Metadata

Title
The Effect of Providing Joint Feedback to Students and Educators to Facilitate Active Learning

Description
This project builds on previous studies conducted in 2021 on Code in Place, an online programming course where we found that automated feedback to instructors can improve their instruction and student satisfaction. The current study was conducted in the spring of 2024 on Code in Place, and its goal is to understand whether providing feedback to students as well, in addition to providing feedback to instructors, influences the quality of their discourse and student outcomes. Feedback to teachers is known to be an effective way to improve their instruction, but few empirical studies have examined the effect of joint feedback to both teachers and students. To answer this question, the study leverages a randomized controlled trial design and computational natural language processing techniques.

Contributors [alphabetical by last name]
Demszky, Dora
Dennison, Deepak
Piech, Chris
Sun, Olivia
Yao, Sarah
Yun, Joy

Category
Project

Subject
Education
Natural Language Processing
Social and Behavioral Sciences

Study information

Research Questions:
List each research question included in this study. When specifying your research questions, it is good practice to use only two new concepts per research question. For example, split up your

questions into a simple format: 'Does X lead to Y?' and 'Is the relationship between X and Y moderated by Z?'. By splitting up the research questions here, you can more easily describe the statistical test for each research question later.

RQ1: Does joint feedback to both students and instructors improve *student outcomes* (attendance, engagement in section, assignment completion) above and beyond feedback provided to instructors alone?
RQ1.1.: Does giving students the option to *also give feedback back* to the instructor further improve these metrics?

RQ2: Does joint feedback to both students and instructors improve *instructor outcomes* (talk time, talk move rate, NPS, instructor engagement with the course) above and beyond feedback provided to instructors alone?
RQ2.1.: Does giving students the option to *also give feedback back* to the instructor further improve these metrics?

RQ3: How do treatment effects vary by instructor and student characteristics (demographics=location,age,gender, baseline engagement before feedback, etc.)?

Exploratory questions:

- How do treatment effects change over time?
- How do treatment effects differ for students whose section membership changed during the course, relative to students who stayed in their assigned section?
- Are there spillover effects on control-group students who sat in sections that also contained treated students?
- How do treatment effects vary across the three talk-move themes of the curriculum (getting ideas on the table, building on ideas, connecting ideas)?

Hypotheses*
For each of the research questions listed in the previous section, provide one or more specific and testable hypotheses. Please make clear whether the hypotheses are directional (e.g., A > B) or non-directional (e.g., A ≠ B). If directional, state the direction. You may also provide a rationale for each hypothesis.

- H1 (directional). Students who receive automated feedback in addition to their instructor receiving feedback show higher attendance, higher in-section engagement (talk time, number of questions asked) and higher assignment completion than control students: StudentFeedback > Control and StudentSLFeedback > Control. Rationale: the student feedback makes each student's own participation visible to them and pairs it with an explicit reminder about the value of asking questions, which is the behavior we measure.

    - H1.1 (directional). Giving students the option to send feedback back to their instructor further improves these outcomes: StudentSLFeedback > StudentFeedback. We expect this incremental effect to be smaller than the effect of student feedback itself, since the added component asks students for effort rather than giving them information about themselves.
- H2 (directional). Joint feedback improves instructor outcomes — engagement with their own feedback (viewing it, time spent), instructor talk time (lower), talk move rates (higher) and the proportion of students engaged (higher) — above and beyond instructor-only feedback. Rationale: instructors in all arms receive the same feedback, so any effect on instructors must operate indirectly, through changes in how their students behave in section and, in the StudentSLFeedback arm, through the student-authored feedback that appears on the instructor's feedback page. We therefore expect instructor-level effects to be smaller than student-level effects. Because most sections contained students from more than one arm (see Prior knowledge), instructor-level estimates are additionally attenuated toward zero and should be read as a lower bound.
    - H2.1 (directional). StudentSLFeedback > StudentFeedback on instructor outcomes. We expect the contrast to be clearer here than for students (H1.1), because this is the only arm in which the instructor receives anything the other arms do not.
- H3 (non-directional). Treatment effects vary by instructor and student characteristics. We make no directional prediction for gender, age, location or returning-instructor status. For baseline engagement we expect a directional pattern: effects are larger for students and instructors with lower first-week participation, who have more room to change.

Data description

Datasets used*

Name and briefly describe the dataset(s), and if applicable, the subsets of the data you plan to use.Useful information to include here is the type of data (e.g., cross-sectional or longitudinal), the general content of the questions, and some details about the respondents. In the case of longitudinal data, information about the survey's waves is useful as well. Mention the most relevant information so that readers do not have to search for the information themselves.

- Instructor demographic information: gender, age, location (country), is returning SL
- Student demographic information: gender, age, location (country)
- Transcripts derived from session recordings
- Automated measures of instructional practice per class recording
    - Talk time
    - Proportion of students engaged
    - Eliciting student ideas
    - Building on student ideas
    - Orienting students to one another
    - Number of learner questions

- View logs of automated feedback page
    - Whether they accessed the feedback
    - Time spent on feedback
- In-platform user input *
    - SL Reflection (binary; content)
    - Num. student personal + SL-facing reflections per session, sent post session
    - For each student, their personal reflection, and their feedback to the SL
- Student attendance data

Randomization at the student level (based on original section assignments). Student experience was consistent throughout – whether they received feedback, and whether they could share feedback with SL was same throughout for a given student, even if they attended a different section.

-

- Student assignment completion data *
- Instructor and student access logs for different resources on the CiP platform
    - Handouts
    - IDE
    - Training modules
- Exit survey responses from instructors and students

Data collection procedures*

If the data collection procedure is well documented, provide a link to that information. If the data collection procedure is not well documented, describe, to the best of your ability, how data were collected. Describe the representativeness of the sample and any possible biases stemming from the data collection.

*You may **attach up to 5 file(s)** to this question. Files cannot total over 5GB in size. Uploaded files will automatically be archived in this registration. They will also be added to a related project that will be created for this registration.*

Sample size: 564 instructors and 6759 students at the time of randomization

In the final data pull, the platform records 7,174 students and 611 instructors across 612 sections; 7,055 students (98.4%) carry an experimental arm assignment (Control n = 2,363; StudentFeedback n = 2,311; StudentSLFeedback n = 2,380). The randomization file itself assigns 7,604 students across 606 pre-course section rosters (Control n = 2,539; StudentFeedback n = 2,498; StudentSLFeedback n = 2,567). The three counts differ because enrollment and section rosters continued to change between randomization and the first section, and because some enrolled students never appear in the course logs. Confirmatory analyses are restricted to students who carry an arm assignment in the final data; we will report the reconciliation between the randomization file and the final data pull, and confirm that attrition between the two is not differential by arm.

Instructors in the final sample are on average 29.3 years old, 30.6% are female, 42.2% are located in the United States, and 25.7% are returning section leaders. Students are on average 32.1 years old, 46.9% are female and 30.1% are located in the United States.
The study was conducted in a free, online 6 week long introductory programming course called Code in Place. Anyone could apply to serve as a volunteer section leader on the course, and then they were selected by the course organizers. Our participant sample consists of all adult (18+) instructors in Code in Place.

Feedback to Instructors
All instructors received automated feedback throughout the course following this curriculum:

The talk move in focus followed a three-theme curriculum, with each theme spanning two weeks: Weeks 0–1 “Getting Ideas on the Table” (eliciting student ideas), Weeks 2–3 “Building on Ideas” (building on student ideas), Weeks 4–5 “Connecting Ideas” (orienting students to one another). Eliciting was detected with the Code in Place eliciting model; building was detected with the uptake model of Demszky et al. (2021).

Home
Code
Section
Forums
SL Training
Lessons
Connections
About
Events

386 online

Hi, Test User!

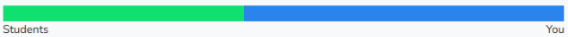
You can view your feedback from your section here. You rock! You've done the most important part of the job: showing up! Here's some feedback to help you improve.

Engagement Overview

Let's begin by looking at some basic numbers from your latest session on active learning. Seeing these numbers go up is a good sign that you're getting better at engaging your students. These stats do not include breakouts so they could be incorrect

Your students talked 43% of the time, and you talked 57% of the time.

The students in your section engaged more than 95% of other sections!



8 of your 10 students engaged at least once.

Theme of the Week

Week 3-4

The theme of the week is Building on Ideas

- Ask a student to clarify or elaborate on their response.
- Follow up on a student comment by pointing out its significance.
- Refer back to a student's idea.

Example: What do you think would happen if...?

Link to Section Leader Training Module

What's coming up next week?

Aside: This feedback is private to you, is for your own benefit, and will not be used for evaluating your performance as a Section Leader. View your full transcript.

Here are some auto-selected highlights from your Zoom transcript when you did a great job **Building on Ideas**.

Student
Yeah, that's a good idea. Actually. Adam. Print.

Section leader
What do you want to print.

Previous Next

Here's some personalized advice from GPT-4 on Building on Ideas. This advice comes from your session transcript and strategies that previous section leaders have found helpful.

During your session, there was a concerted effort to engage with learners' ideas and guide them towards solutions, especially when working on programming problems. However, a recurring opportunity for improvement is directly leveraging the strategy of building upon learners' mistakes to guide them towards the correct solution. For example, in the discussion regarding the planet weight calculation, when navigating through troubleshooting and refining the code, there was a significant amount of trial and error involved. You adeptly navigated the discussion, encouraging participation and sharing of ideas, which is commendable. Yet, there was a chance to more explicitly recognize and validate learners' initial thoughts even when they were incorrect, and then use those as stepping stones towards the correct code structure. A specific instance where this could be applied was when the session touched upon the structure of `if` and `elif` statements and handling of variables (like in the conversation around the implementation of flags and while loops). Embracing these moments to say something like, "You're thinking along the right lines by wanting to loop until we get a valid planet, how can we adjust our approach to ensure the program prompts the user again instead of exiting or entering an infinite loop?" would highlight the learner's contribution before steering towards an improvement. By doing so, you are not just fixing errors but using those errors as natural learning pivot points. In encouraging this development, it's beneficial to recognize the educational value in errors. Each mistake made by a learner and corrected through your guidance is a valuable learning moment. Bringing attention to these moments not only enhances understanding but also fosters a learning environment where it's safe to experiment and learn from mistakes. This technique can significantly enhance the learners' problem-solving skills and confidence in programming.

Your Personal Reflection

What are your reflections from this week's section? We'll show you this next week so you can revisit them.

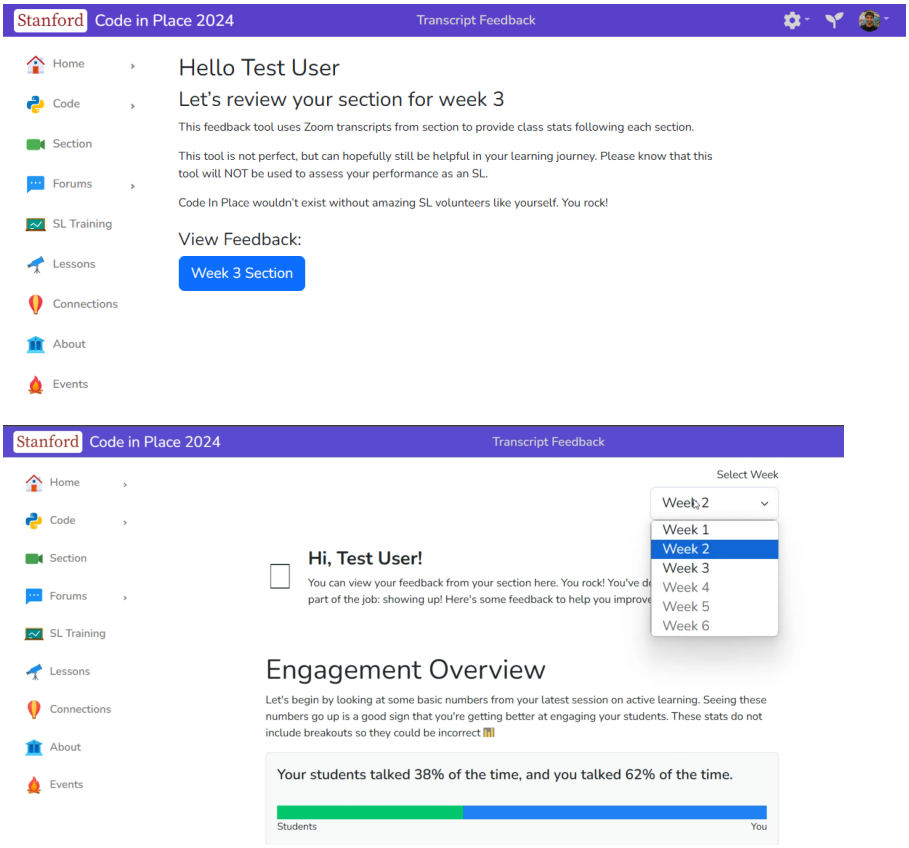
Submit Reflection

Keep up the great work, and stay tuned for next week!

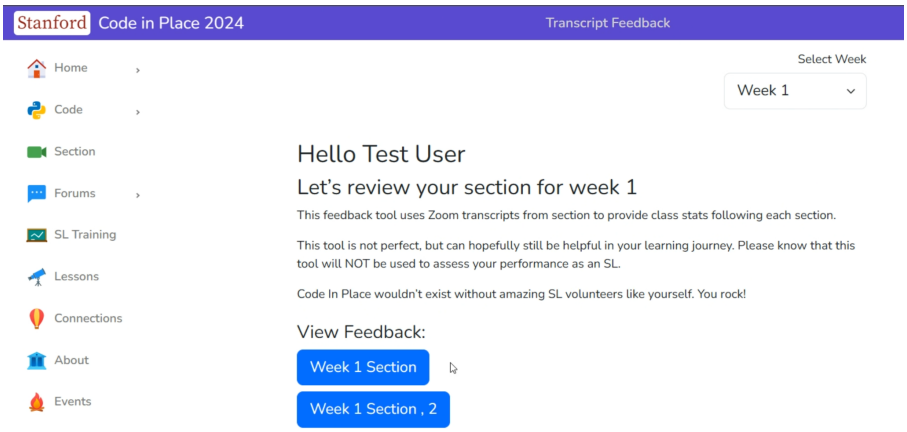
The instructor feedback included several components:

- Introduction to the feedback
- SL talktime percentage + percentile
- Proportion of students engaged
- Description of the talk move in focus
- Link to the relevant training module
- Link to full transcript
- Their talk moves in action (list of talk moves from their transcript)
- Personalized GPT-4 feedback for given talk move
- Opportunity to write a reflection
- Reflection opportunities

Sections took place between Wednesday-Friday of each week. Each Monday starting in the second week, instructors could view the feedback from their prior week by clicking on the corresponding item on their “roadmap” (i.e. checklist before their next section).



If an instructor taught multiple sections (e.g. they served as a substitute for one), they could view feedback from each of those sections.



Randomized Study Setup

We randomized participants at the student level, implemented as a cluster randomization over pre-course section rosters. Concretely, the list of 606 pre-course section rosters was shuffled and split into three equal parts, and every student on a given roster was assigned to that roster's condition. Treatment then travelled with the student for the rest of the course: a student's condition never changed, including in the weeks when they attended a section other than their own.

Two consequences follow for the analysis. First, the unit of randomization is the pre-course section roster, so confirmatory standard errors are clustered at that level rather than at the level of the section a student happened to attend. Second, because rosters were revised between randomization and the first section, and because students moved between sections during the course, sections are not pure with respect to condition (see Prior knowledge below).

Before the course began: We randomized students once they were accepted to the course, before the first section. Third of the students (n=2188 name=StudentFeedback group) were assigned to receive automated feedback in addition to their instructors receiving feedback. Another third of the students (n=2187, name=StudentSLFeedback group) both received feedback, similar to the StudentFeedback group, and also had the option to share feedback back to their instructor. Control group students did not receive feedback.

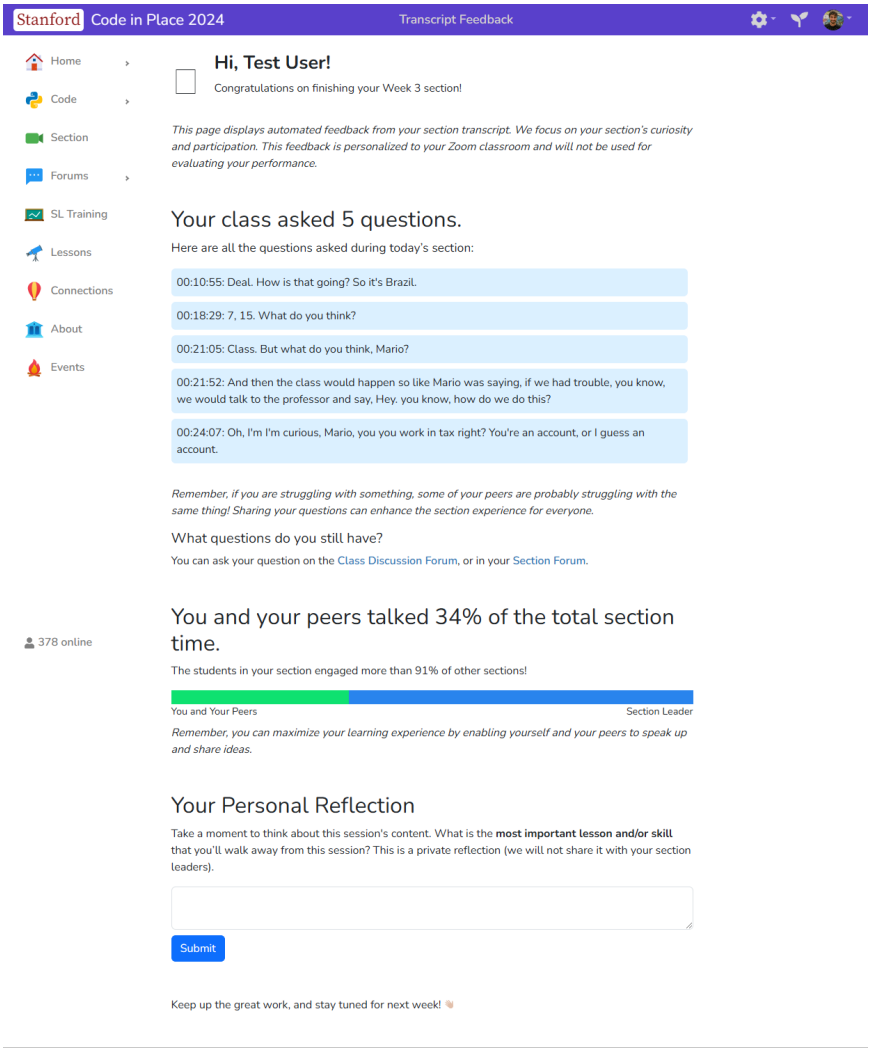
Control-group students were not given personalized feedback. They saw a static acknowledgement message (“thank you for showing up”) in the same place in the roadmap where treated students saw their personalized feedback, so the contrast between arms is the content of the feedback rather than the presence of a roadmap item.

Once the course began: On the Monday following each section, students were able to access the automated feedback within their roadmap (checklist before their next section).

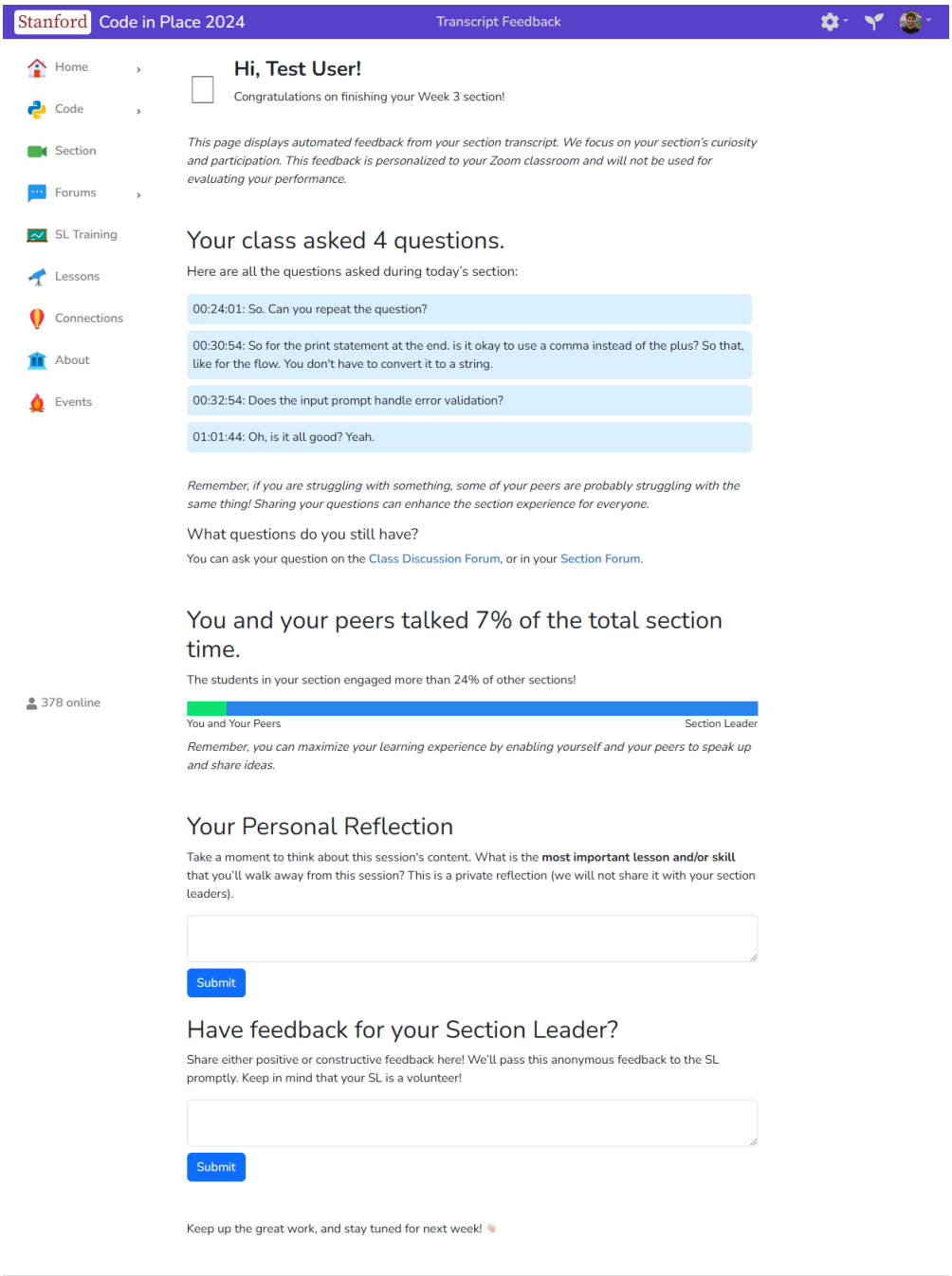
The feedback for the StudentFeedback group looked had the following components:

- Introduction to the feedback
- Number of student questions asked
- List of student questions asked with timestamps
- Reminder about the importance of asking questions
- Link to class and section forum to ask further questions
- Student talktime percentage + percentile
- Opportunity to write a reflection

And looked like this:

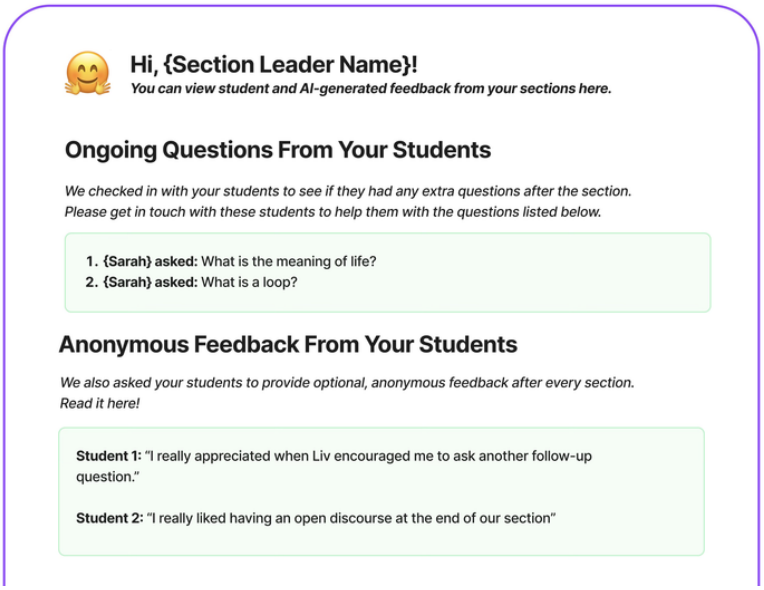


The feedback for the StudentSLFeedback group was similar, with one additional component: students could provide feedback to their SLs



As soon as students submitted feedback, they also appeared inside the instructor feedback page:

Updated SL Feedback View
(V1: Feedback from Students)



If a student did not attend their section, they only saw a reflection box as part of their feedback.

At the end of the course, instructors and students were sent a quick exit survey to complete, which included an NPS rating as well as questions about section quality.

Variables

Manipulated variables

If you are going to use any manipulated variables from the study variables, identify them here. Describe the variables and the levels or treatment arms of each variable. Note that this is not applicable for observational studies and meta-analyses. If you are collapsing groups across variables this should be explicitly stated, including the relevant formula. If your further analysis is contingent on a manipulation check, describe your decisions rules here.

You may attach up to 5 file(s) to this question. Files cannot total over 5GB in size. Uploaded files will automatically be archived in this registration. They will also be added to a related project that will be created for this registration.

The manipulated variable is the student's feedback condition, with three levels assigned by randomization:

- Control (n = 2,363): the student's instructor received automated feedback; the student received only a static acknowledgement message.

- StudentFeedback (n = 2,311): the student's instructor received automated feedback and the student received personalized automated feedback on their own participation (questions asked, talk time, percentile, reflection prompt).
- StudentSLFeedback (n = 2,380): as StudentFeedback, plus the option to write feedback to their instructor, which then appeared on the instructor's feedback page.

Instructor feedback is held constant: every instructor received the same automated feedback in every arm, so the study has no instructor-only manipulation. In regressions we code the condition as two indicators, StudentFeedback and StudentSLFeedback, with Control as the omitted category. For the pooled contrast against control (H1, H2) we also define AnyStudentFeedback = StudentFeedback + StudentSLFeedback. We do not condition any analysis on a manipulation check, but we will report, as a descriptive check, the share of students in each arm who opened their feedback at least once.

Measured variables*

Describe both outcome measures as well as predictors and covariates and label them accordingly. If you are using a scale or an index, state the construct the scale/index represents, which items the scale/index will consist of, and how these items will be aggregated. When the aggregation is based on exploratory factor analysis (EFA) or confirmatory factor analysis (CFA),also specify the relevant details (EFA: rotation, how the number of factors will be determined, how best fit will be selected, CFA: how loadings will be specified, how fit will be assessed, which residuals variance terms will be correlated). If you are using any categorical variables, state how you will code them in the statistical analyses.

Covariates:

- Instructor demographics
    - Age
    - Is female (binary)
    - In US (binary)
    - Returning section leader (binary)
        - Binary indicator indicating if the instructor is a returning section leader in Code in Place.
- Student demographics
    - Age
    - Is female (binary)
    - In US (binary)
- Week (1 to 6; only used for transcript-level analyses)
- Student attendance of first session
    - Is not affected by the treatment since there was no communication between the instructor and students prior to first session
- Discourse characteristics of the first session
    - Talk time, rate of talk moves

Outcome(s):

- **RQ1: treatment effects on student outcomes**
  - number of attended sessions between weeks 2-6
  - student talk time in their sessions
  - number of student questions in their sessions
  - proportion of assignment completed for each week (overall proportions listed below)
    - Overall completion proportions in the final data (n = 7,151 students): assignment 0 = 92.5%, assignment 1 = 42.2%, assignment 2 = 62.1%, assignment 3 = 47.1%, assignment 4 = 25.9%, assignment 5 = 21.6%. Assignment 0 is used as a baseline covariate rather than an outcome, since it precedes any feedback.

      ```
      assignment 1: 0.6605952909817858
      assignment 2: 0.07587738782763218
      assignment 3: 0.048630238412557376
      assignment 4: 0.027642035638481666
      assignment 5: 0.01988005330964016
      assignment 6: 0.007737301939878573
      ```

    -
  - exploratory
    - survey responses (NPS, quality of instruction)
      - the response rate is only 12.6% (886 of 7,055 arm-assigned students), so we may not be able to rely on this data that much
    - engagement with in-platform things, e.g. IDE
- **RQ2: treatment effects on instructor outcomes**
  - engagement with feedback
    - **viewing their feedback before their subsequent section (binary)**
      - At any point during the course
      - By week
    - **time spent on feedback (in seconds)**
  - impact on section practices
    - talk time
    - talk move rate for each of the 3 talk moves
    - proportion of students engaged
  - exploratory
    - self-reported outcomes (e.g. NPS)
      - the response rate is 24.5% (150 of 611 instructors), which may limit the conclusions we can draw from this data
    - engagement with in-platform things, like IDE or teachers lounge
- **RQ3: heterogeneity analyses**
  - same outcomes as above, using the following variables for heterogeneity:
    - gender (binary)
    - returning instructor (binary)
    - age
    - US location (binary)

- baseline practices and engagement (1st week)

Knowledge of data

Prior knowledge*
Disclose any prior knowledge you may have about the dataset that is relevant for the proposed analysis. If you do not have any prior knowledge of it, please state so. Your prior knowledge could stem from working with the data first-hand, from reading previously published research, or from codebooks. Provide prior knowledge for every author separately.

We have conducted preliminary analyses on all the data to understand which variables are useable, but **without looking at the treatment status** in any analyses. Specifically, we did not use condition (treatment vs control) in any of the preliminary analysis.

We found one anomaly: even though we randomized students based on section assignments, it turns out that some students actually ended up in different sections during the course. Specifically, for about 10% of student-session pairings, students were in a different section than they were originally assigned. This mix-up lead to **59% of sections having some mixing across experimental groups** (n=337 sections). Since treatment groups were assigned at the student level, and student experience remained consistent throughout (e.g. they either never got feedback or always got feedback), this only introduced inconsistencies at the section leader level. Specifically, some section leaders had mixed sections where students might have been in different experimental groups. What this means is that the instructor level analyses may be less reliable, since instructors got exposed to a mix of treatment and control students. And this means that we may have to consider outcomes only up until those weeks where no contamination occurred for a particular SL. We may also have to conduct additional separate analyses to understand how this mixing might have affected student and SL outcomes.

In the final data these figures are: 8.9% of student-week attendance records (2,678 of 30,075) are in a section other than the student's assigned section, 361 of 612 attended sections (59%) contain students from more than one arm, and 111 of 599 assigned sections contain students from more than one arm, reflecting roster revisions between randomization and the first section. We will therefore (a) treat student-level analyses as the confirmatory analyses and instructor-level analyses as attenuated, (b) report instructor-level results both on the full sample and on the subset of instructor-weeks with no cross-arm mixing, and (c) report the share of a section's students in each arm as a section-level covariate in instructor analyses.

We also looked at recording durations to understand which recordings may need to be filtered out. We expect each session to be about an hour long, so we will have to remove or correct recordings that fall significantly outside this range. Similarly to our previous study, we will filter out recordings shorter than 30 mins (4.3% of the data) because they indicate that there might have been an issue with that session. We also noticed that some recordings are very long (19.1% are longer than 90 mins, with a maximum of 3.2 hours), indicating that recording might

have been left on. If treatment doesn't affect duration, we will leave the duration intact, otherwise we keep the first 60 minutes. We will compute and report results for both versions of the data.

Unuseable data

- Section-weeks with no usable transcript (165 of 3,623, 4.6%) drop out of all transcript-derived outcomes, and section-weeks with no recorded session duration (265 of 3,623, 7.3%) drop out of any rate-based outcome; both are retained for attendance and assignment outcomes.
- Students with no experimental arm assignment in the final data (n = 119) are excluded from all confirmatory analyses.
- Week 6 contains only 3 section records and no feedback cycle; it is excluded.

Statistical models*

For each hypothesis, describe the statistical model you will use to test the hypothesis. Include the type of model (e.g., ANOVA, multiple regression, SEM) and the specification of the model. Specify any interactions and post-hoc analyses and remember that any test not included here must be labeled as an exploratory test in the final paper.

Differential attrition

We will first conduct an analysis to test differential attrition in the data by condition. For this, we'll use the number of transcripts available for each student. We will use treatment status as a predictor, as well as covariates. If we find differential attrition, we will apply Lee bounds to bound the treatment effects for differential attrition.

Evaluating randomization

We will also evaluate randomization by running two-sample t-tests on each student characteristic, computing individual p values as well as a joint F statistic.

Models

For every research question we estimate ordinary least squares regressions, clustering standard errors at the level of randomization (the pre-course section roster). Treatment is coded as two indicators, StudentFeedback and StudentSLFeedback, with Control omitted; the pooled contrast uses AnyStudentFeedback. Each outcome is estimated in a separate regression. All models control for the covariates listed under Measured variables.

RQ1: Does joint feedback to both students and instructors improve student outcomes above and beyond feedback provided to instructors alone?

For student outcomes measured once per student-week (attendance, talk time in section, number of questions asked, proportion of the week's assignment completed) we fit

$$Y_{iv} = \delta_1 \, \text{StudentFeedback}_i + \delta_2 \, \text{StudentSLFeedback}_i + X_i\beta + \pi_v + \varepsilon_{iv} \quad (1)$$

where i indexes the student and w the week (1–5; week 0 precedes any feedback and enters as a baseline covariate). X_i contains student demographics (age, is female, in US), the demographics of the assigned instructor (age, is female, in US, returning section leader), the student's week-0 attendance and week-0 participation, and the discourse characteristics of the section's first session. π_v are week fixed effects. Standard errors are clustered at the pre-course section roster level. δ_1 and δ_2 test H1; the contrast $\delta_2 - \delta_1$ tests H1.1 (RQ1.1). For outcomes defined once per student (number of sessions attended in weeks 1–5, overall assignment completion) we fit the same specification without π_v at the student level.

RQ2: Does joint feedback to both students and instructors improve instructor outcomes above and beyond feedback provided to instructors alone?

Instructor outcomes are measured at the section-week level (instructor talk time, rate of each of the three talk moves, proportion of students engaged, whether the instructor viewed their feedback before the next section, time spent on the feedback page). Because randomization was on student rosters, the instructor-level treatment variable is the share of the instructor's assigned students in each arm; in the primary specification we use the share treated, and in a secondary specification we restrict to instructors whose assigned roster was entirely in one arm:

$$\mathbf{Y_{jv} = \delta_1\ ShareStudentFeedback_j + \delta_2\ ShareStudentSLFeedback_j + X_j\beta + \pi_v + \varepsilon_{jv}} \quad \mathbf{(2)}$$

where j indexes the instructor. X_j contains instructor demographics, the demographics of the assigned students, week-0 discourse characteristics (talk time, talk move rates, proportion of students engaged), section size and session duration. Talk move counts are converted to hourly rates by dividing by session duration. Standard errors are clustered at the instructor level. For outcomes defined once per instructor (any feedback viewed during the course, total time on feedback) we fit the same specification without π_v. Because 59% of attended sections mixed arms, we report (2) both on the full sample and on the unmixed subset, and treat the full-sample estimate as a lower bound.

RQ3: How do treatment effects vary by instructor and student characteristics?

We add an interaction between each treatment indicator and one moderator at a time to specifications (1) and (2), fitting one model per moderator: gender, age, US location, returning instructor, and baseline (week-0) practice and engagement. We report these as pre-registered heterogeneity analyses, with p-values adjusted for multiple comparisons across the moderator set within each outcome family.

Multiple comparisons and robustness

We group outcomes into families (student engagement, student achievement, instructor practice, instructor engagement with feedback) and apply a Benjamini–Hochberg correction within each family. As robustness checks we (a) re-estimate all models on transcripts truncated to the first 60 minutes, (b) re-estimate excluding sessions shorter than 30 minutes, (c) re-estimate student models clustering at the attended-section level, and (d) re-estimate instructor models on the unmixed subset. We report results for all versions.