Metadata

Title
The Effect of Giving Instructors Agency to Choose Among Different Types of Automated, Natural Language Processing Based Feedback

Description
This project builds on a previous study conducted in 2021 on Code in Place, a five-week-long 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 2023 on Code in Place, and its goal is to understand whether providing instructors agency in choosing the type of automated feedback they would like to receive influences their engagement with and impact of the feedback. Learner agency is thought to enhance engagement and improve outcomes, but few empirical studies have examined its effect in instructor learning settings. To answer this question, the study leverages both manual annotation and computational natural language processing techniques.

Contributors [alphabetical by last name]
Demszky, Dora
Dennison, Deepak
Hill, Heather
Kupor, Ashlee
Piech, Chris
Taylor, Eric

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: Do treatment group instructors *engage* with the feedback more?

RQ2: Do treatment group instructors *like* the feedback more?

RQ 3: Do treatment group instructors *change their instruction* more?

RQ3.1: Pre-feedback, after choice (week 1)
RQ3.2: Post-feedback, after choice (week 2-6)

RQ 4: Do treatment group instructors have better student outcomes?

Exploratory questions:

- How do instructors perceive the automated feedback, including the experimental feedback?
- How do treatment effects vary by instructor demographics and whether the instructor completed the training modules?
- How do treatment effects change over time?
- How does the effectiveness of feedback on talk moves compare within the control group?

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.

RQ1: We expect treatment group instructors to be more likely to engage with the feedback.

RQ2: We expect treatment group instructors to like the feedback more (agency > more control > more positive perception).

RQ3: We expect treatment group instructors to increase their use of talk moves and student talk percentage more both pre-feedback & post-feedback.

RQ4: We do not expect to see a significant treatment effect on student outcomes, given that those are more distal measures and are also noisy in nature. However, we do expect to see a positive trend.

Exploratory hypotheses:

- We expect a greater treatment effect for instructors who completed the training modules (a greater understanding of the talk moves will help them make a more informed choice on the feedback, and act on the feedback better).
- We expect treatment effects to fade over time, as time between when the choice was made (beginning of course) and the feedback increases.
- We do not have specific hypotheses about heterogeneous treatment effects by instructor demographics.

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)
- Student demographic information: gender, age, location (country)
- Instructors' choices for automated feedback
- Transcripts derived from session recordings
- Automated measures of instructional practice per class recording
  - Eliciting student ideas
  - Building on student ideas
  - Orienting students to one another
- Instructor survey responses on automated feedback (link)
- View logs of automated feedback page
- In-platform user input
  - Ratings of feedback
  - Reflections
- Student attendance data
- Student assignment completion data

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.*

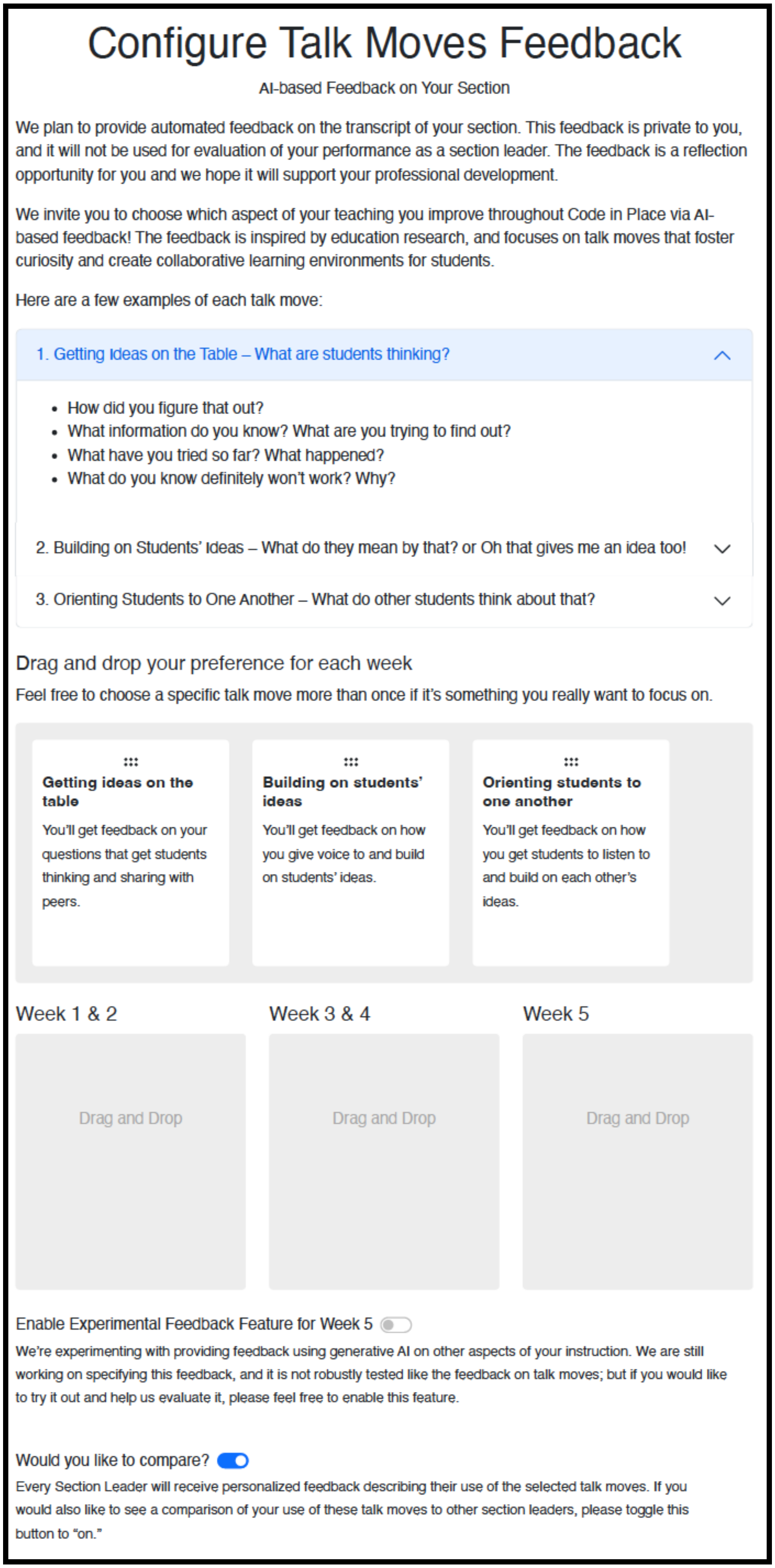
Randomized study setup

Sample size: 588 instructors

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.

Before the course began: We randomized instructors once they were accepted to teach in the course, and before the course began. Half of the instructors got a choice for what type of automated feedback they wanted to receive (see figure below). Instructors were asked to make this choice on the Code in Place website, and this action item was listed on their pre-course checklist. The choice involved feedback on 3 types of talk moves (Getting ideas on the table, Building on student ideas, Orienting students to one another), which they could select for pairs of weeks (1-2, 3-4, weeks 5-6*). Instructors also had the option to enable experimental, GPT-4 based feedback for the last 2 weeks and had the option to compare their metrics with other section leaders for all weeks. We displayed a short definition and an example below each talk move to help inform their choices. We sent email nudges to instructors before the course began to make a choice.

*We had thought that the course would only be 5 weeks long, hence the choice interface only had Week 5 listed for the third box; as we realized the course would be 6 weeks long, we applied their choices for week 5 to week 6 as well.



Configure Talk Moves Feedback

AI-based Feedback on Your Section

We plan to provide automated feedback on the transcript of your section. This feedback is private to you, and it will not be used for evaluation of your performance as a section leader. The feedback is a reflection opportunity for you and we hope it will support your professional development.

We invite you to choose which aspect of your teaching you improve throughout Code in Place via AI-based feedback! The feedback is inspired by education research, and focuses on talk moves that foster curiosity and create collaborative learning environments for students.

Here are a few examples of each talk move:

1. Getting Ideas on the Table – What are students thinking?

- How did you figure that out?
- What information do you know? What are you trying to find out?
- What have you tried so far? What happened?
- What do you know definitely won't work? Why?

2. Building on Students' Ideas – What do they mean by that? or Oh that gives me an idea too!

3. Orienting Students to One Another – What do other students think about that?

Drag and drop your preference for each week

Feel free to choose a specific talk move more than once if it's something you really want to focus on.

Getting ideas on the table

You'll get feedback on your questions that get students thinking and sharing with peers.

Building on students' ideas

You'll get feedback on how you give voice to and build on students' ideas.

Orienting students to one another

You'll get feedback on how you get students to listen to and build on each other's ideas.

Enable Experimental Feedback Feature for Week 5 ○

We're experimenting with providing feedback using generative AI on other aspects of your instruction. We are still working on specifying this feedback, and it is not robustly tested like the feedback on talk moves; but if you would like to try it out and help us evaluate it, please feel free to enable this feature.

Would you like to compare? ●

Every Section Leader will receive personalized feedback describing their use of the selected talk moves. If you would also like to see a comparison of your use of these talk moves to other section leaders, please toggle this button to "on."

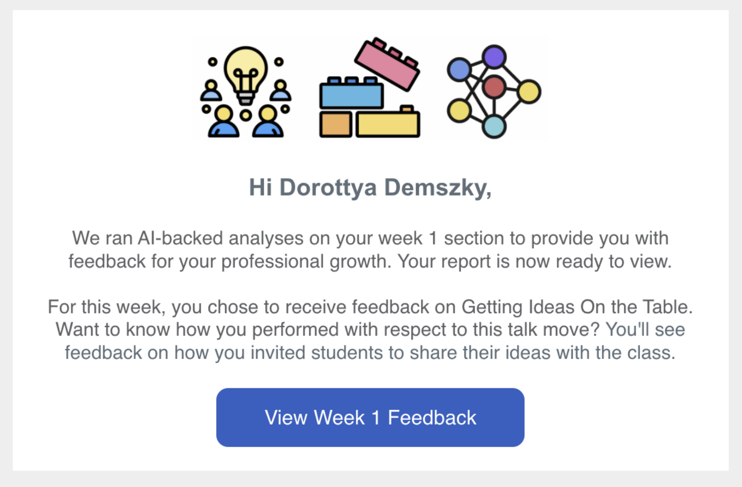
The control group did not get to choose. Instead the control group was randomly assigned to feedback under the constraint that the distribution of feedback patterns in the control group was the same as the distribution in the treatment group. For example, 36% of the treatment group chose the pattern: 2 weeks on getting ideas on the table, 2 weeks on building on student ideas, 2 weeks of experimental feedback. Thus 36% of the control group were assigned to that same pattern. This assignment of the control group means that the only difference between treatment and control, in expectation, is whether the instructor chose their pattern of feedback or were assigned their pattern of feedback.

About 80% of treatment group instructors made a choice of what feedback to receive. The 20% that did not choose got were assigned feedback with a similar method as the control group.

All instructors had access to training modules that explained each talk move and showed animations to illustrate the talk move. Both control and treatment group instructors were encouraged to complete these modules prior to their first section. Completion rates were about 40% for each of the training modules.

Choice of feedback: 36% of instructors choose the following sequence of feedback: Getting ideas on the table, building on student ideas, then the experimental (GPT-4) feedback with the choice to compare their metrics to other section leaders; this sequence of feedback was by far the most popular choice. The rest of the instructors choose a combination of other feedback choices.

Once the course began: After each of their six sessions, instructors received automated feedback based on their chosen or assigned feedback type. The feedback was released to everyone on Saturday (sections taught Wed-Friday) via email. The email varied: the treatment group was reminded that they got a choice and here's the feedback on that topic. The control group did not get this reminder that they had made a choice. Below is an example email that the treatment group received.



Hi Dorottya Demszky,

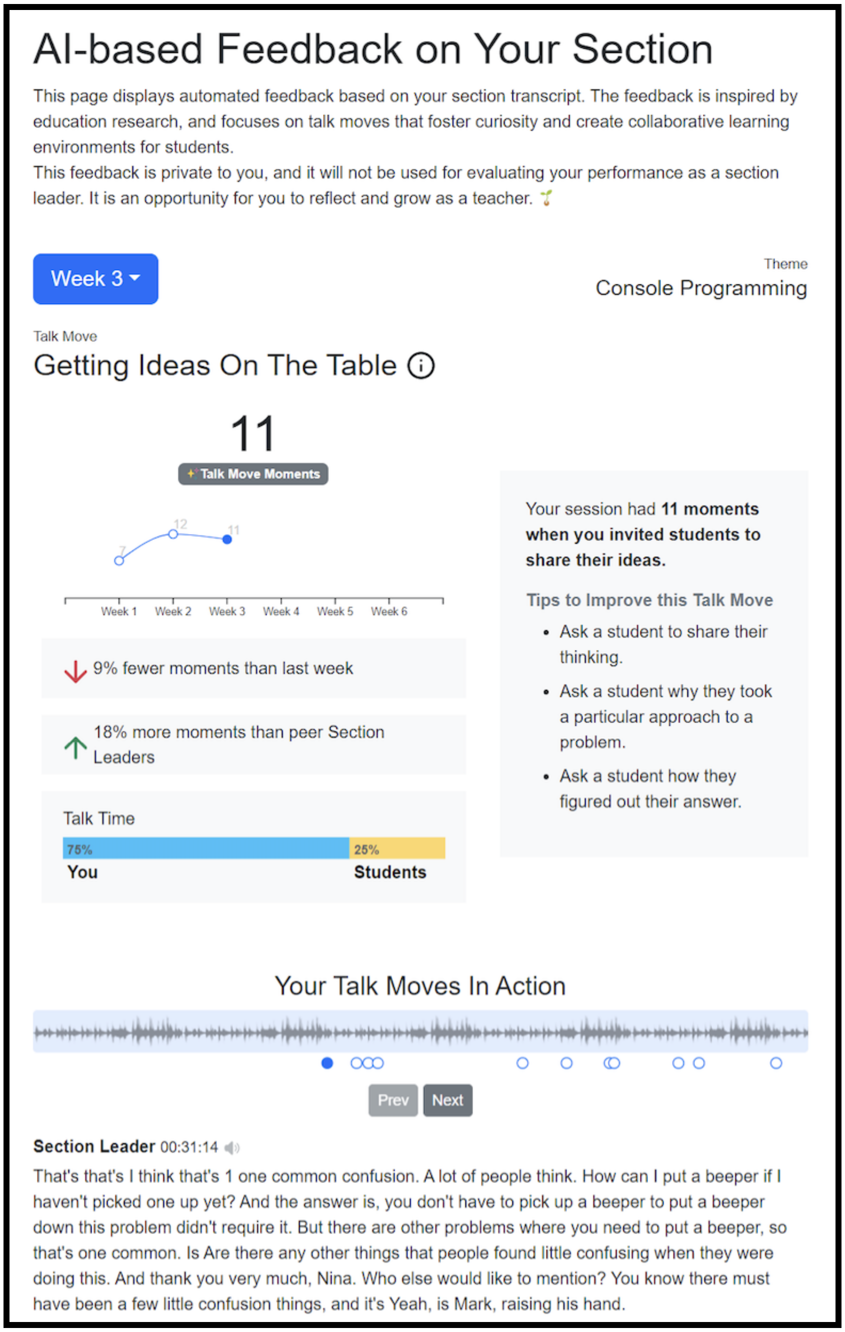
We ran AI-backed analyses on your week 1 section to provide you with feedback for your professional growth. Your report is now ready to view.

For this week, you chose to receive feedback on Getting Ideas On the Table. Want to know how you performed with respect to this talk move? You'll see feedback on how you invited students to share their ideas with the class.

All instructors could view their automated feedback on the Code in Place webpage. The talk move feedback included several components:

- Introduction to the feedback
- Summary statistics for the given talk move
- Definition of each talk move
- Their talk moves in action (list of talk moves from their transcript)
- Link to the relevant training module
- Reflection opportunities

Screenshot of part of the feedback page (we also showed a link to the training module and provided instructors with reflection opportunities).



AI-based Feedback on Your Section

This page displays automated feedback based on your section transcript. The feedback is inspired by education research, and focuses on talk moves that foster curiosity and create collaborative learning environments for students.

This feedback is private to you, and it will not be used for evaluating your performance as a section leader. It is an opportunity for you to reflect and grow as a teacher. 🌱

Getting Ideas On The Table ⓘ

9% fewer moments than last week

18% more moments than peer Section Leaders

Talk Time

You	Students
75%	25%

Your session had **11 moments when you invited students to share their ideas.**

Tips to Improve this Talk Move

- Ask a student to share their thinking.
- Ask a student why they took a particular approach to a problem.
- Ask a student how they figured out their answer.

Your Talk Moves In Action

Section Leader 00:31:14

That's that's I think that's 1 one common confusion. A lot of people think. How can I put a beeper if I haven't picked one up yet? And the answer is, you don't have to pick up a beeper to put a beeper down this problem didn't require it. But there are other problems where you need to put a beeper, so that's one common. Is Are there any other things that people found little confusing when they were doing this. And thank you very much, Nina. Who else would like to mention? You know there must have been a few little confusion things, and it's Yeah, is Mark, raising his hand.

After the course. We sent a final survey (preview here) to all instructors regardless of condition to ask them about their experience with the automated feedback, as well as questions related to self-efficacy. A randomly sampled third (n=200) of instructors were incentivized to complete the feedback with a $5 Amazon gift card. The rest of the instructors were not incentivized. The response rate was about 50% for the incentivized instructors and 40% for all instructors. The responses do not vary significantly based on whether the instructors were incentivized.

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.

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)
    - We use binary as opposed to including all possible values (female, male, non-binary, na) because when running a regression with all values to predict an outcome (viewing the feedback), the model does not converge. The binary value captures most of the variance in this case, as there are few values that do not fall under male or female.
  - In US (binary)
    - We use binary (as opposed to unique continents, or top 10 countries) because it captures most of the variance when running a regression with location as a predictor and viewing the feedback as an outcome.
- Student demographics
  - Age
  - Is female (binary)
  - In US (binary)
    - We use binary values for students too to ensure consistency with instructor covariates and for the same reasons as outlined in the instructor covariate section above.

- 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

Outcome(s):

- **RQ1: engagement with feedback**
    - **Viewing the feedback before their subsequent section (binary)**
        - At any point during the course
        - By week
    - Number of sessions of viewing the feedback
    - **Total time spent on page (in seconds)**
        - Total across weeks
    - Exploratory outcomes
        - Sharing their reflections with other section leaders – between 2-14% total (binary)
- **RQ2: perception of feedback**
    - Aggregated items from the final survey (Factor Analysis showed that only one factor has an eigenvalue greater than >1, following the Kaiser criterion, indicating that a single factor explains most of the variance)
    - NPS score (1-10) from final survey
- **RQ3.1: impact on instruction post-choice, pre-feedback (week 1)**
    - hourly rate of each of the 3 talk moves in the first session
        - getting ideas on the table
        - building on student ideas
        - connecting student ideas
    - student talk percentage
    - potentially other discourse features (distal, not expecting impact)
        - proportion of students participating (saying something or typing something in chat)
        - uptake (Demszky et al., 2021)
        - number of questions asked by instructor
- **RQ3.2: impact on instruction post-feedback (weeks 2-6)**
    - hourly rate of each of the 3 talk moves in a given session
        - getting ideas on the table
        - building on student ideas
        - connecting student ideas
    - student talk percentage
    - potentially other discourse features (distal, not expecting impact):
        - proportion of students participating (saying something or typing something in chat)
        - uptake (Demszky et al., 2021)
        - number of questions asked by instructor
- **RQ4: impact on student outcomes (distal, unlikely to observe impact)**

- number of attended sessions between weeks 2-6
- proportion of assignment completed for each week (overall proportions listed below)

  ```
  week 1 assignment: 0.80
  week 2 assignment: 0.65
  week 3 assignment: 0.61
  week 4 assignment: 0.55
  week 5 assignment: 0.48
  week 6 assignment: 0.35
  ```

- final project completion (overall 25%)

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 looked at response rates to different items, and to see which variables we may need to collapse (e.g. location, gender, survey items), and whether there were anomalies in our sample (e.g. choice sequences in feedback).

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 (5% 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 (15% are longer than 90 mins, with a maximum of ~5 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:

- Ratings of talk move feedback within page (we had a star rating option, but instructors only rated the feedback 4% of the time)
- Ratings of experimental feedback within page (we asked them to rate it for reliability and helpfulness but instructors only rated the feedback 7% of the time)

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 two outcomes: a) number of transcripts available for each instructor, b) likelihood of completing the final survey. 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 demographic characteristic, computing individual p values as well as a joint F statistic.

RQ1: Do treatment group instructors engage with the feedback more?

We use ordinary least squares to fit the following regression specification:

$$Y_{iw} = \delta T_i + X_i\beta + \pi_w + \varepsilon_{iw} \qquad (1)$$

where the indicator variable T_i=1 if the instructor (indexed by *i*) was assigned to the treatment condition. We estimate (1) separately for each outcome, Y_{iw} described in the Measured Variables section. Each observation is nested within an instructor, and we have one observation per week (indexed by *w*) for each instructor (corresponding to a single session recording and unit of feedback).

We cluster standard errors at the instructor level. The vector X_i includes controls for instructor demographics, proportion of students attending the first session, as well as student demographics assigned to the instructors' section (without necessarily attending any of the sections). We also control for the week of observation (1-6) effects, i.e., week fixed effects π_w.

Three sub-questions:

- Ever check?
- Check by week?
- Heterogeneous week?

RQ2: Do treatment group instructors *like* the feedback more?

We use ordinary least squares to fit the following regression specification:

$$Y_i = \delta T_i + X_i \beta + \varepsilon_i \quad (2)$$

We estimate (2) separately for each outcome (aggregated survey responses + NPS) described in the Measured Variables section. Each observation is an instructor in the course. As in (1), we control for instructor demographics, proportion of students attending the first session, as well as student demographics assigned to the instructors' section (without necessarily attending any of the sections).

RQ 3: Do treatment group instructors *change their instruction* more?

Our estimates for RQ3 use ordinary least squares to fit variations on the following regression specification:

$$Y_{iwm} = \delta T_i + X_i \beta + \pi_w + \theta_m + \varepsilon_{iwm} \quad (3)$$

where Y_{iwm} is the instructor's (indexed by *i*) score for a given talk move (indexed by *m*) during a given week (indexed by *w*). The indicator variable T_i = 1 if the instructor was randomly assigned to the treatment condition. The vector X_i is a set of controls for instructor demographics, proportion of students attending the first session of the given instructor, as well as student demographics assigned to the instructors' section. The specification also includes week fixed effects, π_w, and move fixed effects, θ_m. The intent to treat effect estimate is the estimated δ.

We also test for heterogeneity of treatment effects by fitting the following specification, which builds on (3):

$$Y_{iwm} = \delta T_i + \alpha F_{iwm} + \gamma(T_i * F_{iwm}) + X_i \beta + \pi_w + \theta_m + \varepsilon_{iwm} \quad (4)$$

The new term in (4) is the indicator variable F_{iwm} which is = 1 if instructor *i* was given feedback on move *m* in week *w*. Recall, treatment instructors chose what feedback to receive and when, while control instructors were assigned feedback. The estimate of γ reflects whether the treatment effect was larger (or smaller) when the talk move outcome is the topic of feedback chosen or assigned.

RQ3.1: Pre-feedback, after choice (week 1)

Analysis 1: We estimate specification (3) but only using observations from week 1 (only w=1). One observation is a single week 1 transcript. We estimate specification (3) separately for each move *m* described in the Measured Variables section. (Thus the regression specification simplifies to $Y_{i1} = \delta T_i + X_i \beta + \varepsilon_{i1}$.)

Analysis 2: We also estimate a modified version of specification (4), but still limited to only week 1 (w=1) and estimated separately for each move *m*. Additionally, in week 1 instructors have not yet received any feedback (F_{i1} does not exist). In this analysis of week 1 we want to understand if choosing a particular talk move for week 2 induced treatment instructors to use that talk move more compared to the control group even before they received feedback. Thus we use F_{i2} in the specification. The regression specification is thus:

$$Y_{i1} = \delta T_i + \alpha F_{i2} + \gamma(T_i * F_{i2}) + X_i\beta + \varepsilon_{i1} \quad (5)$$

Analysis 3: Finally, we estimate specification (3) and (4) pooling together all moves *m* into one estimation sample. Before pooling we standardize each talk move outcome (mean 0, s.d. 1 using the full sample), so now Y_{iwm} is in standard deviation units. By pooling together we expect to gain more power, at the expense of some interpretability. However, we continue to limit estimation to only week 1 (only w=1). And we continue to use F_{i2} in the heterogeneity specification. Thus specification (3) becomes $Y_{i1m} = \delta T_i + X_i\beta + \theta_m + \varepsilon_{i1m}$, and specification (4) becomes $Y_{i1m} = \delta T_i + \alpha F_{i2m} + \gamma(T_i * F_{i2m}) + X_i\beta + \theta_m + \varepsilon_{i1m}$.)

RQ3.2: Post-feedback, after choice (week 2-6)

Analysis 1: We estimate specification (3) pooling together observations from weeks 2-6. Each observation is a transcript from a given week *w* for a given instructor *i*. We estimate specification (3) separately for each move *m* described in the Measured Variables section. In addition, we add controls in X_i for week 1 discourse features, but will also do the analysis without controlling for week 1 disclosure features. (Thus specification (3) becomes $Y_{iw} = \delta T_i + X_i\beta + \pi_w + \varepsilon_{iw}$.)

Analysis 2: To test for heterogeneity, we estimate specification (4) using the same sample as in Analysis 1. (Thus specification (4) becomes $Y_{iw} = \delta T_i + \alpha F_{iw} + \gamma(T_i * F_{iw}) + X_i\beta + \pi_w + \varepsilon_{iw}$.)

Analysis 3: Finally, we estimate specification (3) and (4) pooling together all moves *m* into one estimation sample, and pooling together weeks 2-6. Before pooling we standardize each talk move outcome (mean 0, s.d. 1 using the full sample), so now Y_{iwm} is in standard deviation units. By pooling together we expect to gain more power, at the expense of some interpretability.We exclude transcripts from this analysis that got the experimental feedback (since the experimental feedback has no associated outcome measure). We control for week 1 discourse features, but will also do the analysis without controlling for week 1.

RQ 4: Do treatment group instructors have better student outcomes?

We use ordinary least squares to fit the regression specification:

$$Y_j = \delta T_{i(j)} + X_{i(j)}\beta + \varepsilon_j \quad (6)$$

Here students are indexed by *j* and each student is assigned to just one instructor, *i* = *i(j).* We estimate (6) separately for each outcome described in the Measured Variables section. Each observation is a student in the course. We control for instructor demographics, whether the student attended their first session (binary), student demographics. We cluster standard errors at the instructor level.

For all research questions, we only use a small number of pre-defined primary estimates to avoid multiple hypothesis testing. We will also consider a wider set of estimates beyond these primary, and apply a Bonferroni correction.