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 and 2023 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 the Schoolhouse.world platform, 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
Hicke, Yann
Karpawich, Max
Olson, Mariah
Yun, Joy

Category
Project

License
CC-By Attribution International

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 providing automated feedback to instructors improve *instructor* practice?

RQ2: Does providing automated feedback to instructors improve *student outcomes*, including their attendance & engagement in section, practice test scores and their experience?

RQ3: Does joint feedback to both students and instructors improve instructor and student outcomes above and beyond feedback provided to instructors alone?
RQ3.1: Does giving prosocial vs self-oriented feedback make a difference?

RQ4: How did instructors and students perceive the feedback and what were their barriers for acting upon them?

Exploratory questions:

- How do treatment effects vary by instructor and student characteristics, including demographics and baseline practices/engagement in the course?
- How do treatment effects change over time?
- How do treatment effects vary between the two bootcamps?

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.

1. Providing automated feedback to instructors improves their teaching practice.
2. Providing automated feedback to instructors improves student outcomes.
3. Providing joint feedback to both students and instructors improves instructor and student outcomes above and beyond feedback to instructors alone.
    a. We may not observe significant differences in messaging (pro-social vs self-oriented), but pro-social feedback may show slightly more positive impacts than self-oriented messaging.
4. Overall, instructors and students have a positive perception of the feedback, but they may raise important barriers to engagement, such as difficulty to act on the feedback, or for instructors, difficulty to engage students.

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 characteristics: gender, tutoring experience on the platform, high school grade, location (zip code)
- Student characteristics: gender, self-reported SAT score range, grade, number of SH sessions before (potentially in other SH programs)
- Transcripts derived from session recordings
- Automated measures of instructional practice per class recording
    - Talk time
    - Number / proportion of students participating
    - Eliciting student ideas
    - Building on student ideas
    - Number of student questions asked
    - Average length of tutor utterance (spoken + chat)
    - Average length of student utterance (spoken + chat)
    - Number of turns
    - Maybe (pending validation):
        - Student reasoning
        - Number of problems discussed
            - worked examples, practice test
- Instructor survey responses on automated feedback
- View logs of on platform
    - Automated feedback page
    - Maybe: number of optional trainings completed
- In-platform user input from tutors
    - Reflections / goal-setting
    - Sharing reflection
    - When tutor clicked OK on feedback vs. when they joined their next session
- Student attendance data
- Student practice test scores
- Student rating of session (1-3)
    - Helpful, not helpful, super helpful (required)
- Every 3 sessions, private NPS rating for students (required)

1st, 3rd and 7th sessions
- Qualitative interviews with instructors and students

Maybe (pending College Board approval and verification):
- Post-SAT scores, aggregated to the cohort level
- Pre-post survey from instructors and tutors, administered by College Board, about their experience

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

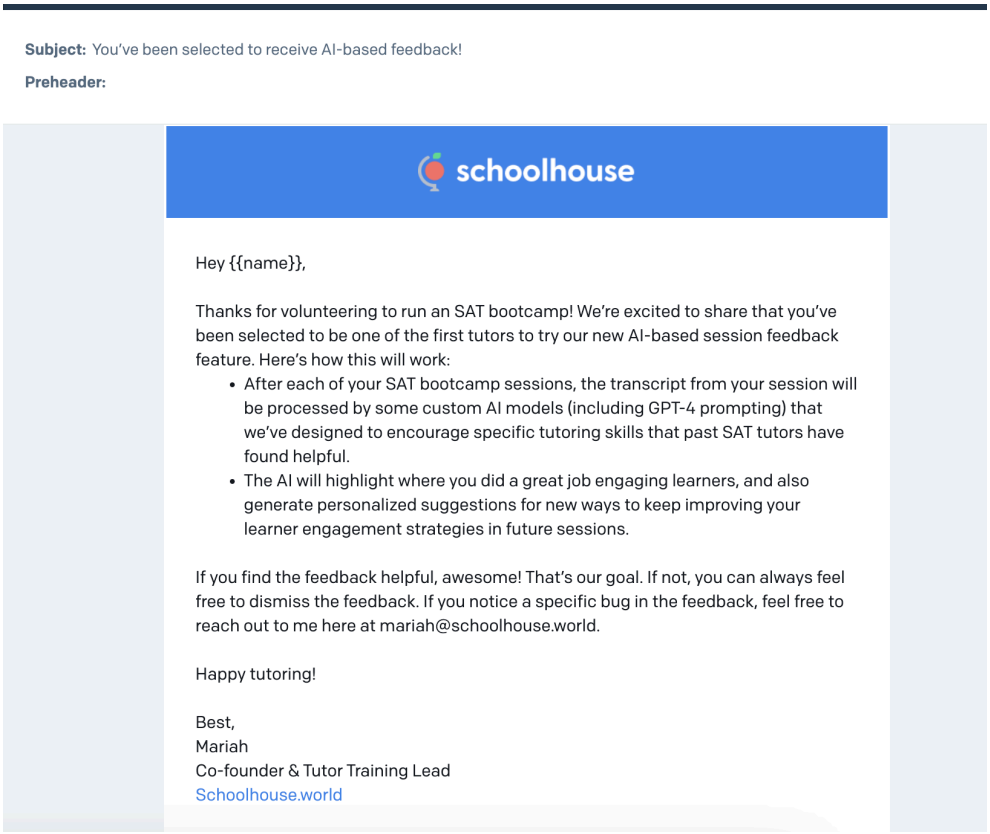
Sample size (unfiltered):

May Bootcamp: 697 tutors

June Bootcamp: 517 tutors

The study was conducted in a free, online 4-week long online peer SAT math tutoring bootcamp on the Schoolhouse.world platform. Anyone with an SAT subject score of 650 or above could apply to serve as a peer tutor for teaching that subject. As long as they complete the Schoolhouse asynchronous tutor training, they are then eligible to teach their first bootcamp. Our participant sample consists of all instructors and students in the April/May 2024 and the June 2024 bootcamps.

In the May bootcamp, tutors in the treatment arms of the RCT received an email prior to the start of the bootcamp informing them they would receive feedback and explaining the relevant parts of the feedback modal. For the June bootcamp, tutors were not primed to receive feedback.



Subject: You've been selected to receive AI-based feedback!

Preheader:

schoolhouse

Hey {{name}},

Thanks for volunteering to run an SAT bootcamp! We're excited to share that you've been selected to be one of the first tutors to try our new AI-based session feedback feature. Here's how this will work:

- After each of your SAT bootcamp sessions, the transcript from your session will be processed by some custom AI models (including GPT-4 prompting) that we've designed to encourage specific tutoring skills that past SAT tutors have found helpful.
- The AI will highlight where you did a great job engaging learners, and also generate personalized suggestions for new ways to keep improving your learner engagement strategies in future sessions.

If you find the feedback helpful, awesome! That's our goal. If not, you can always feel free to dismiss the feedback. If you notice a specific bug in the feedback, feel free to reach out to me here at mariah@schoolhouse.world.

Happy tutoring!

Best,
Mariah
Co-founder & Tutor Training Lead
Schoolhouse.world

Before the first tutoring session, tutors were randomized into one of four conditions:

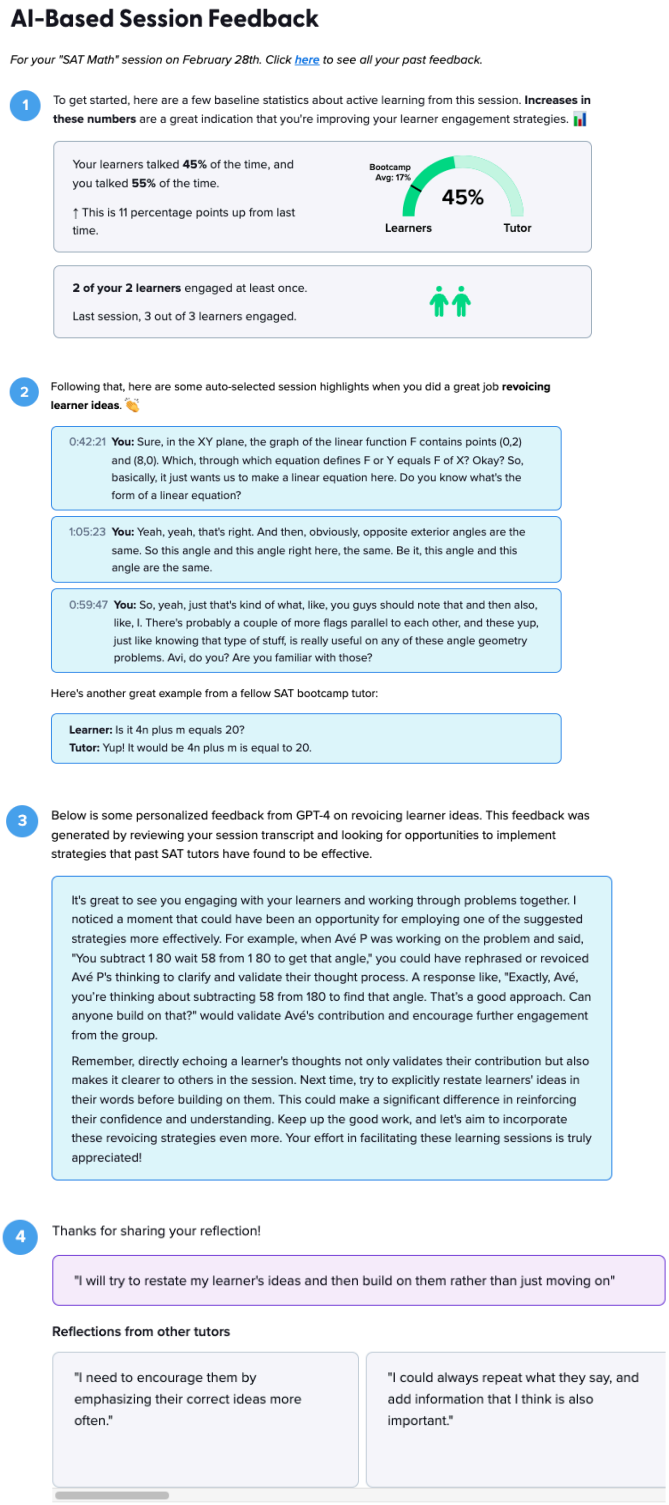
- **Control** = 30% were assigned to control condition, and conducted business as usual
- **TutorFeedback** = 30% were assigned to receive automated feedback on their instruction
- **TutorStudentFeedback$_{Self}$ =** 15% were assigned to receive automated feedback on their instruction AND their students also received automated feedback with self-oriented messaging related to the importance of engaging in section
- **TutorStudentFeedback$_{Social}$ =** 15% were assigned to receive automated feedback on their instruction AND their students also received automated feedback with pro-social messaging related to the importance of engaging in section

Feedback to Instructors

Instructors in the TutorFeedback and TutorStudentFeedback conditions received automated feedback with the following components:

- Introduction to the feedback
- Summary statistics of the session and comparison to the previous session
    - tutor talk percentage
    - proportion of students engaged
- Description of talk move in focus for that session
- Their talk moves in action (list of talk moves from their transcript)
- For first two sessions: link to relevant training module

- GPT-4 Turbo generated actionable suggestions for next session
- Reflection opportunity



The talk moves that were in focus changed over the course of the bootcamp, following a pre-defined curriculum of talk moves for each of the 8 sessions:

Session #
1: Eliciting ideas from students
- 344/363 tutors in the treatment conditions did not receive feedback for session 1 during June bootcamp due to an error

2: Eliciting ideas from students
3: Revoicing student ideas
4: Revoicing student ideas
5: No feedback
6: Prompting for reasoning
7: Prompting for reasoning
8: No feedback - received end-of-bootcamp survey on the AI feedback

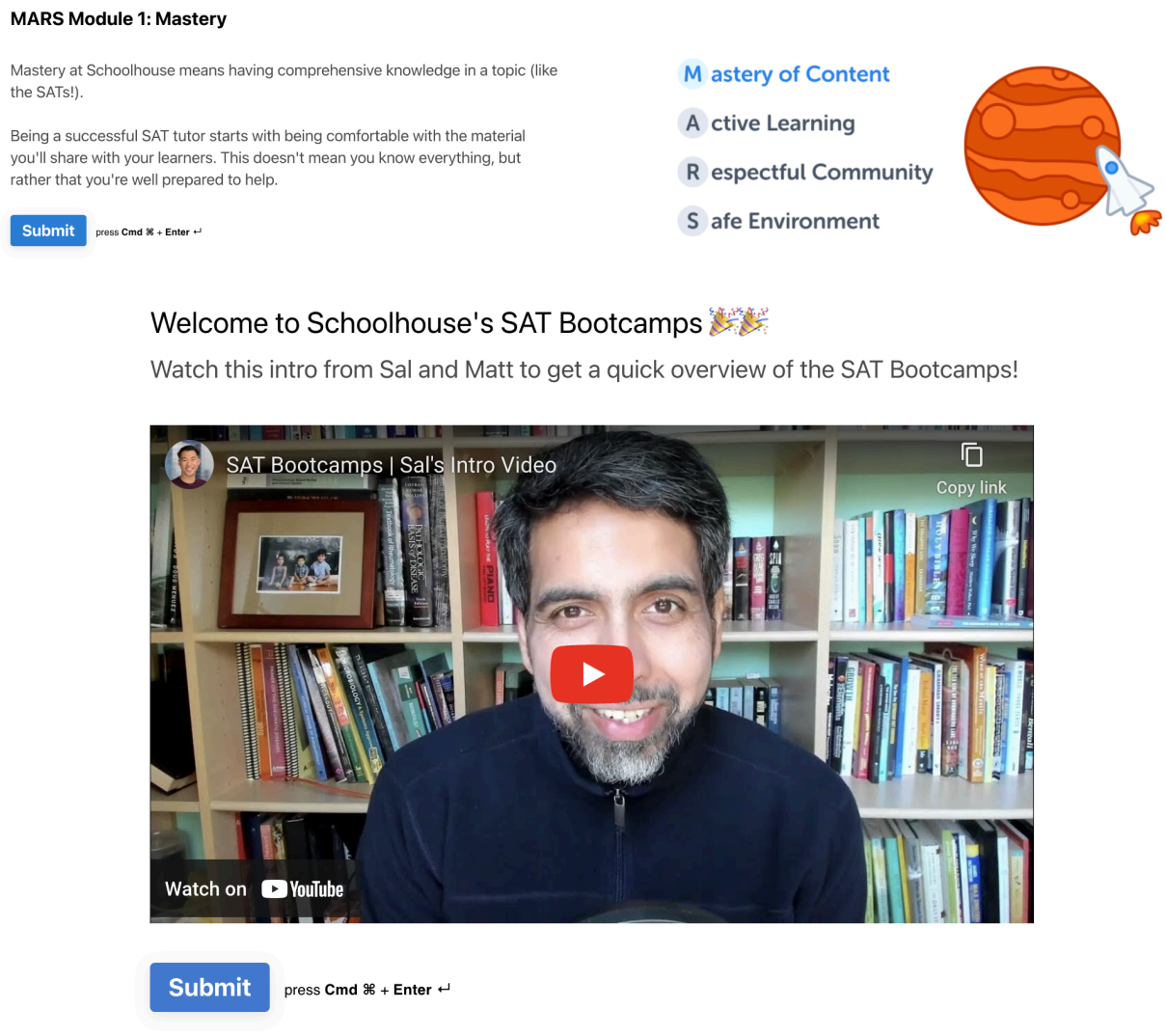
The talk moves are defined as:

- Inviting learner ideas
    - This talk move was identified by first filtering session transcripts with a fine-tuned question detection model, which isolates utterances resembling questions. The questions are then passed to an Electra-base model, fine-tuned to identify the "pressing for reasoning" and "pressing for accuracy" labels from the TalkMoves Dataset by Suresh et al. (2022)
- Building on learner ideas
    - For identifying this talk move, we used the uptake model developed by Demszky et al. (2021). This model analyzes utterances from the session transcripts to pinpoint when tutors effectively engage with and extend student contributions.
- Pressing for reasoning
    - For the sessions with this focus, the same Eliciting model was used as the "inviting learner ideas" sessions.

After a tutor taught their section on Zoom, their transcript was analyzed through our automated analysis pipeline. The analysis was usually completed within a few hours. Once the feedback for the most recent session becomes available, tutors receive an email notifying and encouraging them to log into the Schoolhouse.world platform to view it. The feedback appears as a pop-up modal the next time they log into Schoolhouse.world. All previous feedback can be accessed again from the tutor's personal profile page.

Edge-cases: if a tutor misses a session, they do not receive any feedback. They will receive feedback after the next session they teach. Tutors did not substitute teaching other cohorts.

All instructors were required to complete training about Schoolhouse's MARS rubric (Mastery, Active Learning, Respectful Community, and Safety), general SAT knowledge, and new information about the digital SAT. They also participate in a 1-hour live onboarding session.

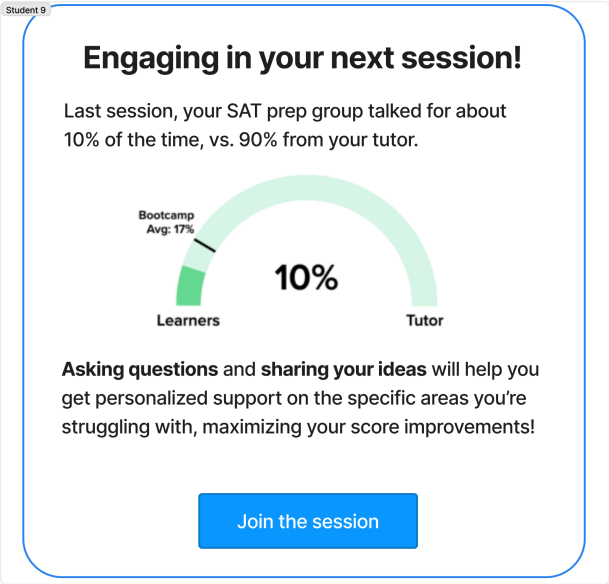


Feedback to Students

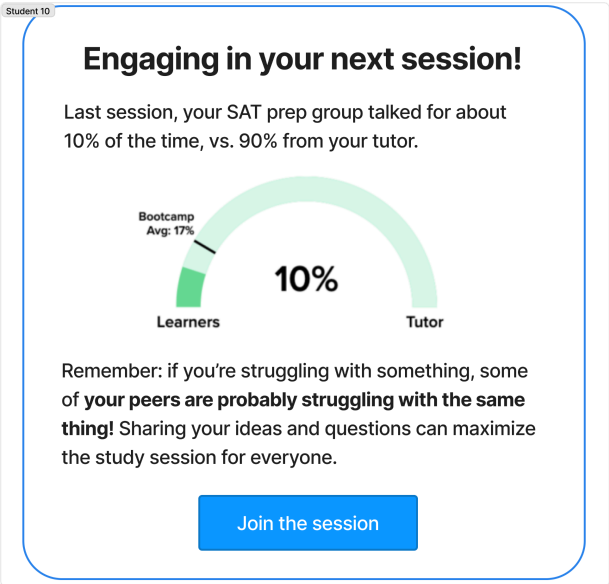
Students in the TutorStudentFeedback groups received automated feedback on their engagement in the tutoring session. The feedback included the following components:

- Student talk time ratio in section
- Motivational message to encourage students to participate in the session

Students were randomized to receive one of two types of motivational messages. Students in the TutorStudentFeedback$_{Self}$ group received a message that encouraged them to participate in order to optimize their own learning:



Students in the TutorStudentFeedback$_{Social}$ group received a message that encouraged them to participate in order to help everyone else learn:



Students received the feedback from their previous session as a pop-up modal right before they join their next session.

If a student did not attend their section, they did not receive feedback. If a student attended a session of a different tutor, they received the treatment condition assigned to that new tutor (i.e.

if they were in control but then dropped-in on another session in the treatment group, they received feedback for that session) -- this did not happen often, if at all.

At the end of the study

After the last session, automated feedback to tutors included a few survey questions to probe their perceptions of the automated feedback:

We also interviewed a sample of tutors and students to gauge their perception of the feedback. A random sample of tutors and students from each treatment arm of the study was emailed after the Bootcamp with an invitation to sign up for an interview in exchange for a $15 Amazon gift card. In total, 17 tutors were interviewed, with 5, 6, and 6 from each arm; 9 learners were interviewed, with 3, 2, and 4 from each arm. The interviews were conducted virtually over Zoom by a member of the Schoolhouse team. In the first phase of the interview, the interviewee was asked about their overall experience in the SAT Bootcamp; in the second phase, the interviewee was shown their automated feedback and asked a set of questions about how they felt about it. Finally, the interviews with tutors were different from the interviews with students; each was tailored to the specifics of the feedback they had received and their role in the Bootcamp.

Tutor Questions

1. Can you tell me what the most challenging thing was for you as a tutor for the bootcamp?
2. In your initial tutor training, you may remember that you were encouraged to engage learners to lead the problem-solving process themselves. Did you feel like you were able to do this? Why or why not?
3. *For sections 1-3 of the tutor feedback:*
    a. How did you feel about section *n*?
    b. Did section *n* encourage you, discourage you, and why?
    c. How was section *n* useful or not for you?
    d. Can you recall ever acting on the feedback in section *n*?
    e. Was it ever challenging to act on the feedback in section *n*?
4. If you had the option, would you turn off any part of the AI feedback — maybe you would turn off all of it — and why?
5. Of all four sections of the feedback, which was the most helpful and why?
6. Which was the least helpful and why?

Learner Questions

1. How much did you engage with each session?
    a. How much did you engage by speaking versus typing in the Zoom chat during your sessions? Which did you do more often, and why?
    b. Do you remember any changes to how you engaged as the Bootcamp progressed? Why or why not?
2. Did you feel that your tutoring group worked well together? Why or why not?
    a. Do you remember any changes to how your group as a whole engaged as the Bootcamp progressed?
    b. Did you see other learners in your group ask more questions as the Bootcamp progressed?
3. Can you remember an example of when you asked a question during a session?

    a. *If they don't*: Were there any questions that you wanted to, but did not ask? Why did you hold back?
    b. *If they do*: Who responded? What did you learn from their response?
4. How much of the interactions were tutor-learner versus learner-learner?
    a. Can you remember an example of when you responded to another student's question or comment?
        i. *If they don't*: Were there any moments when you wanted to respond, but didn't? Why did you hold back?
5. How much did you pay attention when other learners spoke or typed in the chat during sessions, and how much of the time did you tune them out.
6. Did you look at the talk time feedback at the end of your sessions?
7. How did you feel about the feedback?
8. Did it encourage you or discourage you, and why?
9. How was the feedback useful or not for you?
10. If you had the option, would you turn off any part of the AI feedback — maybe you would turn off all of it — and why?
11. Do you remember ever acting on the feedback in future sessions?
12. Was it ever challenging to act on the feedback in future sessions?

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 covariates
    - Gender
    - Number of SAT bootcamps tutored
    - Number of non-SAT tutoring on SHW
    - High school grade level
    - Maybe: zipcode (we are figuring out the useability of this data)
    - Maybe: signal of proactivity (number of optional training modules; we're still figuring out useability of this data)
- Student covariates
    - Gender
    - Self-reported SAT score range
    - Grade level
    - Number of SH sessions taken prior (including other SHW programs)
    - Maybe: zipcode (we are figuring out the useability of this data)
- Session number (1 to 8; only used for transcript-level analyses)
- Features of first session
    - Tutor talk time ratio
    - Number of students attending
    - Proportion of students participating
    - Rate of three key discourse features: inviting, building, reasoning

Outcome(s):

- **RQ1: instructor practice**
    - tutor talk ratio
    - proportion of students participating
    - hourly rate of each of the 3 key talk moves (inviting, building, reasoning)
    - student talk percentage
    - potentially other discourse features (distal, not expecting impact)
        - number of questions asked by tutor
        - number of problems discussed
- **RQ2: student engagement in section**
    - student attendance
    - student participation (via chat or spoken)
    - student number of questions asked
    - potentially other discourse features (distal, not expecting impact)
        - student reasoning
- **RQ2: student practice test scores**
- **RQ2: student experience**
    - section rating
    - NPS
- **RQ3: same as RQ1 and RQ2**
- **RQ4: instructor and student perception of feedback**
    - Final survey results for instructors
    - Qualitative interview responses

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 75 minutes long, so we may 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 (2% 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 (1% are longer than 2 hours), indicating when a tutor stayed on to work on additional practice problems with a student. If treatment doesn't affect duration, we will leave the duration intact, otherwise we keep the first 90 minutes. We will compute and report results for both versions of the data.

<u>Unuseable data:</u>

- We will remove transcripts from session 5, which was dedicated to students doing a practice test. No automated feedback was delivered after this session.

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 instructor. 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: Does providing automated feedback to instructors improve *instructor* practice and experience with the course?

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

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

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 covariates, features of the first (baseline) session, as well as student demographics assigned to the instructors' section (without necessarily attending any of the sections). We also control for session id (1-8) effects, i.e., week fixed effects π_w.

When using tutor experience as an outcome, we conduct the same analysis as above, except at the instructor, rather than at the session-level.

RQ2: Does providing automated feedback to instructors improve *student outcomes*, including their attendance & engagement in section, practice test scores, NPS?

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

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

We estimate (2) separately for each outcome described in the Measured Variables section. We conduct analyses of attendance and engagement at the session-level, clustering standard errors by student and tutor. We conduct analyses at the student-level for practice test scores and NPS. As in (1), we control for instructor demographics, student demographics, proportion of students attending the first session, student demographics assigned to the instructors' section (without necessarily attending any of the sections), and baseline discourse features as well as session id for session-level analyses.

RQ3: Does joint feedback to both students and instructors improve instructor and student outcomes above and beyond feedback provided to instructors alone?
RQ3.1: Does giving prosocial vs self-oriented feedback make a difference?

Within the group of instructors who got automated feedback, we will compare the two groups that received or did not receive joint feedback (also sent to students). We will use similar analyses as the ones listed above.

RQ4: How did instructors and students perceive the feedback and what were their barriers for acting upon them?

We will provide descriptive analyses of instructors' survey responses about the automated feedback. As for the interview data, we will conduct qualitative coding to identify key themes in the interviews and report those results.