PEAF 2026 · WORKSHOP ON PEDAGOGICAL EVALUATION OF AUTOMATED FEEDBACK

Look past usage volume — measure the revision loop.

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

Lifeng Han
Department of Mathematics, Tulane University

5-minute lightning talk

THE SETUP

Everyone uses the tool — so “whether” tells us nothing.

An instructor-built AI grader-and-tutor in one statistics section. Students upload a draft worksheet, get itemized rubric feedback, and may revise and re-upload — all **voluntary**.

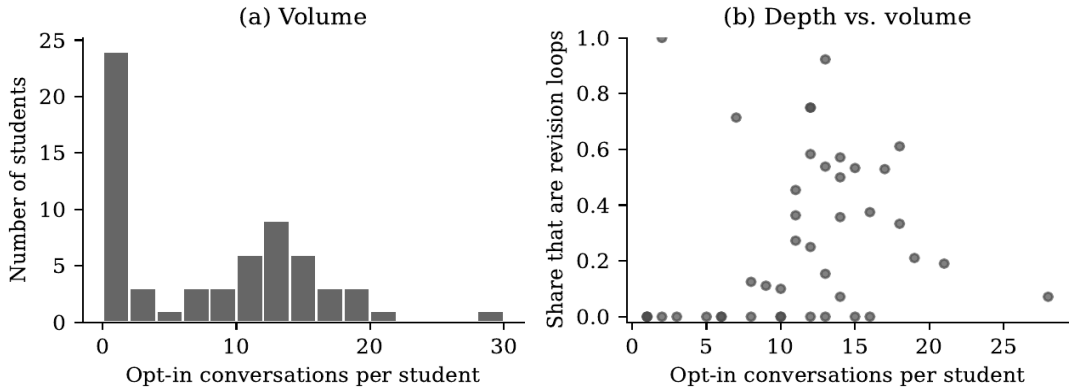
97%
of students logged in at least once

913
substantive conversations logged

49%
of opt-ins continued past the first reply

The real question isn’t whether — it’s *how*.

Does the student treat the AI as a one-shot oracle, or revise their draft and re-upload — a formative-feedback loop?



Volume is bimodal — and depth of engagement is not predicted by volume. Two students with the same number of opt-ins can revise very differently.

n = 63 enrolled

THE CENTRAL RESULT

Volume between students is noise. Revision within a student is signal.

BETWEEN STUDENTS

Cross-sectional OLS · exam ~ usage + Midterm 1

No usage measure is significant

opt-in count	p = 0.54
follow-up rate	p = 0.76
revision-loop share	p = 0.63

Selection dominates: opt-in students simply started stronger (Midterm 1: 77% vs 62%).

WITHIN STUDENT

Week-by-week panel · student fixed effects

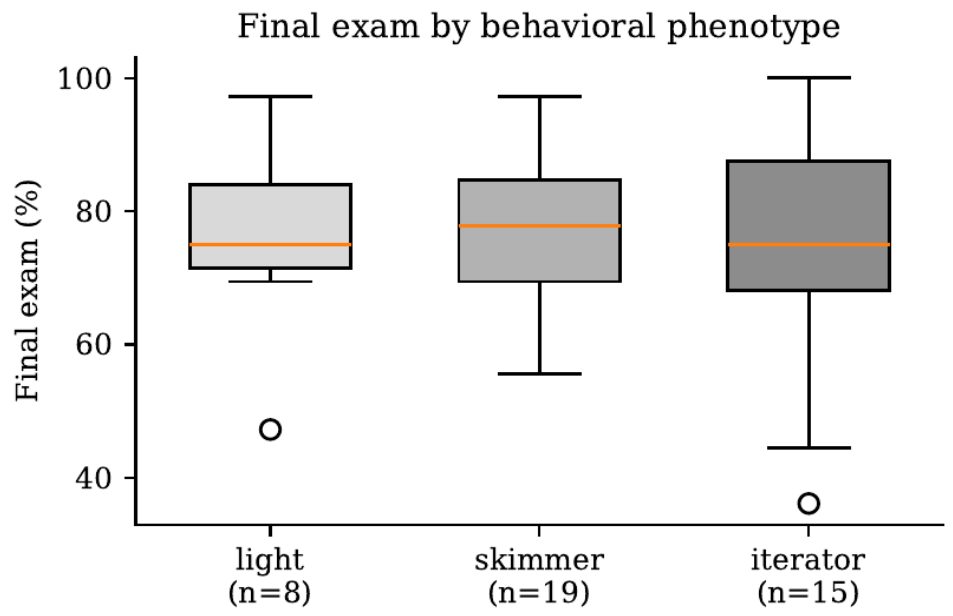
Each student is their own control

A week they opt in	+0.90 pts	p = .005
Each revision loop that week	+0.68 pts	p = .012
Any revision loop (1/0)	+0.76 pts	p = .013

…on the TA-graded recitation worksheet (out of 10). 788 student-week pairs.

WHO ACTUALLY BENEFITS

Same tool, three behavioral phenotypes.



Raw final-exam scores are statistically identical across groups (F = 0.16, p = 0.85).

Light

low volume, rarely follows up

−0.2
vs baseline

Skimmer

high volume, but one-shot — 12% revise

−0.6
vs baseline

Iterator

high volume & multi-turn — 60% revise

+0.9
vs baseline

Iterators were ***not*** the strongest going in — yet they're the **only group that beats its own baseline.**

TAKEAWAY

Evaluate automated feedback by revision, not by volume.

1. **Volume comparisons are dominated by selection.** Within-student revision behavior is the sharper, fairer signal.

2. **Revision telemetry is free.** Any platform that supports re-submission already logs it — no surveys, no extra instrumentation.

3. **It maps onto theory.** Revision loops ≈ ICAP "constructive" engagement — exactly where learning gains are predicted to live.