entered the course with substantially higher Midterm 1 scores than the 21 non-opt-in students ($76.7\% \pm 16.8$ vs $62.0\% \pm 17.4$, Welch's t=3.20, p=0.003), and remained higher on Midterm 2 ($86.3\% \pm 10.8$ vs $78.7\% \pm 14.2$, p=0.038) and the Final Exam ($75.7\% \pm 14.9$ vs $63.1\% \pm 13.1$, p=0.001). The Midterm 1 gap is present before most opt-in usage has accumulated, so it indexes pre-existing ability rather than a tool effect.

No cross-sectional association. Table 1 shows OLS coefficients on Final Exam (out of 108) regressed on a usage measure plus Midterm 1. No usage variable is statistically significant; Midterm 1 absorbs essentially all explainable variation. The same null pattern holds for Midterm 2 (not shown).

Within-student panel signal. The same usage measures, evaluated within-student week-by-week against the relevant TA-graded recitation worksheet, yield positive and significant effects (Table 2). Weeks in which a student opts in are associated with recitation scores 0.90 points higher out of 10 than weeks they do not (p=0.005). Each additional revision-loop conversation that week adds 0.68 points (p=0.012), and the indicator for any revision loop adds 0.76 points (p=0.013). The within-R^2 values are small (≤ 0.012), as expected — most variance in weekly recitation grades is idiosyncratic

Table 1
Cross-sectional OLS on Final Exam (out of 108) with Midterm 1 control. $n = 42$ opt-in students. Coefficients are unstandardised; standard errors are HC3.

Predictor	Coefficient (SE)	p
n_{optin} (alone)	$0.20\ (0.32)$	0.54
follow-up rate (alone)	$-1.96\ (6.40)$	0.76
revision-loop share (alone)	$3.15\ (6.59)$	0.63
n_{optin} + rev. share	$0.17\ (0.31)$ & $2.52\ (6.13)$	0.57, 0.68
Midterm 1 (control, always included)	$0.45\ (0.16)$	0.004

Table 2
Within-student panel OLS on TA-graded recitation worksheet score (out of 10), 788 student-week pairs from 61 students, student fixed effects via demeaning, HC3 SEs. Each row is a separate regression.

Predictor (within-student)	Coefficient (SE)	p
n_{optin} that week	$0.13\ (0.20)$	0.52
any opt-in that week (1/0)	$0.90\ (0.32)$	0.005
revision-loop count that week	$0.68\ (0.27)$	0.012
any revision loop that week (1/0)	$0.76\ (0.31)$	0.013
follow-up message count that week	$0.45\ (0.21)$	0.032

— but the effects are stable across specifications. The contrast between Tables 1 and 2 is the central methodological message of this paper: volume between students is too contaminated by selection to be informative at this sample size; volume within a student is a sharper signal.

Behavioral phenotypes. The $k{=}3$ clustering yields three centroid profiles that we name descriptively after inspection: a low-volume, near-zero-follow-up cluster (*light*, mean 3.1 opt-ins, near-zero follow-ups); a high-volume cluster whose conversations rarely continue past the first AI response (*skimmer*, mean 14.0 opt-ins but only 36% follow-up rate and 12% revision share); and a comparably high-volume cluster whose conversations are mostly multi-turn revision dialogues (*iterator*, mean 12.2 opt-ins, 75% follow-up rate, 60% revision share). The cohort baseline differs across phenotypes (mean Midterm 1: iterator 71.3, light 72.5, skimmer 82.7), but raw final-exam scores do not (Figure 2; $F{=}0.16$, $p{=}0.85$). Adjusting for Midterm 1 changes the ordering: regressing the exam-average composite (mean percent across Midterm 1, Midterm 2, and the Final Exam) on Midterm 1 alone, and averaging the residuals within each phenotype, yields $+0.9$ points for iterators, -0.2 for light users, and -0.6 for skimmers. Iterators are the only group with a positive residual — they were not the strongest students to begin with, but they are the only ones modestly outperforming what their baseline would predict. Two design choices underpin this comparison: the clustering features are aggregated over the full thirteen-week deployment, so a phenotype reflects whole-semester usage style rather than any single week, and Midterm 1 again serves as the pre-tool baseline.

5. Discussion

The findings bear on three questions about evaluating and deploying automated feedback:

- *Evaluation methods:* most evaluations of automated feedback tools rely on student surveys or accuracy comparisons against human graders [3], leaving silent the question of whether the student revised; behavioral telemetry of revision is a computable, passively collected metric that indexes the formative-feedback signal those evaluations miss.
- *Impact on teaching and learning:* the within-student panel shows an association: in weeks a student opts in or revises, their TA-graded recitation grade is higher than predicted by their

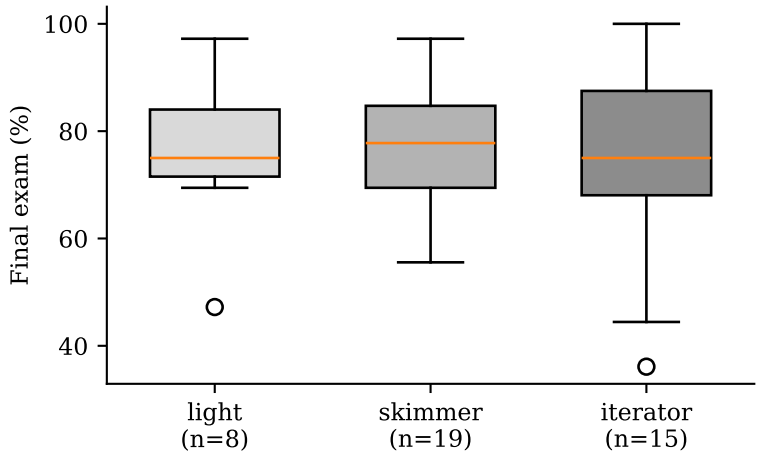

Figure 2: Final exam score by behavioral phenotype. Raw final-exam distributions are statistically indistinguishable across the light (n=8), skimmer (n=19), and iterator (n=15) groups (F=0.16, p=0.85); phenotype differences emerge only after adjusting for Midterm 1 baseline ability.

other-week performance — the pattern formative feedback is intended to produce.

- *Personalization:* the same tool is consumed in qualitatively different ways by the three phenotypes, and skimmer-style consumption is not associated with above-baseline performance even at high volume.

The personalization pattern admits a clean theoretical reading. Chi and Wylie's ICAP framework [4] predicts that learning gains rise from *passive* through *active* and *constructive* to *interactive* engagement. Under this mapping, single-shot opt-ins approximate *active* behavior, whereas revision loops approximate *constructive* behavior. The positive Midterm-1-adjusted residual concentrated in the iterator phenotype is consistent with ICAP's prediction that learning gains live above the active-constructive boundary. More broadly, treating passively-logged interaction traces as engagement proxies follows established learning-analytics precedent [5, 6].

Several threats to causal interpretation deserve emphasis. *Selection:* opting in correlates with conscientiousness, consistent with the cross-sectional null. The within-student panel is more credible but still assumes that within-student weekly usage is not jointly determined with weekly performance shocks (illness, competing course load); the direction of any residual bias is ambiguous. *Mechanism:* we cannot separate the AI's feedback quality from the act of revising itself — a student who revises after any feedback might gain — so a clean test would randomize who delivers the draft feedback (AI vs. human). *Rubric alignment:* opt-in students may gain from a "teaching to the rubric" effect rather than substantive understanding; the panel partially mitigates this by using each student as their own baseline but cannot fully separate the two.

The pilot's reach is also narrow: one section, one course, one semester (63 enrolled, 42 opt-in). The model and rubric are held fixed throughout, so we cannot speak to how feedback quality varies with either. The four behavioral features index *whether* a student revised, not the discourse quality of the revision turns; qualitative coding against frameworks for collaborative talk [7, 8] is one follow-up, and a pre-registered multi-section design with as-good-as-random variation in opt-in friction, currently in preparation, is another. We offer this pilot not as a verdict on whether LLM feedback works but as a demonstration that revision telemetry is worth measuring, and a prompt for the design-controlled studies that can test what it means.

Declaration on Generative AI

During the preparation of this work, the author used Claude to paraphrase and reword. The author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al., *ChatGPT for good? On opportunities and challenges of large language models for education*, Learning and Individual Differences 103 (2023) 102274.

[2] D. Smerdon, *AI in essay-based assessment: Student adoption, usage, and performance*, Computers and Education: Artificial Intelligence 7 (2024) 100288.

[3] M. Messer, N. C. C. Brown, M. Kölling, M. Shi, *Automated grading and feedback tools for programming education: A systematic review*, ACM Transactions on Computing Education 24 (1) (2024) Article 10.

[4] M. T. H. Chi, R. Wylie, *The ICAP framework: linking cognitive engagement to active learning outcomes*, Educational Psychologist 49 (4) (2014) 219–243.

[5] G. Siemens, *Learning analytics: the emergence of a discipline*, American Behavioral Scientist 57 (10) (2013) 1380–1400.

[6] C. Romero, S. Ventura, *Educational data mining and learning analytics: an updated survey*, WIREs Data Mining and Knowledge Discovery 10 (3) (2020) e1355.

[7] N. Mercer, *Words and Minds: How We Use Language to Think Together*, Routledge, London, 2000.

[8] R. Wegerif, *Dialogic Education and Technology: Expanding the Space of Learning*, Springer, Boston, 2007.