Can Automated Feedback Improve Teachers' Uptake of Student Ideas? Evidence From a Randomized Controlled Trial in A Large-Scale Online Course *

Dorottya Demszky Jing Liu Heather C. Hill Dan Jurafsky
Chris Piech

June 3, 2023

Abstract

Providing consistent, individualized feedback to teachers is essential for improving instruction but can be prohibitively resource-intensive in most educational contexts. We develop M-Powering Teachers, an automated tool based on natural language processing to give teachers feedback on their uptake of student contributions, a high-leverage dialogic teaching practice that makes students feel heard. We conduct a randomized controlled trial in an online computer science course (n=1,136 instructors), to evaluate the effectiveness of our tool. We find that M-Powering Teachers improves instructors' uptake of student contributions by 13% and present suggestive evidence that it also improves students' satisfaction with the course and assignment completion. These results demonstrate the promise of M-Powering Teachers to complement existing efforts in teachers' professional development.

Keywords: randomized controlled trial, natural language processing, teaching practices, online learning

*Dorottya (Dora) Demszky (corresponding author, ddemszky@stanford.edu) is an Assistant Professor in Education Data Science at Stanford University. Jing Liu (jliu28@umd.edu) is an Assistant Professor in Education Policy at the University of Maryland and a research affiliate at the IZA Institute of Labor Economics. Heather Hill (heather_hill@gse.harvard.edu) is the Hazen-Nicoli Professor in Education at the Harvard Graduate School of Education. Dan Jurafsky (jurafsky@stanford.edu) is Professor of Linguistics and Professor of Computer Science at Stanford University. Chris Piech (piech@cs.stanford.edu) is Assistant Professor of Computer Science Education at Stanford University. The authors are grateful to Michael Chang, Brahm Capoor, Julie Zelenski and other members of the Code in Place team for their help with implementing the project. They are also thankful to Greg Walton, Betty Malen, Gábor Orosz, János Perczel, Christine Kuzdzal, Kelsey Kinsella, Max Altman, Grace Hu, Sterling Alic and the attendants of seminars at Brown, Google, Cornell and Stanford for their feedback. The project was partially funded by the HAI Hoffman-Yee grant (#203116). Demszky acknowledges support of the Melvin and Joan Lane Stanford Graduate Fellowship.

1 Introduction

Causal evidence suggests that providing teachers formative feedback can improve both their instruction (Kraft et al., 2018) and their students' outcomes (Taylor & Tyler, 2012; Steinberg & Sartain, 2015). Formative feedback is nonevaluative, supportive, timely, and specific, with the intention to modify teachers' thinking or behavior to improve their teaching (Shute, 2008). Yet, the average teacher in the U.S. may have limited access to such feedback. In many schools, the most regular feedback to teachers occurs via principals, particularly following reforms to U.S. teacher evaluation systems in the early 2010s. Teachers often report such feedback as having low utility (Hellrung & Hartig, 2013) and researchers find mixed evidence regarding the efficacy of evaluative feedback on instruction and student outcomes (for a review, see Firestone & Donaldson (2019); Rigby et al. (2017)). Further, only roughly 40% of schools provide teachers access to a math or reading coach (Taie & Goldring, 2017), and some studies suggest that many coaches spend limited time working directly with teachers to improve instruction (Bean et al., 2010; Gibbons & Cobb, 2016; Scott et al., 2012). A major reason is that coaches' roles include a variety of duties, including locating and generating curricula for teachers and facilitating data collection and grade-level team meetings, crowding out time for 1:1 feedback to teachers (Bean et al., 2010; Kane & Rosenquist, 2019; Gibbons & Cobb, 2017).

High-quality formative feedback can thus be effective, but it is likely that few educators experience such feedback on a regular basis. This suggests the need to improve the availability and utility of such feedback. We identify two key challenges in accomplishing this goal using the current system of human observation and feedback. First, generating formative feedback tends to be resource-intensive (Kraft & Gilmour, 2016). Experts in instruction must form relationships with teachers, observe classrooms, prepare comments, and meet to review and reflect with teachers — limiting the number of teachers an individual may serve. Second, the quality of feedback varies. Even the most formal classroom observation rating systems tend to have low rater consistency (Ho & Kane, 2013), and descriptive studies find feedback strongly influenced by the perspective of the observers (Donaldson & Woulfin, 2018). Kraft & Gilmour (2016) also found principal feedback associated with a new teacher evaluation system prone to upward bias (see also (Ho & Kane, 2013)), perhaps as principals sought to avoid conflict, further limiting the utility of feedback as an improvement mechanism (Kraft & Gilmour, 2016). Though feedback quality is best documented in studies of teacher evaluation, it is likely that similar variability in coach feedback exists.

In this study, we address these challenges and show that it is possible to provide useful and effective feedback to teachers via automated tools. Leveraging recent advances in natural language processing (NLP), we developed M-Powering Teachers, a tool to provide automated feedback to teachers on their uptake of student contributions — namely, instances when a teacher acknowledges, revoices, and uses students' ideas as resources in their instruction. We focus on uptake because it is a fundamental teaching skill (Collins, 1982) associated with dialogic instruction (Nystrand et al., 1997; Wells, 1999), whose positive association with student learning and achievement has been widely documented across learning contexts (Brophy, 1984; O'Connor & Michaels, 1993; Nystrand et al., 2000; Wells & Arauz, 2006; Herbel-Eisenmann et al., 2009; Demszky et al., 2021). Improving uptake has proven to be among the most difficult teaching practices to change (Cohen, 2011; Kraft & Hill, 2020) perhaps due to its cognitive complexity (Lampert, 2001). Applying our tool to a practice that has been shown difficult to alter can help demonstrate its potential to improve instruction through providing feedback to teachers.

We employed M-Powering Teachers to provide feedback to 1,136 instructors as part of Code in Place, a five-week free online computer science course organized by Stanford University. This course teaches introduction to programming to ~12k students worldwide, in small sections with a 1:10 teacher-student ratio, all but nine of which use English as the language of instruction (Piech et al., 2021). Three features make Code in Place an ideal setting for our study. First, the instructors in this course are volunteers and many do not have prior experience in teaching. Thus, they are likely more responsive to the automated feedback we provide than experienced teachers who may already know how to uptake student ideas. Second, the instruction took place in an online video conferencing platform, which facilitates the recording of high-quality classroom audio compared to an in-person setting. While our ultimate goal is to implement our feedback tool in in-person classrooms, a virtual context like this serves as a useful first step to test out the feasibility of our approach. Third, as informal teaching settings are now growing in an unprecedented speed, partially due to the Covid-19 pandemic, conducting our study in a virtual context can help contribute to the emerging literature on the efficacy of online teaching.

We provided automated, personalized feedback on each instructor's uptake of student contributions at the end of the week following their teaching session (within 2-4 days). To create variation on checking the feedback, we randomly selected half of the instructors to receive email reminders after the weekly feedback was released. Our results suggest that the email intervention increases treated instructors' likelihood of checking the feedback (i.e., opening the feedback web page) at least once four times and improves their uptake of student contributions by 7% each week compared to the control group. Treatment on the treated analysis shows much larger effects – checking the automated feedback results in a 13% average increase in instructors' uptake of student contributions. We also find that this improvement in uptake is not driven by instructors' simple repetition of student contributions but instead by more sophisticated instructional strategies such as follow-up questioning. Heterogeneity analysis shows that female, returning instructors, and instructors who are not in the U.S. respond more strongly to the feedback than their counterparts. We also find suggestive evidence that instructors' checking the feedback improves students' assignment completion and satisfaction with the course.

1.1 Measuring Teachers' Uptake of Student Contributions

When teachers take up student contributions by, for example, revoicing them, elaborating on them, or asking a follow-up question, they amplify student voices and give students agency in the learning process. Given its documented positive association with student learning and achievement (Brophy, 1984; O'Connor & Michaels, 1993; Nystrand et al., 2000; Wells & Arauz, 2006; Herbel-Eisenmann et al., 2009; Demszky et al., 2021), many scholars consider uptake a core teaching strategy and an important part of classroom observation instruments. Uptake is associated with various discourse strategies (Clark & Schaefer, 1989). In education, especially effective uptake strategies include cases when a teacher follows up on a students' contribution via a question or elaboration (Collins, 1982; Nystrand et al., 1997). Repetition is considered to be a less sophisticated uptake strategy in education, but can still serve as a way for teachers to demonstrate that they are listening to students (Tannen, 1987).

The most widely used classroom observation instruments in the U.S. such as the Framework for Teaching (Danielson, 2007) and CLASS (Pianta et al., 2008) include items that measure uptake. These items, along with many others that capture similarly complex teaching strategies, are coded manually by experts through a cognitively demanding and labor-intensive process. Wells & Arauz (2006) developed an even more fine-grained hierarchical coding scheme for manually evaluating uptake. Although their scheme allows for the measurement of sophisticated uptake patterns, including various sub-categories such as follow-up questions and rejection/acceptance of student contributions, it has as many as 230 code combinations, which makes its use too resource-intensive to scale.

Recent efforts to measure uptake at scale have sought to generate scores for this construct automatically using NLP methods. Samei et al. (2014) and Jensen et al. (2020) use automated classification to detect uptake in elementary English language arts (ELA) and math classrooms. Their approach involved hiring experts to manually code several thousand teacher utterances for uptake, training a machine learning classifier on the annotated utterances, and then applying this classifier to detect uptake in new teacher utterances. Although this approach shows promise, the relationship of their measure to educational outcomes is yet to be explored.

In this work, we use a fully automated measure to identify uptake, one which has been validated using educational outcomes across domains Demszky et al. (2021). This measure, described in greater technical detail in Section 3.1.3, also uses machine learning but it does not require manual annotation because it learns to identify uptake based on turn-taking patterns. The uptake measure captures the extent to which a teacher's response is specific to the student's contribution; that connection serves as evidence that the teacher understands and is building on the student's idea (Clark & Schaefer, 1989). Demszky et al. (2021) find that this measure captures a wide range of uptake strategies, including revoicing, question answering, and elaboration, and that it correlates strongly with expert annotations for uptake (Spearman $\rho = 0.54$, $p < 0.001$). The authors also conducted a cross-domain validation and found that their measure correlates positively with instructional quality and student satisfaction across three different contexts of student-teacher interaction, including elementary math classrooms, small group English Language Arts virtual classrooms, and a text-based math and science tutoring setting.

1.2 Providing Automated Feedback to Teachers

Recent technological advances are giving momentum to a growing number of efforts to build automated feedback tools for educators. Such tools can provide teachers with objective insights on their practice in a scalable and consistent way and thereby offer complementary advantages to expert feedback, which is challenging to scale due to resource constraints and teachers' buy-in to inherently subjective information on their teaching (Kraft et al., 2018).

The majority of automated tools provide teachers with analytics on student engagement and progress and allow teachers to monitor student learning and intervene when needed (Alrajhi et al., 2021; Aslan et al., 2019, among others). Few tools provide teachers with feedback that can serve as a vehicle for self-reflection and instructional improvement. To help address this gap, researchers have developed measures to detect teacher talk moves linked to dialogic instruction, a pedagogical approach that involves students in a collaborative construction of meaning and is characterized by shared control over the key aspects of classroom discourse (Samei et al., 2014; Donnelly et al., 2017; Kelly et al., 2018; Jensen et al., 2020). For example, Kelly et al. (2018) propose an NLP measure trained on human-coded transcripts of live classroom audio to identify the number of authentic questions a teacher asks in her classroom. Moving beyond measurement to teacher feedback, Suresh et al. (2021) introduce the TalkMoves application that provides teachers with information on the extent to which they use dialogic talk moves, including pressing for accuracy and revoicing student ideas. However, their pilot study did not show a statistically significant impact of using TalkMoves on later teacher practice (Jacobs et al., 2022).

1.3 Our Contributions

Building on the aforementioned literature, our work makes two key contributions. First, we are among the first to evaluate the impact of automated feedback on teacher instruction through a large-scale randomized controlled trial. Our study took place in an online, informal teaching setting and it provides evidence that automated feedback can improve instructors' uptake of student ideas – a high-leverage teaching practice that thus far has proven difficult to change. We believe that this study opens up a new strand of inquiry that examines how to best leverage cutting-edge NLP techniques for enhanced instruction and student learning, and lays the foundation for experimenting with this approach in new learning contexts, such as in-person K-12 classrooms.

Second, M-Powering Teachers is reproducible and scalable because it primarily uses open-source software. In an online setting, our tool requires minimal resources because it uses a low-cost automated speech recognition service and a fully automated measure for uptake. Our user interface, developed in consultation with experts in human-computer interaction and educational interventions as well as teachers themselves, is intuitive to use and is non-evaluative. We share the details on the tool and the decisions we made so that researchers and practitioners can readily reproduce, build on and integrate it into their own educational platforms.

Lastly, the specific context of an online, voluntary computer science course closely mimics many emerging teaching settings such as virtual tutoring[1] where instructors tend to be less trained. As a proof of concept, our study demonstrates the potential of using automated feedback to improve teaching practices in virtual classrooms. It also creates avenues for future research to adapt M-Powering Teachers to a wider range of teaching contexts and integrate it into a scalable professional development framework for teachers.

2 Background

We ran the study as part of Code in Place, a 5-week-long, large-scale, free online introductory programming course organized by Stanford University (Piech et al., 2021). The mission of the course is to democratize access to teaching and learning how to code. The course was taught for the first time in Spring 2020 as a response to the COVID pandemic; due to its popularity, it was offered again in Spring 2021, which is when we conducted the experiment.

Instruction primarily took place in OhYay, an online video calling platform. Each week instructors were provided with a link for their own virtual OhYay room for meetings with their students, which occurred between Wednesday-Friday of each week. Instructors also had the option to use a different platform (e.g. Zoom). The course materials were prepared in advance by the course organizers and thus are uniform across different instructors.

The 2021 course recruited 1,136 volunteer instructors from across the globe. Instructors applied for the position by submitting both a programming exercise and a 5-minute video of themselves teaching. Each accepted instructor was assigned to teach a section with 10 students. The sections met weekly for an hour to discuss key topics in the course. We exclude instructors who did not use English in their instruction, instructors who did not use OhYay and who thus did not receive our automated feedback, and those who failed to teach their assigned section, resulting in a total of 918 instructors and 10,794 students. Table 1

[1] `https://www.chalkbeat.org/2022/6/29/23186973/virtual-tutoring-schools-covid-relief-money`

shows the basic demographics of our analytic sample.

[Insert Table 1]

Instructors. Based on the limited demographic information Code in Place has collected, the instructors are diverse in terms of gender, age, and their location while teaching the course. 65% of our instructor sample described themselves as male, 32% as female and 1% as non-binary. Instructors ranged in age from 18–81, with an average of roughly 30 years old. They were located in 82 unique countries with the majority (63%) being in the U.S. 79% were first-time instructors for Code in Place 2021. Based on their open-ended responses about their background, the majority of instructors were young professionals working in the technology industry with limited teaching experience. The rest of the instructors included college students, researchers and former K-12 teachers. The top three motivation for volunteering were to give back through community service, to improve their teaching ability and a love for teaching programming.

Student demographics and assessment. The course enrolled 12,210 students and collected gender, age and location information from them at the time of application. 37% of the students were female and the majority were under the age of 30 (70%).[2] Students were located in 164 unique self-reported countries, with those in India (32%) and the U.S. (30%) accounting for over 60% of the student body.[3].

This course did not administer an end-of-course test to assess student learning, but students did have three optional assignments that were autograded. The first assignment was released on the day of the first section (Wednesday of week 1) and due a week later. The second assignment was released immediately after the due date of the first assignment

[2] Unlike instructor applicants, who were asked to report their specific age, student applicants were asked to select their age ranges.

[3] 3% in Canada, 2% each in Bangladesh, Germany and the UK, 1% each in Nigeria, Turkey, Singapore, Australia, Pakistan, Brazil, Philippines, Japan, Nepal, Russia, Serbia, Kenya, Indonesia, and 16% total in other countries

and due on the Monday of week 3. The third assignment was released immediately after the due date of the second assignment and due on the Friday of week 5.

Online setup and recording. All instructors consented to being recorded when choosing to use OhYay at the time they signed up for the course. Code in Place automatically recorded each section in OhYay. For sections that were offered in a different platform, Code in Place does not have access to the recordings. We thus conduct our study only on sections recorded via OhYay.

3 The M-Powering Teachers Tool

3.1 Workflow for Generating Feedback

Our workflow for generating feedback for instructors is fully automated; it does not require human intervention at any step. Below we explain the details of each step.

3.1.1 Step 1: Recording.

OhYay recorded each class section automatically. We focus on measuring teaching practices in whole class interaction, as it is our primary research interest. Also, in practice, teachers spent on average only 1% of class time in breakout rooms, likely due to the small class size.

3.1.2 Step 2: Transcription and anonymization

We transcribed and algorithmically anonymized recordings using Assembly.ai, a service we chose because of its accuracy, cost-effectiveness ($1 per 1 hr of audio) and ease of use. We separated speakers (also referred to as diarization) by aligning speaker timestamps obtained from OhYay with word-level timestamps obtained from Assembly.ai. To make sure our transcripts do not contain any sensitive data, we anonymized transcripts automatically via Assembly.ai by redacting all words that could potentially refer to people, organizations, locations, phone numbers or credit card numbers. We also replaced all speaker IDs with identifiers such as "Teacher", "Student 1", "Student 2", etc.. One important limitation of this step is that automated speech recognition (ASR) is known to be less accurate for speakers whose native language is not Standard American English (Koenecke et al., 2020), and we do find disparate accuracies in our data as well. However, we have found evidence that the tool does not impact instructors outside the U.S. more negatively – see Appendix A for details. Before scaling up the use of our tool, it is our highest priority to evaluate and address speech recognition issues by leveraging technological improvements in this area.

3.1.3 Step 3: Transcript analysis

We algorithmically analyzed the transcripts to identify various discourse-related phenomena. The core measure of the feedback is **teachers' uptake of student contributions**. We identified teacher uptake using the automated measure described in Demszky et al. (2021). This measure is a machine learning model that is trained on a combination of three large corpora of interactions: (i) the NCTE transcript dataset of elementary math classrooms (Demszky & Hill, 2022), (ii) the Switchboard dataset of phone conversations, widely used in NLP research on dialog (Godfrey et al., 1992) and (iii) a one-on-one math and science text-based tutoring dataset from a company. The model is *unsupervised*: instead of learning from human coding, it learns to distinguish actual student-teacher adjacency pairs (e.g. S: "I added 30 to 70.", T: "Where did the 70 come from?"), from randomly paired student-teacher utterance pairs (e.g. S: "I added 30 to 70.", T: "Please turn to your partner"). Using this simple training objective, the model learns to estimate the extent to which a teacher's response is specific to a students' contribution.

At inference time, the model scores new student-teacher utterance pair between 0 and 1, which can be interpreted as the probability of the teacher utterance being a response to the given student utterance. This probability score is used as an estimate for uptake. For example, if a student says "I added 30 to 70.", "Okay." as a teacher's response would score low on uptake, as it can be a response to many student utterances, and "Where did the 70 come from?" would score high on uptake, since it is specific to the student's contribution. The measure is applicable exclusively to utterance pairs where the student utterance is at least five words long. This is because uptake hinges on the previous contribution to be substantive enough so it can be taken up. We considered a predicted score greater than 0.8 as an example of uptake, a threshold we set as (a) it is close to the center of the binomial distribution of the predictions (in other words, it separates the high vs low uptake examples) and (b) it yielded a precision on par with human agreement (0.62, based on the annotated dataset of Demszky et al. (2021)). As mentioned in the introduction, the measure captures multiple uptake strategies, including repetition, elaboration and follow-up questions and has been extensively validated using data from a range of instructional settings, and proved to have meaningful correlations with student learning outcomes.

We used three additional automated discourse measures to enrich our understanding of changes in instruction relevant to uptake. Given that uptake hinges on students contributing to the classroom discourse, we quantified **teacher talk time** using timestamps from the transcript. We also detected **teacher questions** by relying on question marks and a classifier that we trained to identify questions in the absence of question marks. The question detector can help us identify follow-up questions, which tend to be the best examples of uptake, as they both build on and probe students' ideas. We also captured the extent to which the **teacher repeats student words** using Demszky et al. (2021)'s method who found repetition to be a core component of uptake. The repetition measure computes the percentage of student words that are repeated by the teacher in their subsequent utterance, ignoring stopwords and punctuation. Appendix B provides more details on these measures and their correlation with the uptake measure.

3.1.4 Step 4: Generating the feedback

We display feedback to teachers on a web application, showing them statistics on their uptake, examples of strong uptake from their transcript, and tips for improvement. We also invite teachers to reflect on their instruction and plan for the next lesson. We introduce the design principles and features of the feedback below.

3.2 Design Principles for the Automated Feedback

Our primary objective is to encourage teachers to reflect on their practice, and thereby improve their uptake of student contributions during class sessions. To this end, we designed M-Powering Teachers with several principles in mind and drew on insights from experts and relevant literature in education, social psychology and human computer interaction.

We provided non-judgmental information about teachers' instruction in a way that respects their agency and authority over their practice (Wills & Haymore Sandholtz, 2009; Priestley et al., 2015; Oolbekkink-Marchand et al., 2017). Specifically, we conveyed the feedback privately to each teacher, and explicitly stated that the feedback is not used to evaluate them, but rather to support their professional development. We also included open-ended reflection questions to elicit teachers' own interpretation of the statistics and examples and to encourage them to give advice to themselves, following the "saying is believing" principle (Higgins & Rholes, 1978) widely recognized in social psychology.

Second, we took several steps to make the feedback concise, specific and actionable. With only one page of information, we used figures to visualize high-level statistics on their frequency of taking up student ideas and on student talk time. To substantiate these statistics and encourage teachers to reflect on their instruction, we highlighted examples of uptake from their transcript and asked teachers to reflect on the strategies they used in these examples. To help teachers see how their practice evolves over time and set goals for themselves, we included tabs that allowed them to revisit their feedback from earlier class sessions. We also provided advice on and examples of uptake as well as links to further resources including papers and blog posts on uptake and dialogic instruction.

Finally and most importantly, we delivered the feedback in a timely and regular manner. To ensure that teachers still had a fresh memory of what they did and to make the feedback more relevant and exciting (Shute, 2008), we shared feedback with teachers within 2-4 days after their class sessions, and always before their next class. We delivered feedback to teachers after each recorded class, with hopes that sustained work in this area would lead to improved practice over time.

3.3 User Interface of the Feedback Application

[Insert Figure 1]

[Insert Figure 2]

Figures 1 and 2 show the components of the one-page feedback application. On the top of the page, a brief paragraph introduces the feedback to users, emphasizing that the feedback is private and the goal of it is to support the user's professional development. Then, users can see statistics about talk time, and examples from their transcript when their questions elicited a long student utterance. Below that, users can see the number of uptakes (i.e., examples when they built on student contributions) and examples from their transcript identified by our algorithm. As we noticed that the best examples of uptake occur in the context of a teacher asking a follow-up question, we show and count teachers' uptake examples that co-occur with the teacher asking a question. We also provide an input box for users to reflect on these examples and plan for the next session. At the bottom of the page, we share resources, including blog posts and papers on dialogic instructional practices. Finally, we provide the entire transcript to users for review.

4 Randomized Controlled Trial

We conducted a randomized controlled trial to evaluate the effectiveness of the M-Powering Teachers tool. The key idea of our study design is to generate an exogenous variation of checking the feedback, by sending email reminders to a random group of instructors. For ethical reasons, we offered all instructors access to the feedback through a link on the course website. However, the link to the feedback was in an inconspicuous place, listed among many other teaching-related resources, and hence we expected most instructors would not check the feedback unless they received our email reminder.[4]

[Insert Figure 3]

Before the start of the course, we randomly assigned half of the instructors to treatment (n=568) and the other half to control (n=568) groups. We sent instructors in the treatment group a weekly email reminder about the feedback, resulting in a total of five reminders. The instructors in the control group did not receive such emails. In order to ensure that the intervention effect is mediated by the content of the automated feedback rather than the content of the email, we made the email short and generic (Figure 3), with only a link to the feedback and two non-personalized sentences encouraging instructors to follow the link. Our system logged whether an instructor opened the feedback page in their browser, which we used as a binary variable to measure whether the teacher checked the feedback.

[Insert Figure 4]

Figure 4 shows the timeline of the intervention in relation to the course sections and the three assignments administered in the course. Sections took place between Monday-Wednesday of each week, and we sent the email reminders on the Sunday of each week.

[4]We do not have evidence for spillover effects. Since instructors were located across the world, their primary way to communicate was through the course forum. We moderated the forum by making all instructor posts about the automated feedback private, visible only to the course organizers. We also asked course organizers to not advertise the automated feedback to instructors. We took these steps to prevent advertisement about the automated feedback to control group instructors.

4.1 Measures of Outcomes

Teaching practices. As discussed above, we use the transcripts that are generated automatically based on section recordings from OhYay to measure and track instructors' uptake of student contributions.[5] We conduct a descriptive analysis to show the predictors and the variance components of uptake using pre-intervention data and data from the control group — see Appendix C for details.

Besides uptake, we also track other discourse features correlated with uptake, including the number of questions asked by an instructor, the number of times an instructor repeats students' utterances, and instructors' talk time. We use these three measures as additional outcome variables to provide some evidence on what instructional strategies drive the changes we see in instructors' use of uptake. See Section 3.1.3 for details on how we measure them.

Assignment completion. We use the percentage of questions completed in each assignment as our key outcome metric. We only use data from assignments 2 and 3 because the first assignment was due between the first and the second class section, which means that our feedback to instructors could not have yet affected the completion rate of the the first assignment. The choice of outcome metric (whether the assignment was attempted, whether the assignment was fully completed, etc.) does not significantly affect the results. Based on this metric, the average completion rates are 54% for assignment 2 (SD=48%) and 34% for assignment 3 (SD=47%). The relatively low completion rates are likely explained by the fact that this is a free online course and the assignments are optional.

Endline survey to instructors and students. We incentivized a randomly selected group of 200 instructors to fill out a short survey about the feedback tool. The survey asked instructors to report their perception of the tool, the effects this tool had on their

[5] We removed recordings shorter than 30 minutes to ensure that our sample only includes transcripts where meaningful instruction took place. Recordings shorter than 30 minutes usually indicate technical issues. As a result, our analytic sample consists of a total of 4,056 section recordings with an average duration of 64 minutes.

teaching and suggestions for improving the tool. We include the survey in Appendix D. Instructors were sampled irrespective of treatment status, received up to three reminders and were incentivized with a chance to win one of ten $40 Amazon gift cards. The survey achieved a 71% response rate (n=142), which does not differ by treatment group (p=0.303).

Code in Place also administered a short survey to all students (16% response rate, n=1,958). The survey asked students to indicate how likely they are to recommend the course to friends and how helpful different elements of the course were, including sections, assignments, course forum, etc. The lack of reminders and incentive explains the low response rate for the student survey. We include the survey in Appendix E. We constructed two measures from the survey as outcomes for our analyses: a binary indicator on whether a student responded to the survey and students' raw ratings of their likelihood to recommend the course to others on a 1-10 scale.[6] All survey data were de-identified before analysis and linked through anonymous research IDs.

4.2 Validating Randomization

To verify whether our randomization was successful, we evaluate whether the demographics of instructors in the treatment and control groups differ statistically. We also compare instructors' discourse features measured in their first class session, prior to receiving feedback. As Table 2 shows, other than average instructor age we do not find statistically significant differences between conditions in any of the instructor demographics and discourse features of the first section. The joint significance test that considers all these baseline variables shows a F statistic of 0.81, failing to reject balance between the two conditions. This analysis validates our randomization and suggests that any differences we observe later in the course are likely due to the effects of the intervention.

We also conduct an attrition analysis to examine whether instructors exhibited differential

[6]The results are very similar if we use students' ratings of how helpful the sections are so we omit them in our main analysis.

attrition patterns between the two study arms. Attrition can be caused by multiple factors — instructors might be using a different platform instead of OhYay (e.g. Zoom) or dropped out of the course; we do not have information to identify the cause behind a missing recording.[7] To formalize the attrition analysis, we regress a binary variable that indicates whether we are able to observe an instructor teaching in a particular week on the treatment status and control for instructor characteristics. Results in Appendix F Table A2 suggest that other than a marginally significant coefficient on the treatment status in week 2, there is no evidence that instructors attrited differently in the treatment and control groups across the span of the course.

[Insert Table 2]

5 Empirical Strategy

We use the exogenous variation generated from our randomized email intervention to estimate the impact of checking the NLP-based automated feedback on teaching practices and student outcomes. As the feedback is provided on a weekly basis and the course is five weeks long, we can observe how teaching practices evolve from week two to week five. However, given that the randomization was conducted at the individual level rather than at the individual-by-week level, whether an instructor changed their behavior in a given week may be affected by random assignment not only through whether they checked the feedback in that week but also through whether they checked the feedback in prior weeks. Thus, to account for the longitudinal nature of our experiment, we define our primary independent variable of interest as *checking the NLP-based feedback at least once* prior to the instructor's subsequent section. Specifically, we estimate the following two-stage least squares (2SLS)

[7] The Code in Place team did not document the cause of missing recordings but they suspect that the majority of them are caused by an instuctor switching to Zoom or another platform. If an instructor did not show up to teach, the organizers did their best to find a substitute instructor. In cases when they weren't able to find substitutes, they would share a recorded section by another instructors with the students from the same week. However, we do not have the recordings for the substituted sections, nor do we know if the section had substitutes.

estimator:

$$Feedback_{it} = \pi_0 + \pi_1 T_i + \pi_2 X_i + \epsilon_{it} \tag{1}$$

$$Y_{it} = \beta_0 + \beta_1 \widehat{Feedback}_{it} + \beta_2 X_i + \mu_{it} \tag{2}$$

where i indicates instructors and t indicates an instructional week, which takes the value of 2, 3, 4, and 5. In Equation (1), we model whether instructor i opened the feedback page prior to their subsequent section at least once up to a given week t as a function of the treatment status (T_i) and a series of time-invariant covariates (X_i). These covariates include instructor demographics (female, age, age^2, in the U.S., first-time CiP instructor), pre-intervention discourse features (number of uptakes per hour, number of questions per hour, number of repetitions per hour, teacher talk time proportion), and classroom demographics (proportion of female students, proportion of students in the U.S., proportion of students in each age group listed in Table 1). We then use the predicted value for checking the feedback at least once up to week t as the independent variable in the second stage and estimate Equation (2). β_1 is our parameter of interest that captures the local average treatment effects of our intervention. We consider several outcomes (Y_{it}) to capture various aspects of instructor behavioral changes: the number of uptakes per hour is our primary outcome as it is what the intervention is designed for, but we also consider the number of questions asked per hour, the number of repetitions per hour, and percentage of talk time to further examine the mechanisms of change. To further verify that our randomization is successful, we also estimate a version of the model without any covariates.

We estimate the model first by pooling together all the weeks and then by each week to examine how instructors' responses to the feedback evolve over time. We further conduct heterogeneity analysis by instructor gender, whether they are first-time instructors in Code in Place, whether they are in the U.S., and whether they demonstrated high or low uptake in their first week of instruction. Lastly, we estimate how instructors' checking the feedback affects student assignment completion, class attendance, whether they respond to the endline survey, and their satisfaction of the course. To do this, we can no longer conduct the analysis at the weekly level as we only observe student outcomes at the end of the course. We thus use whether an instructor checked the feedback, prior to their subsequent section, at least once during the five weeks of teaching as the primary independent variable and conduct the analysis at the student level.

6 Results

6.1 First Stages

We present results from the first stages in Table 3. The first column shows estimates based on Equation (1) for the entire sample and the other columns show estimates for each week. We also report the percent of instructors in the control group who opened the feedback page prior to their subsequent section at least once up to week t so we can properly interpret the effect sizes of our intervention. Overall, our first stages are quite strong, with F statistics above 34 when using the entire sample and above 17 when using data from each week.

We find that our email reminder successfully improves treated instructors' likelihood of opening the feedback page. Across all instruction weeks, the email reminder increases treated instructors' likelihood of checking the feedback at least once to 71.2%, four times the rate in the control group (17.6%). It appears that the intervention has the strongest effect in week 2 (i.e., after the first email reminder). While the coefficients get bigger over time, the incremental change is at a smaller margin. Specifically, the first email reminder increases treated instructors' likelihood of interacting with the feedback 4 times more compared to the control group. Namely, nearly 61.4% of all treated instructors have interacted with the feedback at this point. In later weeks, the ratio of the treatment and control group's likelihood of checking the feedback decreases to 3 (week 3), 2.7 (week 4) and 2.8 (week 5). This is understandable, as over time, fewer and fewer instructors in each group are left in the category that has not interacted with the feedback at all. We also find that instructors who are older and those who are outside of the U.S. are more likely to interact with the feedback.

[Insert Table 3]

6.2 Impact on Instructors' Uptake of Student Contributions

In Table 4, for comparison purposes, we report results from both the intent to treat (ITT) and TOT analyses. We also run the analyses for all the four outcomes of teaching practices, including uptake, questions, repetition, and talk time, to probe both the overall effects on uptake and the associated discourse features that might be changed due to the feedback we provided to instructors.

[Insert Table 4]

The ITT results, which are reported in Panel A of Table 4, suggest that our intervention improved instructors' use of uptake. On average, treated instructors increased their use of uptake by 0.60 times per hour of instruction ($p < 0.05$), which is about 7% of the magnitude of the control mean on uptake (8.58). We also find that treated instructors significantly increased their use of questioning, by 1.70 times per hour (6% of control mean). This is likely because teachers are asking more follow-up questions as a strategy to take up student ideas. In contrast, we do not observe any significant effects on instructors repeating student language or decreasing their own talk time. Overall, the ITT results provide suggestive evidence on how our intervention, a simple weekly email reminder that encourages instructors to check the feedback page, is able to improve their teaching practices.

The TOT analysis answers the question of how checking the feedback changes instructors' teaching behavior and is of more policy relevance. We report the results in Panel B of Table 4. Not surprisingly, the effect sizes are much bigger compared to those in the ITT analysis. Specifically, instructors who were induced to check the feedback page at least once by our randomized email reminders improved their use of uptake by 1.13 times per hour (13.2%, $p < 0.05$). Similarly, we find that instructors who checked the feedback asked roughly 3.20 (11.4%) more questions per class ($p < 0.05$), but did not repeat student contributions more frequently nor did they talk less. These results, along with the ITT ones, suggest that the improvement in uptake is driven primarily by more sophisticated strategies such as increased questioning rather than repetition or talk time. We also replicate all these results without any controls other than than the binary weekly indicators in Appendix Table A3. All the coefficients are very close to those in Table 4 but have slightly larger standard errors, providing further evidence that our randomization was done successfully and the control variables only improve the precision of our inferences.

To understand how instructors' responses to the feedback evolve over time, we also run the TOT analysis for each week. The results are reported in Table 5. We find that it takes some time for instructors to utilize the feedback and improve their instructional strategies. While our first stage analysis (Table 3) shows that more than four times as many treated instructors checked the feedback after our first email reminder compared to the control group, the feedback did not immediately lead to any changes in the four discourse features we examine. In fact, the most significant instructional changes took place in week 3, with coefficient sizes close to double those of the second week for both uptake and questioning ($p < 0.05$). While there is a marginally significant coefficient for repetition, we also observe a drop of instructors' talk time (4.9%, $p < 0.01$). In week 4 and 5, the coefficients for the uptake outcome decrease while remaining statistically significant at the conventional level, suggesting our intervention still improves instructors' use of uptake but not as strongly as week 3. While we still see a significant and positive effect on questioning in week 4, all the coefficients are no longer statistically significant for other outcome measures during the last two weeks.

[Insert Table 5]

6.3 Heterogeneity Analysis

Instructors from different backgrounds or with different characteristics might respond to the feedback differently. We thus conduct heterogeneity analysis by gender, teaching experience with Code in Place, whether they are based in the U.S., and whether they demonstrated high or low uptake in their first week of instruction. The results are shown in Table 6.

[Insert Table 6]

While female instructors increase the number of times they take up student ideas slightly more as a result of the feedback compared to males, the coefficients for both groups are marginally significant and the differences are small. We find more pronounced variability by teaching experience and location. Returning instructors in Code in Place and those who are not based in the U.S. increased their uptake of student contributions by roughly 2 instances per hour; three to four times as much as their counterparts whose coefficients are below 1 and are statistically insignificant. We see similar patterns for the use of questions. Instructors who are outside the U.S. also significantly increased their use of repetition and reduced their overall talk time, suggesting that these instructors adopted more than one strategy to improve their performance on uptake and were more amenable to changes. Due to our limited data on instructors' background, we are not able to further pinpoint why non-U.S. instructors are so responsive to the automated feedback. One possible explanation is that non-U.S. instructors have more motivation to learn from the course, as they volunteered to teach a course organized by another country and to teach in a language that may not be their mother tongue.

Interestingly, we also see some suggestive evidence that instructors who exhibited more uptake in the first week of instruction benefit more from our feedback than their counterparts. One conjecture we have is that it might be easier for instructors who already use uptake to some extent (so they have some of the uptake skills) to improve their teaching practices with the help of feedback. Alternatively, it is possible that the information we provided to instructors who were initially low on uptake somewhat discouraging (e.g., one might receive feedback showing they demonstrated zero uptake in week 1) so the feedback did not achieve the expected positive benefits. It will be valuable to investigate how to generate positive effects regardless of an instructor's initial use of uptake in future studies.

6.4 Impact on Student Learning Outcomes and Satisfaction

So far, we have provided evidence on how the automated feedback can improve instructors' uptake of student ideas. However, it is unclear whether this instructional improvement can translate to student learning gains. Since Code in Place did not administer an end-of-course test to the students, we use their assignment completion and survey data to provide suggestive evidence on student learning and satisfaction. We fit the same 2SLS models as discussed before, using student level data. We report the results in Table 7.

TOT estimates suggest that instructors' checking the feedback at least once increased students' completion of the second assignment by 6.6% compared to the control mean ($p < 0.10$). There is no significant change for the third assignment. This is partially explained by the fact that the last assignment was distributed toward the end of the course and students overall had low motivation to finish it. In fact, students taught by the control-group instructors on average finished 52.9% of the second assignment, but this number is only 33.3% for the third assignment. We also do not find evidence that the feedback increased student proportion of classes attended.

There is suggestive evidence that the feedback improved students' course satisfaction. In Column (4), we find that instructors' checking the feedback significantly improved their students' likelihood to respond to the survey by 20% compared to the control mean ($p < 0.05$). While we do not see statistically significant result for course ratings for students who responded to the survey (Column (5)), this is likely driven by the fact that students who responded to the survey were also the ones who were most satisfied with it. As Appendix Figure A5 shows, 96% of student respondents rated the course at 8 or above out of a scale of 10, and 99% rated the course 7 or a above.[8] Overall, while our data on student learning outcomes and satisfaction are not as rich as we would hope and the survey results suffer from the overall low response rate, they provide some evidence on how teaching practices induced by the feedback have the potential to improve student outcomes.

[Insert Table 7]

6.5 Instructor Feedback

Since the instructor feedback is self-reported and we only administered our survey to a random sample of 200 instructors due to limited resources, it constitutes a weaker outcome than the analyses above. That being said, the survey responses do indicate that the feedback had many positive benefits for instructors. Simple descriptions of the survey responses suggest that the majority of instructors who checked the feedback found the feedback helpful and reported that it generated insights into their teaching and helped them become a better teacher (for details, see Appendix J). These findings provide suggestive evidence that the automated feedback enhanced teachers' self-efficacy, which might have contributed to the positive student outcomes we observe. Follow-up studies may help better understand the relationship among the email reminders, instructors' perception of the feedback and student outcomes.

7 Discussion

Our study investigated whether it is possible to effectively deliver feedback to teachers at scale using automated tools. We developed M-Powering Teachers, a fully automated tool to provide feedback to teachers on their uptake of student contributions, one of the most

[8] We do not have a reason to believe that these differences are due to instructors in the treatment group directly telling students to respond to the survey, since instructors were not aware of the intervention and most of them were also not aware of student endline surveys. Thus, we can reasonably assume that these differences are due to an indirect effect of teaching practice on student satisfaction.

important discourse phenomena associated with dialogic instruction, and tested the effectiveness of this tool in a large-scale online programming course. In doing so, we demonstrated that feedback on instruction, typically a labor-intensive process and one that is unavailable to many teachers, can be delivered widely and can stimulate improvements in instructional practice. Importantly, scale does not come at the cost of efficacy: our effect sizes are similar to or greater than those obtained in other professional learning interventions (e.g. Kraft et al., 2018; Gonzalez et al., 2022).

We found that the automated teaching insights in our tool increased instructors' uptake of student contributions by 13%, a result likely driven by instructors' increased use of more sophisticated strategies beyond repetition, such as follow-up questioning. There is also suggestive evidence that students whose teachers looked at the feedback completed a greater percentage of their second assignment and were more satisfied with the course. Finally, the majority of instructors found the feedback helpful. These results together suggest that M-Powering Teachers has a positive impact on instruction.

The success of this intervention suggests four avenues for future work. One is extending M-Powering Teachers to capture other teaching strategies – for instance, using models to parse and provide feedback on teachers' questioning strategies (Alic et al., 2022), use of academic language, or equity-focused talk moves (Wilson et al., 2019). Once we have a set of robust classroom indicators, we can design more robust feedback systems based on teachers' strengths and areas for improvement. A second avenue is extending feedback to new platforms and settings within the online learning sector. At least two states incentivize online course completion prior to high school graduation (Georgia[9] and Florida[10]), and the number of open online courses and degree programs continue to grow. Automated feedback in these settings is simple to implement and relatively easy to study.

A third avenue would take advantage of these research opportunities to gain insight into how feedback can best be crafted to elicit teachers' attention and behavioral change. Qual-

[9] `https://www.legis.ga.gov/api/legislation/document/20112012/127888`

[10] `https://www.fldoe.org/core/fileparse.php/5606/urlt/Virtual-Sept.pdf`

itative studies of teacher perceptions of and actions in response to automated feedback can help prioritize and shape later experimental A/B tests of feedback that varies in tone (e.g., largely positive vs. positive + constructive), in referents (e.g., prior personal performance as the reference vs. a comparison to other teachers), and calls to action (e.g., asking teachers to formulate their own plans for change vs. asking teachers to take up expert-recommended strategies). Such studies could also test other constructs thought to be critical ingredients in adult learning, for instance teacher agency, the personalization of feedback, or social accountability for change.

A fourth avenue involves extending M-Powering Teachers to the K-12 public school sector. Several factors suggest this technology may gain a foothold in public schools. First, the feedback is very low cost, at $1 per session once fixed costs of system setup are paid. Second, automated feedback can occur in settings where coaches are not present and where principals do not have the time or inclination to provide high-quality evaluative feedback. Third, the privacy associated with such feedback may also engage teachers who are hesitant to work with coaches, or who already perceive their instruction to be satisfactory.

However, we think it unlikely that the effects we observed in this experiment would translate directly to K-12 schools without significant additional supports. Code in Place employed mostly novice, all-volunteer instructors; these instructors likely had few other resources for improving their instruction and lots of room to improve. K-12 teachers, by contrast, often have well-established classroom interaction patterns, many opportunities to improve their craft, and some already use highly interactive instructional methods. Further, whereas instruction is seamlessly recorded in online settings, classroom recordings require the setup of recording devices and the upload of files to the cloud, extra tasks that teachers may not want to engage in during the course of their busy workday. Further, teacher and student talk may not be audible if recorded on typical handheld devices (e.g., phones or tablets), and automated speech recognition software may thus fail to generate transcripts usable in NLP analyses. Solving these problems encompasses advances in ASR technology as well as advances in making automated feedback both appealing to and easily used by teachers.

Before this technology can work at scale, several other issues must be resolved. At a high level, we need to create oversight mechanisms for the ethical development, evaluation and use of automated teacher feedback technologies (Madaio et al., 2020, 2022). Teachers and other educators should play an integral role in ethical tool design and evaluation, but we know of no active efforts to set standards and guidelines for the use of this technology in schools. This need is particularly acute in the area of teacher and student privacy, where, in the extreme, the possibility exists for the constant monitoring of classrooms as well as the use of classroom data for marketing purposes. There are also concrete technical issues that we need to address: ASR is less accurate for noisy classroom audio and for speakers whose native language is not Standard American English, and differences in accuracy across these linguistic groups may continue to propagate inequities in teachers' professional development and students' opportunities to learn. Thus, we need to improve and carefully evaluate ASR tools, as well as all other natural language processing methods that build upon it, to make our tool robust and fair (Kizilcec & Lee, 2022).

Despite its limitations, this study constitutes an important step towards our ultimate goal of developing an effective, scalable feedback tool for all teachers. With the development of new NLP-based measures of instruction, we can extend our tool to generate insights on multiple aspects of teaching (Liu & Cohen, 2021). Future efforts should continue to improve, validate and apply M-Powering Teachers to explore its full potential to support teaching and improve student learning outcomes across educational contexts.

References

Alic, S., Demszky, D., Mancenido, Z., Liu, J., Hill, H., & Jurafsky, D. (2022). Computationally identifying funneling and focusing questions in classroom discourse. *BEA 2022*, 224.

Alrajhi, L., Alamri, A., Pereira, F. D., & Cristea, A. I. (2021). Urgency analysis of learners' comments: An automated intervention priority model for mooc. In *International conference on intelligent tutoring systems* (p. 148-160).

Aslan, S., Alyuz, N., Tanriover, C., Mete, S. E., Okur, E., D'Mello, S. K., & Arslan Esme, A. (2019). Investigating the impact of a real-time. In I. Proceedings (Ed.), *multimodal student engagement analytics technology in authentic classrooms* (p. 1-12). of the 2019 CHI conference on human factors in computing systems.

Bean, R. M., Draper, J. A., Hall, V., Vandermolen, J., & Zigmond, N. (2010). Coaches and coaching in reading first schools: A reality check. *The Elementary School Journal*, *111*(1), 87–114.

Bird, S. (2006). Nltk: the natural language toolkit. In *Proceedings of the coling/acl 2006 interactive presentation sessions* (pp. 69–72).

Brophy, J. E. (1984). *Teacher behavior and student achievement (no. 73).* Michigan State University: Institute for Research on Teaching.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive science*, *13*(2), 259-294.

Cohen, D. K. (2011). *Teaching and its predicaments.* Harvard University Press.

Collins, J. (1982). Discourse style, classroom interaction and differential treatment. *Journal of Reading Behavior*, *14*, 429-437.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching.* ASCD.

Demszky, D., & Hill, H. (2022). The ncte transcripts: A dataset of elementary math classroom transcripts. *arXiv preprint arXiv:2211.11772*.

Demszky, D., Liu, J., Mancenido, Z., Cohen, J., Hill, H., Jurafsky, D., & Hashimoto, T. (2021). Measuring conversational uptake: A case study on student-teacher interactions. In *Proceedings of the 59th annual meeting of the association for computational linguistics (acl-ijcnlp).*

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Donaldson, M. L., & Woulfin, S. (2018). From tinkering to going "rogue": How principals use agency when enacting new teacher evaluation systems. *Educational Evaluation and Policy Analysis*, *40*(4), 531–556.

Donnelly, P. J., Blanchard, N., Olney, A. M., Kelly, S., Nystrand, M., & D'Mello, S. K. (2017). Words matter: Automatic detection of teacher questions in live classroom discourse using linguistics, acoustics and context. 218–227: Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17.

Firestone, W. A., & Donaldson, M. L. (2019). Teacher evaluation as data use: What recent research suggests. *Educational Assessment, Evaluation and Accountability*, *31*(3), 289–314.

Gibbons, L. K., & Cobb, P. (2016). Content-focused coaching: Five key practices. *The Elementary School Journal*, *117*(2), 237–260.

Gibbons, L. K., & Cobb, P. (2017). Focusing on teacher learning opportunities to identify potentially productive coaching activities. *Journal of teacher education*, *68*(4), 411–425.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on* (Vol. 1, pp. 517–520).

Gonzalez, K., Lynch, K., & Hill, H. C. (2022). *A meta-analysis of the experimental evidence linking stem classroom interventions to teacher knowledge, classroom instruction, and student achievement.* EdWorkingPaper.

Hellrung, K., & Hartig, J. (2013). Understanding and using feedback–a review of empirical studies concerning feedback from external evaluations to teachers. *Educational Research Review*, *9*, 174–190.

Herbel-Eisenmann, B., Drake, C., & Cirillo, M. (2009). "muddying the clear waters": Teachers' take-up of the linguistic idea of revoicing. *Teaching and Teacher Education*, *25*(2), 268-277.

Higgins, E. T., & Rholes, W. S. (1978). "saying is believing": Effects of message modification on memory and liking for the person described. *Journal of Experimental Social Psychology*, *14*(4), 363-378.

Ho, A. D., & Kane, T. J. (2013). The reliability of classroom observations by school personnel. research paper. met project. *Bill & Melinda Gates Foundation.*

Jacobs, J., Scornavacco, K., Harty, C., Suresh, A., Lai, V., & Sumner, T. (2022). Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, *112*, 103631.

Jensen, E., Dale, M., Donnelly, P. J., Stone, C., Kelly, S., Godley, A., & D'Mello, S. K. (2020). Toward automated feedback on teacher discourse to enhance teacher learning. In *Proceedings of the 2020 chi conference on human factors in computing systems.* 1–13.

Kane, B. D., & Rosenquist, B. (2019). Relationships between instructional coaches' time use and district-and school-level policies and expectations. *American Educational Research Journal*, *56*(5), 1718–1768.

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, *47*, 7. Retrieved from `https://doi.org/10.3102/0013189X18785613`

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In *The ethics of artificial intelligence in education* (pp. 174–202). Routledge.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., ... Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, *117*(14), 7684–7689.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88(4)*, 547–588. Retrieved from `https://doi.org/10.3102/0034654318759268`

Kraft, M. A., & Gilmour, A. F. (2016). Can principals promote teacher development as evaluators? a case study of principals' views and experiences. *Educational Administration Quarterly*, *52*(5), 711–753.

Kraft, M. A., & Hill, H. C. (2020). Developing ambitious mathematics instruction through web-based coaching: A randomized field trial. *American Educational Research Journal*, *57*(6), 2378-2414.

Lampert, M. (2001). *Teaching problems and the problems of teaching*. Yale University Press.

Liu, J., & Cohen, J. (2021). Measuring teaching practices at scale: A novel application of text-as-data methods. *Educational Evaluation and Policy Analysis, 0*, *1623737211009267*.

Madaio, M., Blodgett, S. L., Mayfield, E., & Dixon-Román, E. (2022). Beyond "fairness": Structural (in) justice lenses on ai for education. In *The ethics of artificial intelligence in education* (pp. 203–239). Routledge.

Madaio, M., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–14).

Nystrand, M., Gamoran, A., Kachur, R., & Prendergast, C. (1997). *Opening dialogue*. New York: Teachers College Press.

Nystrand, M., Wu, L. L., Gamoran, A., Zeiser, S., & Long, D. (2000). *Questions in time: Investigating the structure and dynamics of unfolding classroom discourse.* National Research Center on English Learning and Achievement (CELA).The University of Wisconsin-Madison.

O'Connor, M. C., & Michaels, S. (1993). Aligning academic task and participation status through revoicing: Analysis of a classroom discourse strategy. *Anthropology & Education Quarterly*, 318-335.

Oolbekkink-Marchand, H. W., Hadar, L. L., Smith, K., Helleve, I., & Ulvik, M. (2017). Teachers' perceived professional space and their agency. *Teaching and teacher education*, *62*, 37-46.

Pianta, R. C., Paro, L., M., K., & Hamre, B. K. (2008). *Classroom assessment scoring system™: Manual k-3*. Paul H Brookes Publishing.

Piech, C., Malik, A., Jue, K., & Sahami, M. (2021). Code in place: Online section leading for scalable human-centered learning. In *Proceedings of the 52nd acm technical symposium on computer science education* (p. 973-979).

Priestley, M., Biesta, G. J. J., Philippou, S., & Robinson, S. (2015). The teacher and the curriculum: Exploring teacher agency. *The SAGE handbook of curriculum, pedagogy and assessment*, 187-201.

Rigby, J. G., Larbi-Cherif, A., Rosenquist, B. A., Sharpe, C. J., Cobb, P., & Smith, T. (2017). Administrator observation and feedback: Does it lead toward improvement in inquiry-oriented math instruction? *Educational Administration Quarterly*, *53*(3), 475–516.

Samei, B., Olney, A. M., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N., ... Graesser, A. (2014). *Domain independent assessment of dialogic properties of classroom discourse.* Retrieved from `https://eric.ed.gov/?id=ED566380`

Scott, S. E., Cortina, K. S., & Carlisle, J. F. (2012). Understanding coach-based professional development in reading first: How do coaches spend their time and how do teachers perceive coaches' work? *Literacy research and instruction*, *51*(1), 68–85.

Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, *78*(1), 153-189.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? experimental evidence from chicago's excellence in teaching project. *Education Finance and Policy*, *10*(4), 535-572.

Suresh, A., Jacobs, J., Lai, V., Tan, C., Ward, W., Martin, J. H., & Sumner, T. (2021). *Using transformers to provide teachers with personalized feedback on their classroom discourse: The talkmoves application. arxiv.* (preprint)

Taie, S., & Goldring, R. (2017). Characteristics of public elementary and secondary school teachers in the united states: Results from the 2015-16 national teacher and principal survey. first look. nces 2017-072. *National Center for Education Statistics*.

Tannen, D. (1987). Repetition in conversation: Toward a poetics of talk. *Language*, 574-605.

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, *102*(7), 3628-51.

Wells, G. (1999). *Dialogic inquiry: Towards a socio-cultural practice and theory of education.* Cambridge University Press.

Wells, G., & Arauz, R. M. (2006). Dialogue in the classroom. *The journal of the learning sciences*, *15*(3), 379-428.

Wills, J. S., & Haymore Sandholtz, J. (2009). Constrained professionalism: Dilemmas of teaching in the face of test-based accountability. *Teachers college record*, *111*(4), 1065-1114.

Wilson, J., Nazemi, M., Jackson, K., & Wilhelm, A. G. (2019). Investigating teaching in conceptually oriented mathematics classrooms characterized by african american student success. *Journal for Research in Mathematics Education*, *50*(4), 362–400.

Author Bios

Figures

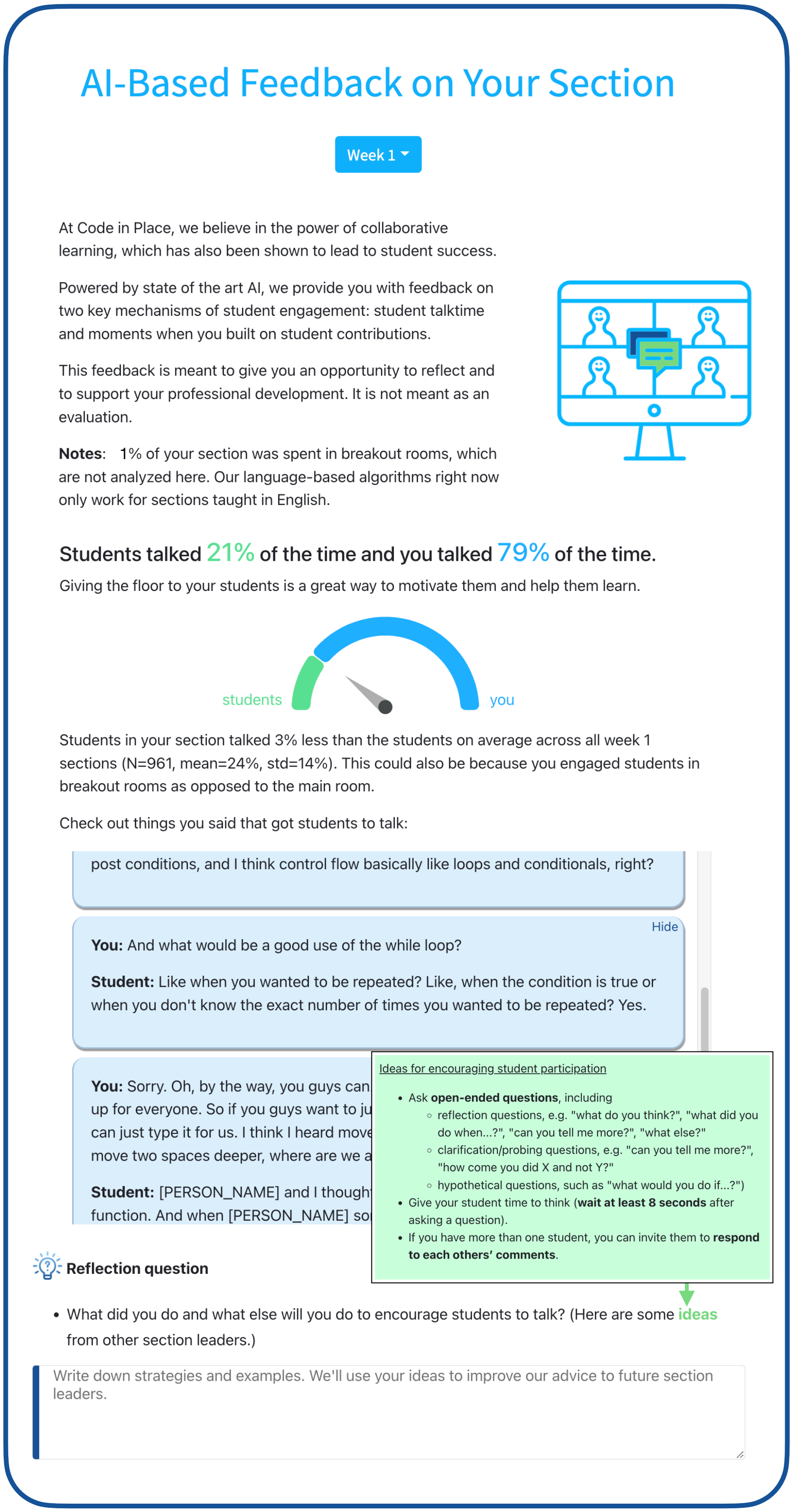


Figure 1: Components of the M-Powering Teachers Web Application (Part 1)

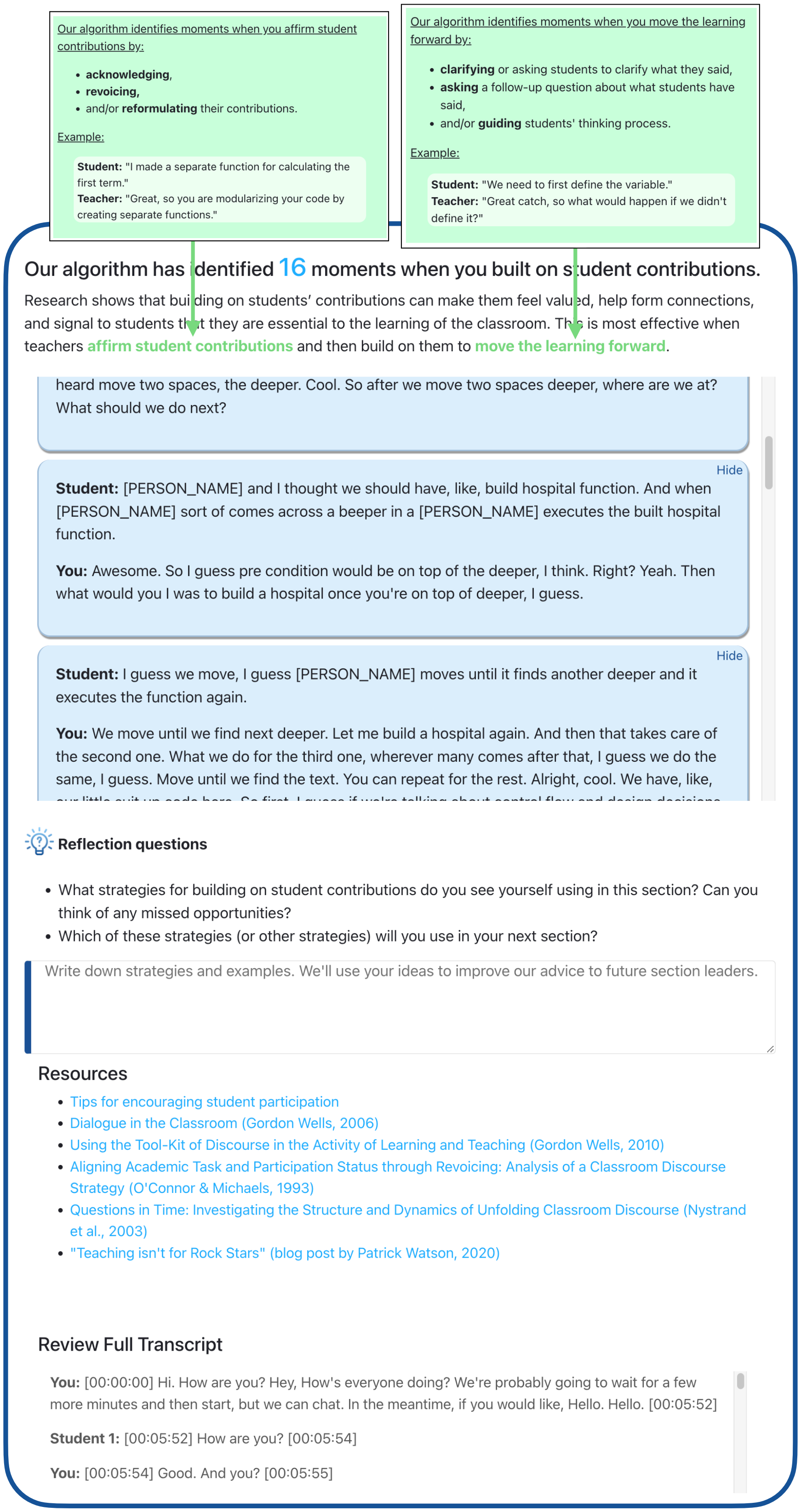


Figure 2: Components of the M-Powering Teachers Web Application (Part 2)

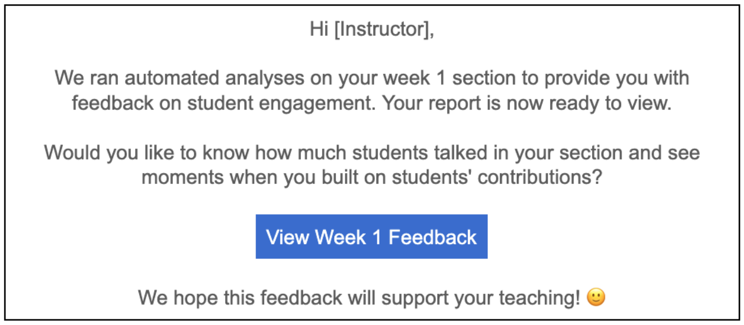


Figure 3: Generic Email Encouraging Instructors to Check the Feedback

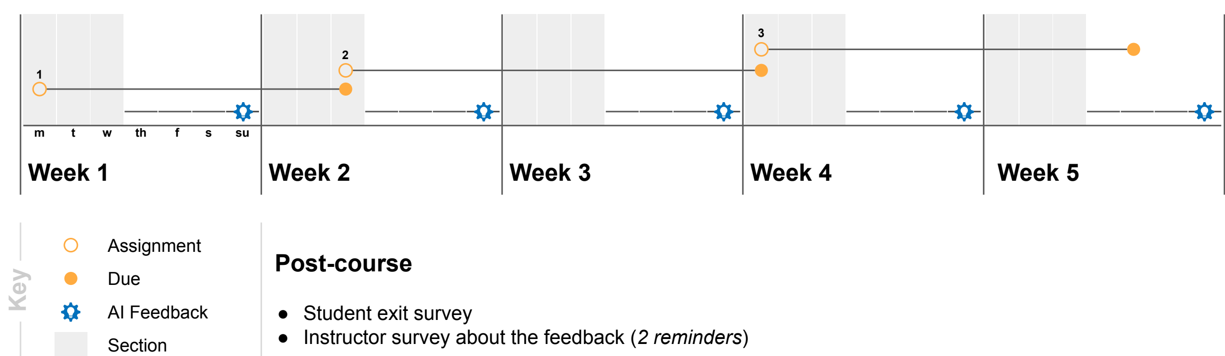


Figure 4: Timeline of the Study

Tables

Table 1: Descriptive Statistics of Analytic Sample

	Mean	SD
A. Instructor Characteristics		
Female	0.318	
Age	29.665	11.252
First-Time Instructor	0.788	
In Africa	0.015	
In Asia	0.159	
In Australia	0.017	
In Europe	0.111	
In North America	0.644	
In South America	0.011	
# of Unique Instructors	918	
B. Student Characteristics		
Female	0.371	
Age		
18-21	0.305	
22-25	0.212	
26-30	0.18	
31-35	0.127	
36-40	0.067	
40+	0.108	
In Africa	0.04	
In Asia	0.446	
In Australia	0.012	
In Europe	0.127	
In North America	0.347	
In South America	0.025	
# of Unique Students	10,794	
C. Student Outcomes		
% of Assignment 1 Completed	0.715	0.419
% of Assignment 2 Completed	0.544	0.486
% of Assignment 3 Completed	0.338	0.467
Class Sections Attended	1.653	0.823

Note: Data come from Code in Place in spring 2021. First-time instructor indicates instructors who taught the first time in Code in Place. Students were asked to choose their age ranges so we do not have their exact ages. Assignment 3 has two versions, one with images and another accessible assignment for visually impaired students. If a student worked on both versions, we use the version a student made more progress on. We only have student attendance information for sections that were conducted in OhYay.

Table 2: Randomization Check

	Control Mean	Treatment Mean	P Value	N
Female	0.33	0.31	0.52	918
Age	28.88	30.41	0.04	917
First-Time CIP Instructor	0.8	0.78	0.41	918
In Africa	0.02	0.02	0.87	918
In Asia	0.16	0.18	0.37	918
In Australia	0.01	0.02	0.36	918
In Europe	0.12	0.11	0.44	918
In North America	0.68	0.66	0.54	918
In South America	0.01	0.01	0.82	918
Offered Week 1 Section	0.96	0.96	0.63	918
Number of Uptakes Per Hour (Week 1)	11.28	10.94	0.41	880
Number of Questions Per Hour (Week 1)	32.73	32.28	0.66	880
Number of Repetitions Per Hour (Week 1)	34.54	34.23	0.77	880
Teacher Talk Time Proportion (Week 1)	0.76	0.76	0.96	880

Note: Joint F-stat is 0.81. First-time instructor indicates instructors who taught the first time in Code in Place. As this course is voluntary, 38 instructors did not show up in the first section (post randomization) and we thus exclude them from our analysis. We also do not have their week-1 discourse features.

Table 3: First Stages

	Instructor Ever Checked Feedback				
	(1) All weeks	(2) Week 2	(3) Week 3	(4) Week 4	(5) Week 5
Email Reminder	0.536**	0.490**	0.537**	0.555**	0.570**
	(0.027)	(0.030)	(0.031)	(0.031)	(0.032)
Female	0.035	0.042	0.034	0.015	0.046
	(0.029)	(0.032)	(0.033)	(0.034)	(0.034)
Age	0.030**	0.027*	0.031**	0.036**	0.024*
	(0.008)	(0.011)	(0.011)	(0.010)	(0.011)
Age^2	-0.000**	-0.000*	-0.000**	-0.000**	-0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
First-Time Instructor	0.050	0.025	0.037	0.079*	0.072+
	(0.032)	(0.037)	(0.038)	(0.039)	(0.040)
In USA	-0.076*	-0.094**	-0.073*	-0.064+	-0.071*
	(0.030)	(0.033)	(0.034)	(0.035)	(0.035)
Number of Uptakes Per Hour (Week 1)	0.003	0.006	0.006	0.000	-0.002
	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Number of Repetitions Per Hour (Week 1)	-0.001	-0.001	-0.001	-0.000	0.000
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Number of Questions Per Hour (Week 1)	-0.000	-0.002	-0.002	0.001	0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Teacher Talk Time Proportion (Week 1)	-0.169	-0.220	-0.284	-0.118	-0.031
	(0.152)	(0.167)	(0.180)	(0.181)	(0.180)
Week=3	0.071**				
	(0.012)				
Week=4	0.113**				
	(0.013)				
Week=5	0.116**				
	(0.014)				
Constant	-0.253	-0.116	-0.088	-0.298	-0.206
	(0.209)	(0.248)	(0.264)	(0.259)	(0.255)
Control Means	0.176	0.124	0.179	0.203	0.204
F Statistics	34.151	17.991	19.837	20.697	21.482
R^2	0.320	0.282	0.310	0.337	0.353
Observations	2962	797	768	710	687

Note: Standard errors are in parentheses. + $p<0.10$ * $p<0.05$ ** $p<0.01$ *** $p<0.001$. These models estimate the effect of the email reminder (treatment) on whether the instructor checked their feedback from the previous week's class session, prior to their subsequent session. Model (1) includes data across all intervention weeks, while columns (2), (3), (4) and (5) show weekly effects of the email reminder on checking the feedback for weeks 2-5, respectively. In addition to the covariates listed, all models include classroom demographics listed in Section 5.

Table 4: Effects of Automated Feedback on Teaching Practices

	(1) Uptake	(2) Question	(3) Repetition	(4) Talk Time
	Panel A: Intent-to-Treat Results			
Email Reminder	0.603*	1.699*	1.044	-0.009
	(0.265)	(0.724)	(0.865)	(0.007)
R^2	0.275	0.345	0.279	0.241
	Panel B: Treatment-on-the-Treated Results			
Ever Checked Feedback	1.125*	3.169*	1.947	-0.016
	(0.491)	(1.344)	(1.606)	(0.013)
Control Mean	8.580	27.849	31.927	0.805
R^2	0.273	0.343	0.278	0.240
Observations	2962	2962	2962	2962

Note: Standard errors, clustered at the instructor level, in parentheses. + $p<0.10$ * $p<0.05$ ** $p<0.01$ *** $p<0.001$. Panel A shows the effects of the email reminder (treatment) on teaching practices. Panel B shows the effects of checking the feedback from the previous class session and prior to their subsequent section on teaching practices estimated via two-stage least squares regression to control for the experimental condition. First stage results are reported in Table 3. The dependent variables are: the number of uptakes per hour (1), number of questions per hour (2), number of repetitions per hour (3) and proportion of teacher talk time (4). All models include the same covariates as Table 3 Model (1): teacher demographics, pre-intervention teaching practices and student demographics, as well as controls for each week.

Table 5: TOT Effects on Teaching Practices by Week

	(1) Uptake	(2) Question	(3) Repetition	(4) Talk Time
		Week 2 (N=797)		
Ever Checked Feedback	0.622	2.233	0.460	-0.004
	(0.741)	(1.864)	(1.993)	(0.015)
Control Mean	9.136	29.867	30.894	0.818
R^2	0.290	0.368	0.346	0.314
		Week 3 (N=768)		
Ever Checked Feedback	1.465*	4.239*	3.962+	-0.049**
	(0.677)	(1.849)	(2.106)	(0.018)
Control Mean	9.010	30.105	33.130	0.801
R^2	0.260	0.319	0.269	0.226
		Week 4 (N=710)		
Ever Checked Feedback	1.233+	3.366*	1.607	0.014
	(0.677)	(1.693)	(2.185)	(0.018)
Control Mean	8.174	25.532	31.579	0.806
R^2	0.308	0.346	0.278	0.233
		Week 5 (N=687)		
Ever Checked Feedback	1.132+	2.868+	1.762	-0.023
	(0.676)	(1.730)	(2.233)	(0.018)
Control Mean	7.826	25.189	32.189	0.793
R^2	0.240	0.304	0.241	0.208

Note: Standard errors in parentheses. + p<0.10 * p<0.05 ** p<0.01 *** p<0.001. The effects of checking the feedback on teaching practices estimated week-by-week via two-stage least squares regression to control for the experimental condition – first stage results are reported in Table 3. The dependent variables are: the number of uptakes per hour (1), number of questions per hour (2), number of repetitions per hour (3) and teacher talk time ratio (4). All models include the same covariates as Table 3: teacher demographics, pre-intervention teaching practices and student demographics.

Table 6: Heterogeneous TOT Effects on Teaching Practices

	(1) Female	(2) Male	(3) First-Time Instructor	(4) Returning Instructor	(5) in U.S.	(6) Not in U.S.	(7) High Wk1 Uptake	(8) Low Wk1 Uptake
Uptake	1.450+	0.958	0.799	2.369*	0.577	2.010**	1.343+	0.930
	(0.856)	(0.597)	(0.556)	(1.108)	(0.648)	(0.706)	(0.715)	(0.665)
Questions	3.586	2.958+	2.213	6.224*	1.489	5.971**	3.506+	2.938
	(2.454)	(1.608)	(1.525)	(2.958)	(1.697)	(2.057)	(1.931)	(1.843)
Repetition	5.347*	0.534	1.019	5.527	-0.496	5.836*	3.131	0.259
	(2.592)	(1.989)	(1.833)	(3.465)	(2.018)	(2.573)	(2.161)	(2.324)
Talk Time	-0.034	-0.007	-0.013	-0.027	0.007	-0.052**	-0.015	-0.019
	(0.023)	(0.016)	(0.016)	(0.025)	(0.017)	(0.020)	(0.018)	(0.019)
N	952	2010	2350	612	1919	1043	1467	1495

Note: Standard errors in parentheses. + $p<0.10$ * $p<0.05$ ** $p<0.01$ *** $p<0.001$. Heterogeneous treatment effects of checking the feedback on teaching practices estimated via two-stage least squares regression to control for the experimental condition – first stage results are reported in Table 3. The dependent variables are: the number of uptakes per hour (1), number of questions per hour (2), number of repetitions per hour (3) and teacher talk time ratio (4). All models include the same covariates as Model (1) in Table 3: teacher demographics, pre-intervention teaching practices and student demographics.

Table 7: TOT Effects on Student Outcomes

	(1)	(2)	(3)	(4)	(5)
	Assn. 2	Assn. 3	Proportion of Classes Attended	Responded to Survey	Course Rating
Ever Checked Feedback	0.035+	0.009	0.021	0.031*	0.111
	(0.021)	(0.019)	(0.024)	(0.015)	(0.155)
Control Mean	0.529	0.333	0.380	0.156	9.386
R^2	0.019	0.012	0.029	0.020	0.018
Observations	9658	9658	9704	9704	1623

Note: Standard errors in parentheses. + p<0.10 * p<0.05 ** p<0.01 *** p<0.001. As assignment 2 was released after week 2's instruction and due on the first day of week 4, we only use whether an instructor checked the feedback at least once prior to week 4 as the independent variable in the first stage of our regression. For the other outcomes, we aggregate data from week 2-4 to construct the independent variable on checking feedback. All models include the same covariates as the instructor-level analyses (e.g. Table 3): teacher demographics, pre-intervention teaching practices and student demographics. Since the data is aggregated across weeks, we also include controls capturing whether an instructor had a transcript for each week.

Appendix

A Automated Speech Recognition by Location

One important limitation of this step is that automated speech recognition is known to be less accurate for speakers of English varieties besides Standard American English (Koenecke et al., 2020). Differences in speech recognition accuracy based on teacher and student demographics are problematic because they may continue to propagate inequities in teachers' professional development. Part of the reason why we selected Assembly.ai is that their service was most accurate based on our manual inspection of a sample of transcripts across English varieties, compared to other speech recognition services. However, at the end of our study, we still found that the confidence scores for transcribed words from Assembly.ai were lower for instructors who were not located in the U.S. (see Figure A1). Further, the main suggestion for improvement that instructors reported about the feedback was to improve transcription quality – see Appendix J for details. That being said, instructors who were *not* in the U.S. rated the feedback tool significantly higher than instructors in the U.S. (Figure A2). Furthermore, the feedback had a significantly more positive impact on instructors' practice who were not in the U.S. compared to those who were in the U.S. (Section 6.3). These results suggest that issues with transcription quality did not impact instructors outside the U.S. more negatively. Since we do not have information about race/ethnicity, we could not conduct the same analysis along this important demographic dimension. Before scaling up the use of our tool, it is our highest priority to evaluate and address speech recognition issues by leveraging technological improvements in this area.

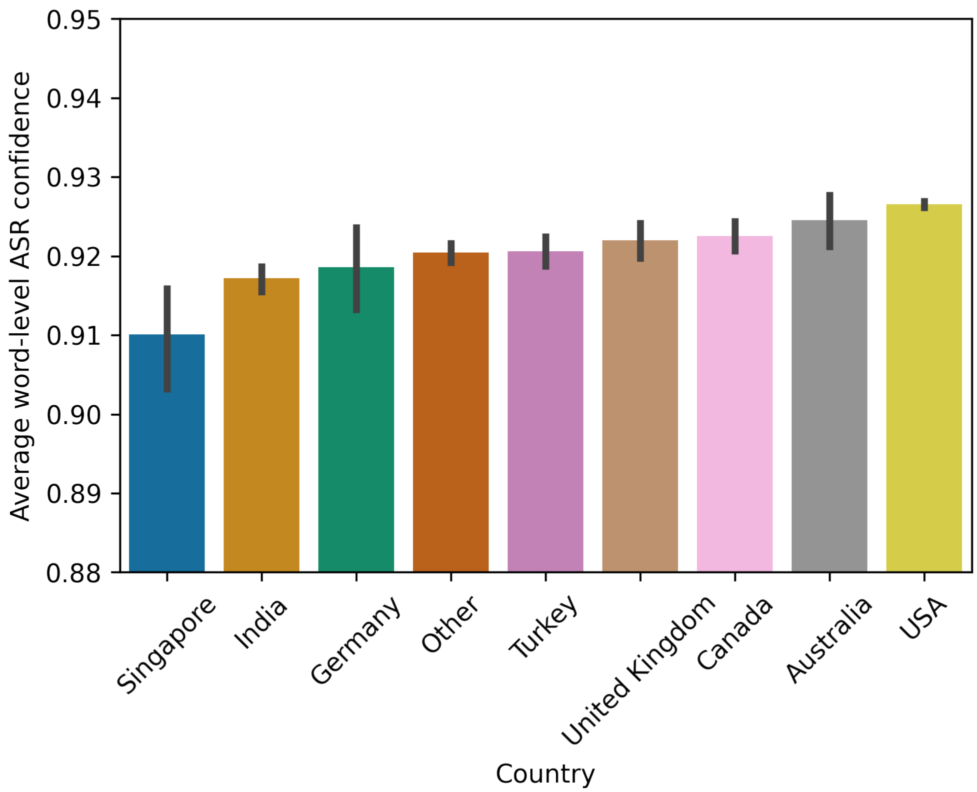


Figure A1

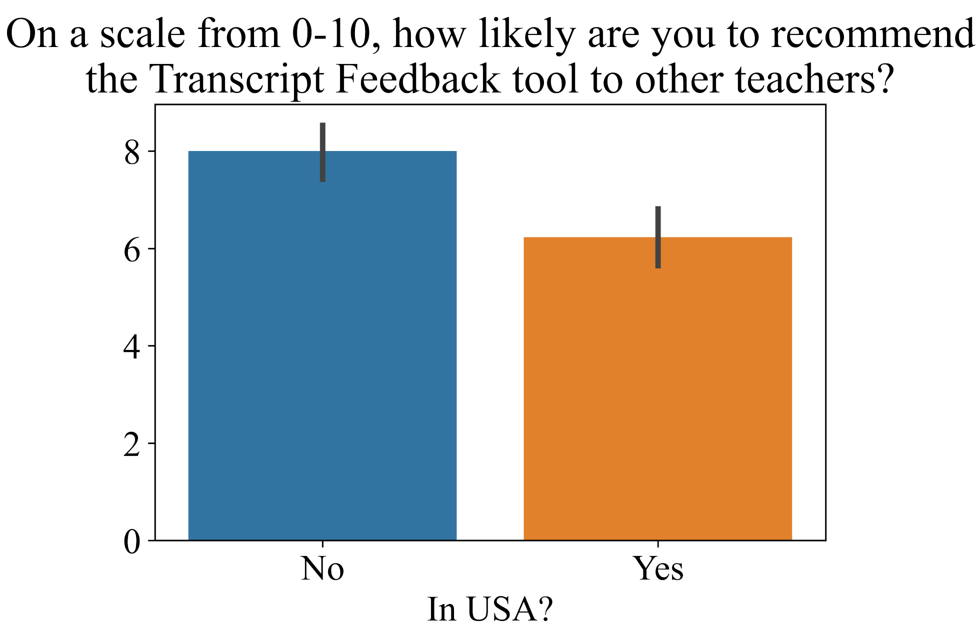


Figure A2

B Details of Natural Language Processing Algorithms

B.1 Student and teacher talk time

We quantify teacher and student talk time using timestamps from the transcripts. Specifically, we sum up the duration of each teacher utterance and compute talk time in minutes for our analyses.

B.2 Teacher questions

We build a question detector to identify teacher questions. The question detector flags an utterance as containing a question either if 1) it contains a question mark, or 2) if our NLP model identifies a question in it, since punctuation from Assembly.ai may not always be accurate. We develop this NLP model using Switchboard (Godfrey et al., 1992), a large corpus of manually transcribed phone conversations that is used often for dialog-related analyses in NLP. We strip all question marks from Switchboard and use those question marks as labels to fine-tune BERT (Devlin et al., 2018), a state-of-the-art NLP model to predict the presence of question marks based on the utterances that are stripped of question marks. This model achieves an accuracy above 90%, and hence we rely on it to catch potential false negatives for teacher questions that we could not detect by purely checking for question marks in our transcripts.

B.3 Teacher repetition

We use the %-IN-T measure from Demszky et al. (2021) to detect instances where the teacher repeats parts of the student utterance. This measure computes the percentage of student words that are part of the teacher utterances, ignoring stopwords and punctuation. We identify stopwords using NLTK's list of stopwords for English (Bird, 2006).

B.4 Correlation Between Uptake and Other Discourse Features

We used pre-intervention transcripts from Code in Place to correlate uptake with these related discourse features. The correlation coefficients are 0.847 ($p < 0.001$) for the number of questions per hour, 0.771 ($p < 0.001$) for the number of repetitions per hour, and -0.412 ($p < 0.001$) for teacher talk time ratio.

C Predictors and Variance Decomposition of Uptake

In order to understand how instructor demographics relate to our uptake measure, we analyze pre-intervention transcripts. We regress the number of uptakes an instructor used in their first section on their demographics and student characteristics. The results are reported in Table A1. We do not find any differential use of uptake by gender, age, or whether an instructor is teaching for Code in Place for the first time. The only statistically significant predictor is whether an instructor is based in the U.S.; instructors who are in the U.S. are more likely to uptake student contributions than those who are not. This difference may be due to transcription errors for non-native speakers rather than actual differences in practices (Appendix A).

Table A1: Predictors of Uptake in Week 1 (Pre-Intervention)

	(1)	(2)
Female	-0.012	-0.026
	(0.442)	(0.448)
Age	0.113	0.007
	(0.097)	(0.147)
Age^2	-0.001	0.000
	(0.001)	(0.002)
First-Time CiP Instructor	-0.320	-0.324
	(0.512)	(0.516)
In USA	0.988*	0.915*
	(0.432)	(0.443)
Student Demographics		X
Constant	8.411**	9.482**
	(1.806)	(2.483)
R^2	0.012	0.019
Observations	879	866

Note: Standard errors in parentheses. + p<0.10 * p<0.05 ** p<0.01 *** p<0.001. Models estimate the effect of instructor demographics on the number of uptakes per hour, using pre-intervention transcripts. Dependent variable is the number of uptakes in instructor's week 1 transcript. Model (2) includes classroom demographics as covariates, mean-aggregated to the transcript level. Student covariates are: proportion of female students, proportion of students in the USA, proportion of students in each age range (18-21, 22-25, 26-30, 31-35, 36-40, 40+).

To further describe the variation of uptake, we decompose the variance of our uptake measure for instructors in the control group using data from all instructional weeks. Our analysis shows that 52% of the variance lies between instructors and 48% of the variance lies between lessons within instructors, suggesting considerable variability of the uptake behavior both between and within an individual. We then run a growth curve model regressing the number of uptake per hour on the week of instruction for the control group (Singer & Willett, 2003). We found that on average, each additional week of instruction is associated with a reduction of about 0.5 uptake per hour (p¡0.01). Thus, we do see evidence that in our online setting, instructors tend to use fewer uptakes over time, which might be a result of decreasing engagement from the students, the instructors, or both. The topic of the section might also influence the number of uptakes — while in the beginning the sections focused more on answering student questions, at the end it focused more on reviewing the whole course material.

D Final Survey for Instructors About the Automated Feedback

We shared the following final survey about the automated feedback tool with a randomly selected sample of 200 instructors. To encourage a high response rate, these instructors received the incentive of a chance to win one of ten $40 Amazon gift cards and we also sent 3 email reminders about the survey.

Transcript Feedback Survey

AI-Based Feedback on Your Week 1 Section

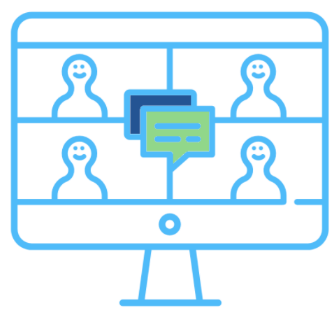
Demo

At Code in Place, we believe in the power of collaborative learning, which has also been shown to lead to student success.

Powered by state of the art AI, we provide you with feedback on two key mechanisms of student engagement: student talktime and moments when you built on student contributions.

This feedback is meant to give you an opportunity to reflect and to support your professional development. It is not meant as an evaluation.

Notes: 20% of your section was spent in breakout rooms, which are not analyzed here. Our language-based algorithms right now only work for sections taught in English.



The Transcript Feedback component of Code in Place was part of a pilot research project. The goal of this project is to understand the usefulness of AI-powered transcript feedback to teachers like you. Thus, your feedback is essential to our project. :)

We are looking for honest feedback, which will help us decide if we should use this tool again and how we can improve it if we do. Your responses are confidential: they will never be linked with your name (only with an anonymous research ID) and they will never be shared or used in any way to reveal your identity, not even to researchers on the Code in Place team.

How often did you engage with the Transcript Feedback?
Select one response.

- Not at all.

- Once or twice.
- Regularly (most weeks).

If they selected "Not at all":

Could you tell us why you didn't engage with the Transcript Feedback?
Select all that apply

- I didn't know about it.
- It wasn't available to me (e.g. I didn't use Ohyay / my section wasn't in English / I had substitute section leaders).
- I didn't have the time.
- I didn't think it would be helpful.
- Other (*please explain*)

Submit

If they selected "Once or twice" :
Could you tell us why you engaged with the Transcript Feedback only once or twice?
Select all that apply

- I only learned about it later in the course.
- It wasn't available to me after each section (e.g. I didn't use Ohyay / my section wasn't in English / I had substitute section leaders).
- I didn't have the time.
- I didn't find it helpful.
- Other (*please explain*)

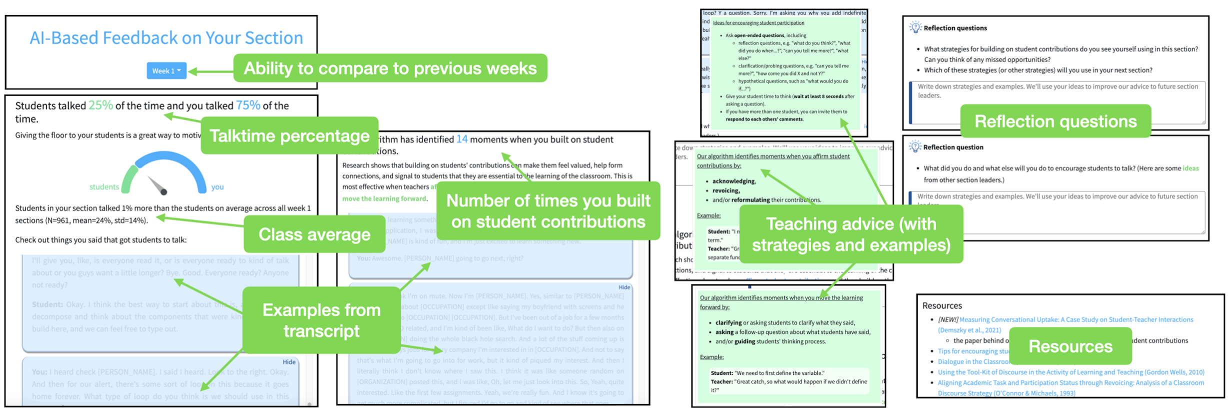
If they selected "Once or twice" or "Regularly most weeks"):

To what extent do you agree with the following about the Transcript Feedback?
Please select one option for each: "Strongly disagree", "Disagree", "Neither agree nor disagree", "Agree", "Strongly agree".

- The feedback has helped me become a better teacher.
- The feedback made me realize things about my teaching that I otherwise would not have.
- The feedback was difficult to understand.
- The feedback made me pay more attention to who was getting a voice in my class than I otherwise would have.
- I tried new things in my teaching because of this feedback

On a scale from 0-10, how likely are you to recommend the Transcript Feedback tool to other teachers?

Please select between 0-10



Please select the MOST helpful elements of the feedback.

Please select between 0-3 elements

- Ability to compare to previous weeks
- Talktime percentage
- Number of times you built on student contributions
- Class average for talktime
- Examples from your transcript for things you said that got students to talk
- Examples from your transcript for moments when you built on student contributions
- Teaching advice (with strategies and examples)

- Reflection questions
- Resources
- Other (*please explain*)

Please select the LEAST helpful elements of the feedback.
Please select between 0-3 elements

- Ability to compare to previous weeks
- Talktime percentage
- Number of times you built on student contributions
- Class average for talktime
- Examples from your transcript for things you said that got students to talk
- Examples from your transcript for moments when you built on student contributions
- Teaching advice (with strategies and examples)
- Reflection questions
- Resources
- Other (*please explain*)

Do you have any suggestions for how we could improve this feedback tool? *(open ended response)*

Do you have any other thoughts / comments? :) *(open ended response)*

Submit

E Final Survey for Students About the Course

Code in Place Survey

We truly appreciate that you took time for Code in Place. It has been so wonderful to go on this adventure of a course with you.

Now that we're wrapping up, we'd like to ask you for a very short reflection on your time with Code in Place. We are always working on improving our own teaching, and the experience we provide students. Filling out this anonymous feedback form will help us decide if we should do this again and how we can improve it if we do.

1. **What did you like about Code in Place?**

2. **What would you improve about Code in Place?**

3. **On a scale from 0-10, how likely are you to recommend being a student in Code in Place to a friend who wants to learn to program?**

4. **Which of these course elements were helpful?**
   Please select one option for each: "Did not use", "Not very helpful", "Somewhat helpful", "Very helpful".

   - Course lectures
   - Small group sections
   - Ed discussion forum
   - Course Assignments
   - Worked Examples

5. **Leave a message for a student thinking of applying to Code in Place!**

Have a story to tell? Email us!

If you feel like something exceptionally positive happened to you that you would like to highlight, please do email codeinplacestaff@gmail.com

Submit

F Attrition Analysis

Table A2: Attrition Analysis

	(1)	(2)	(3)	(4)	(5)
	Had A Transcript in...				
	Week 1	Week 2	Week 3	Week 4	Week 5
Email Reminder	0.032	0.050+	0.003	-0.007	-0.019
	(0.024)	(0.026)	(0.026)	(0.028)	(0.028)
Female	-0.036	-0.036	-0.046	-0.025	-0.016
	(0.026)	(0.027)	(0.028)	(0.030)	(0.030)
Age	0.017**	0.018**	0.026**	0.027**	0.030**
	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)
Age^2	-0.000*	-0.000*	-0.000**	-0.000**	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
First-Time CiP Instructor	0.014	-0.016	0.006	0.019	0.034
	(0.030)	(0.032)	(0.033)	(0.034)	(0.035)
In USA	-0.023	-0.024	-0.004	0.042	0.002
	(0.026)	(0.027)	(0.028)	(0.029)	(0.030)
Constant	0.428**	0.393**	0.203+	0.078	0.020
	(0.111)	(0.116)	(0.120)	(0.126)	(0.127)
Control Means	0.752	0.710	0.698	0.649	0.638
R^2	0.029	0.032	0.046	0.049	0.052
Observations	1129	1129	1129	1129	1129

Note: The outcome variables for the five columns indicate whether there is a transcript for an instructor in a particular instruction week. The variable email reminder indicates the treatment status. Standard errors in parentheses. + $p<0.10$ * $p<0.05$ ** $p<0.01$ *** $p<0.001$.

G Table 4 Without Controls

Table A3: Effects of Automated Feedback on Teaching Practices

	(1)	(2)	(3)	(4)
	Uptake	Question	Repetition	Talk Time
	Panel A: Intent-to-Treat Results			
Email Reminder	0.556+	1.521	0.979	-0.006
	(0.329)	(0.938)	(1.048)	(0.008)
Control Mean	8.606	27.965	31.874	0.804
R^2	0.010	0.022	0.004	0.007
	Panel B: Treatment-on-the-Treated Results			
Ever Checked Feedback	1.051+	2.875	1.850	-0.012
	(0.620)	(1.768)	(1.978)	(0.016)
Control Mean	8.580	27.849	31.927	0.805
R^2	0.009	0.022	0.004	0.008
Observations	3002	3002	3002	3002

Note: Standard errors, clustered at the instructor level, in parentheses. + p<0.10 * p<0.05 ** p<0.01 *** p<0.001. The table replicates Table 4 without any controls other than binary weekly indicators. Panel A shows the effects of the email reminder (treatment) on teaching practices. Panel B shows the effects of checking the feedback from the previous class session on teaching practices estimated via two-stage least squares regression to control for the experimental condition. The dependent variables are: the number of uptakes per hour (1), number of questions per hour (2), number of repetitions per hour (3) and proportion of teacher talk time (4).

H Visualizing Coefficients from Tables 5 and 6

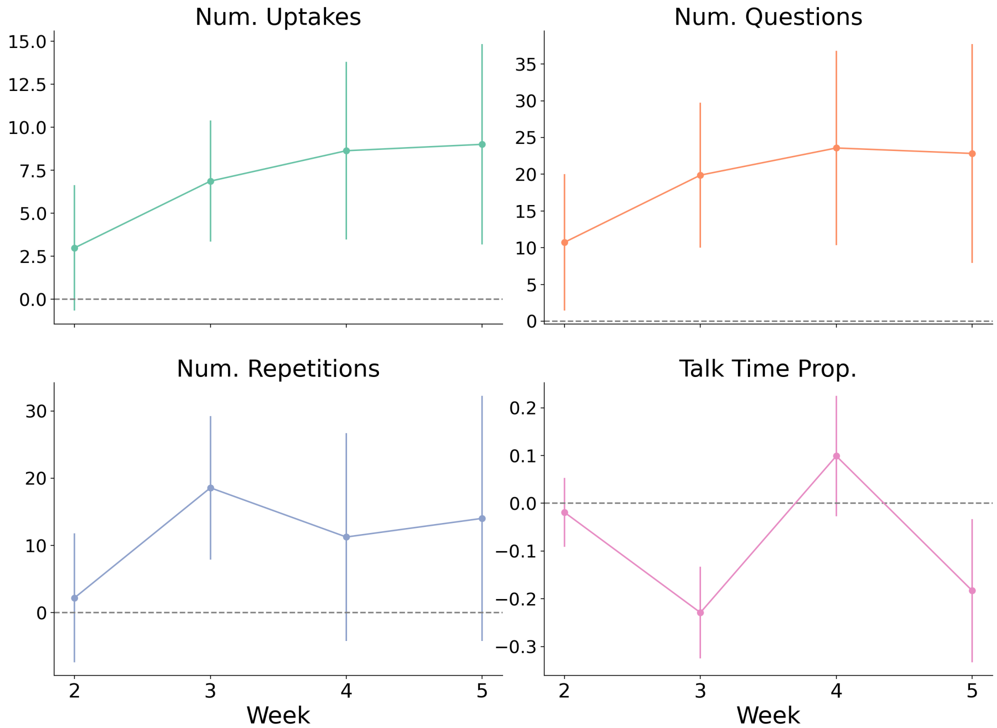


Figure A3: TOT Effects on Teaching Practices by Week (Table 5)

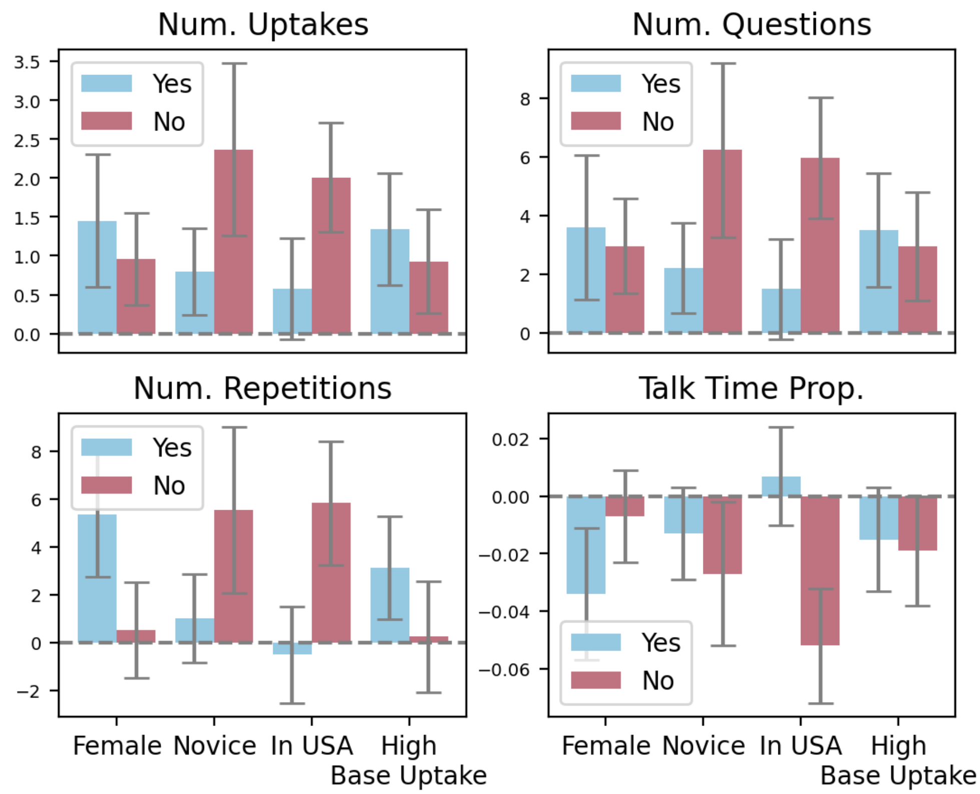


Figure A4: Heterogeneous TOT Effects on Teaching Practices (Table 6)

I Student Survey Responses

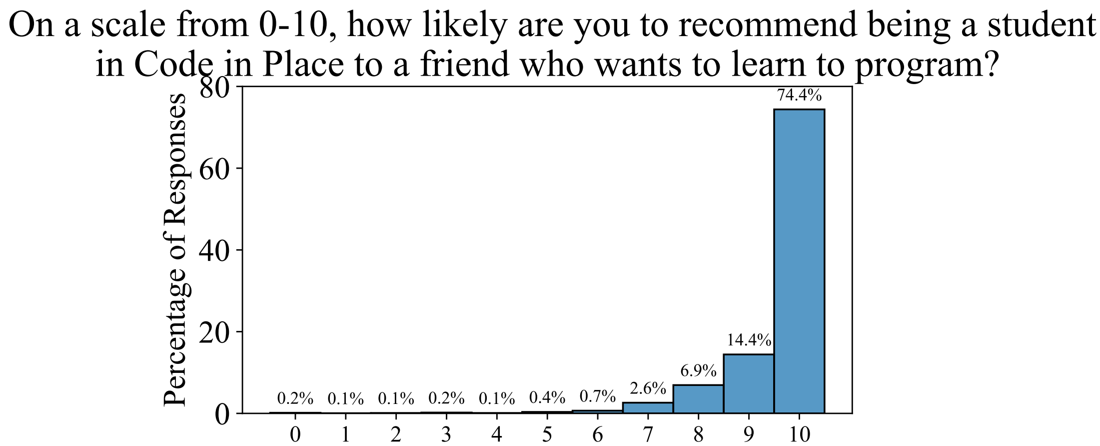


Figure A5

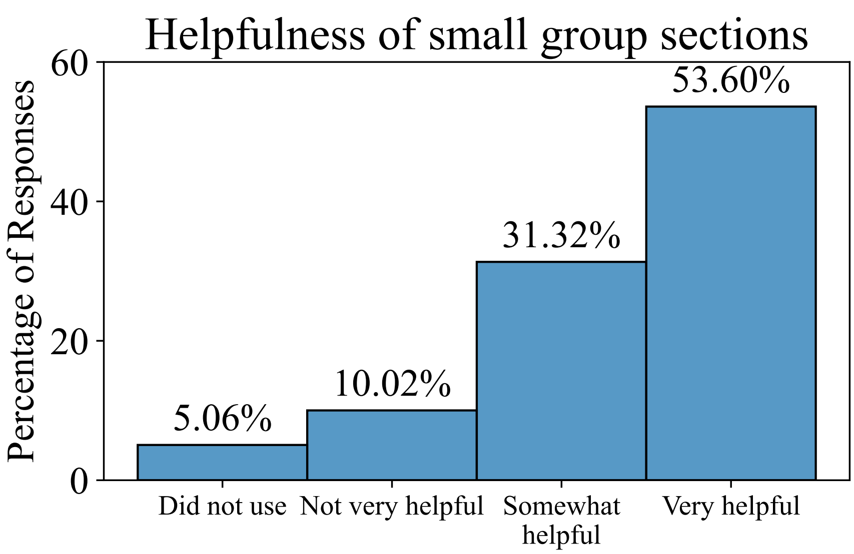


Figure A6

J Instructor Survey Responses

We analyze instructors' responses to the confidential endline survey (Appendix D) to understand if they found the feedback helpful (n=142). Instructors were strongly encouraged to report their honest opinion as a way to help improve the tool. We found that overall, instructors reported that the feedback was helpful: the majority of instructors reported that the tool 1) helped them become a better teacher (57%, Figure A11), 2) made them realize things about their teaching that they otherwise would not have (76%, Figure A12), 3) made them pay more attention to who was getting voice in their class (57%, Figure A13, 4) tried new things in their teaching as a result of the feedback (53%, Figure A14) and that 5) the feedback wasn't difficult to understand (64%, Figure A15). Instructors gave an average score of 7 out of 10 for how likely they are to recommend the tool to other teachers Figure A10. In the open-ended questions, the most frequently reported suggestions for improvement (n=62) relate to improving the transcription (n=20) and incorporating the chat into the analysis (n=8).

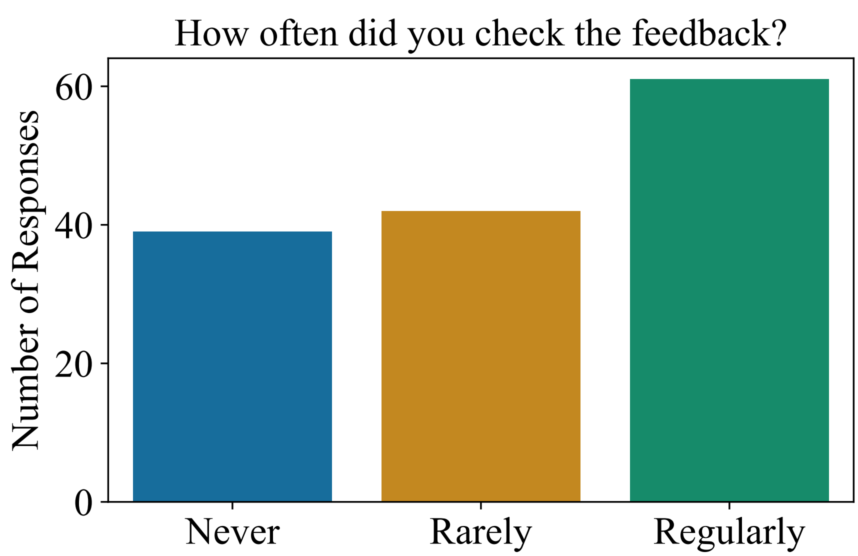


Figure A7

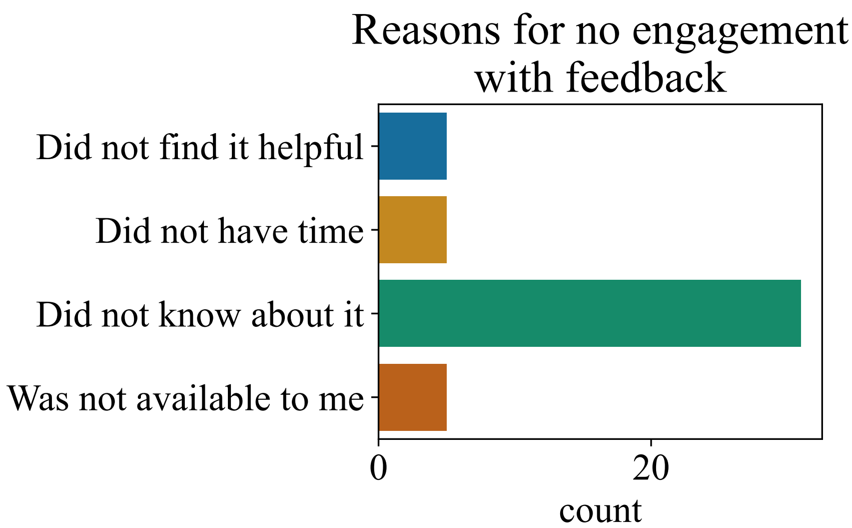


Figure A8

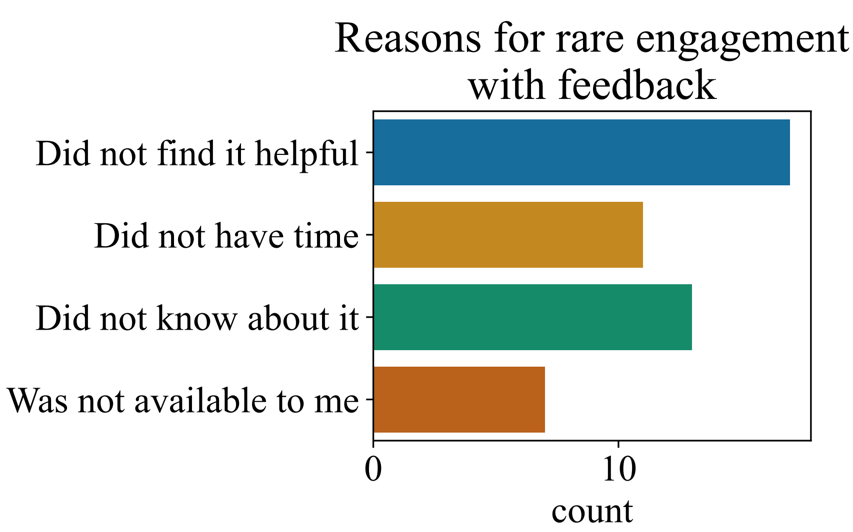


Figure A9

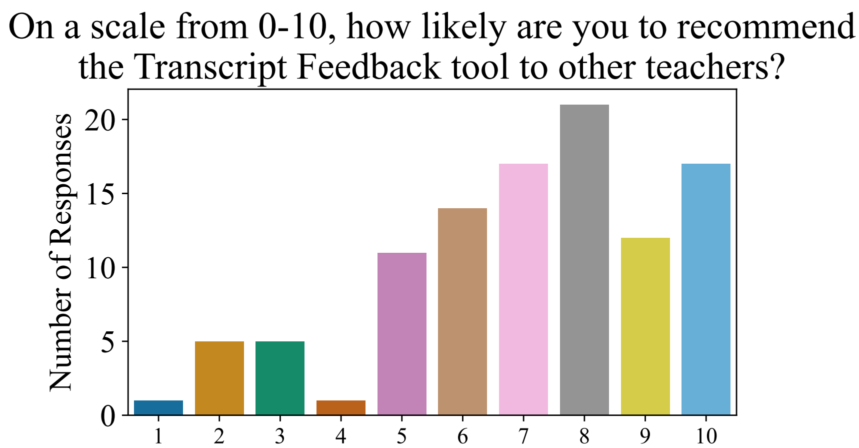


Figure A10

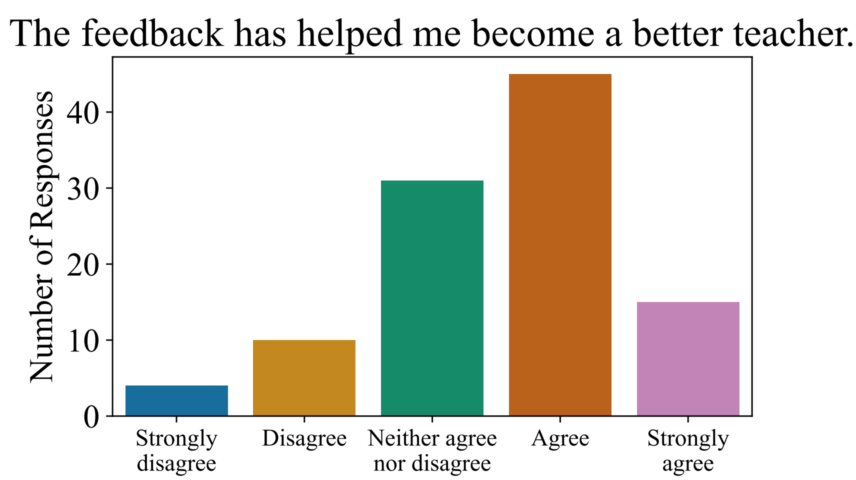


Figure A11

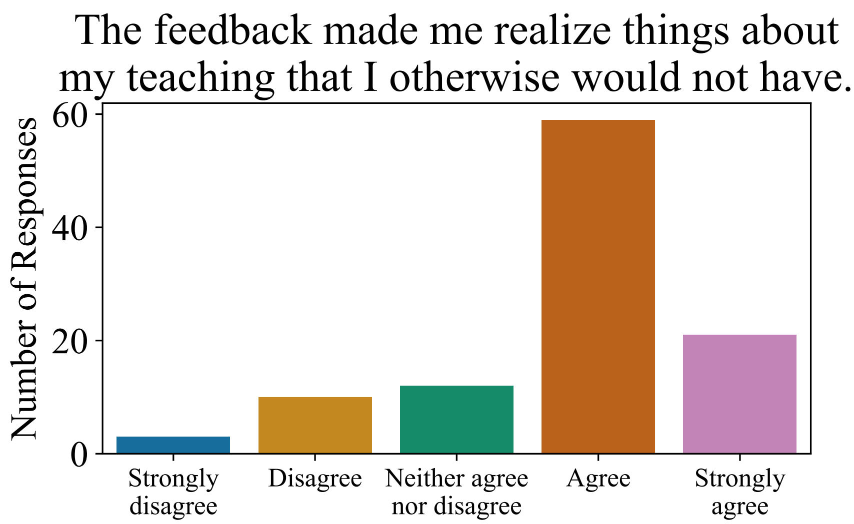


Figure A12

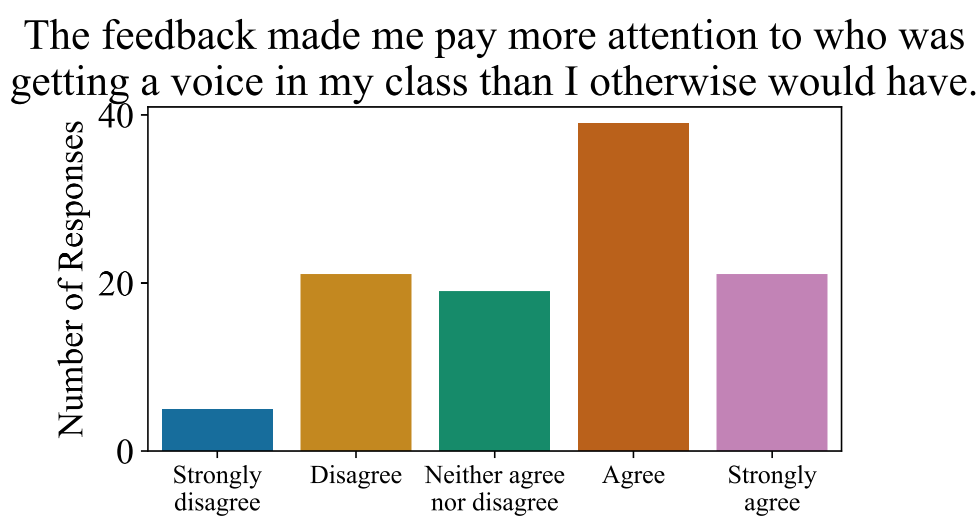


Figure A13

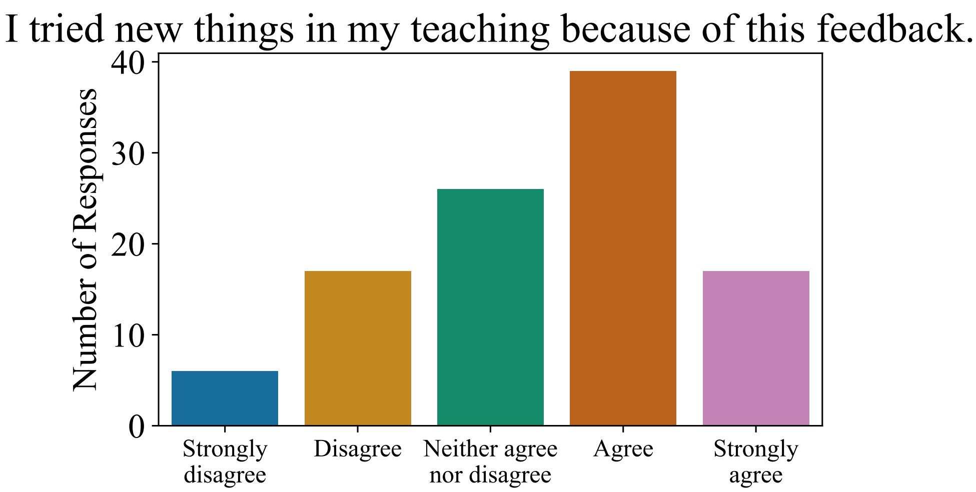


Figure A14

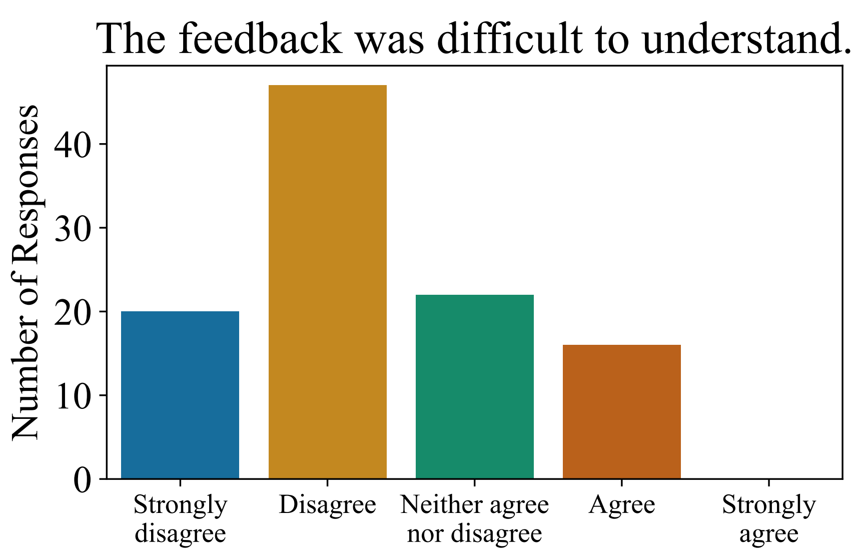


Figure A15

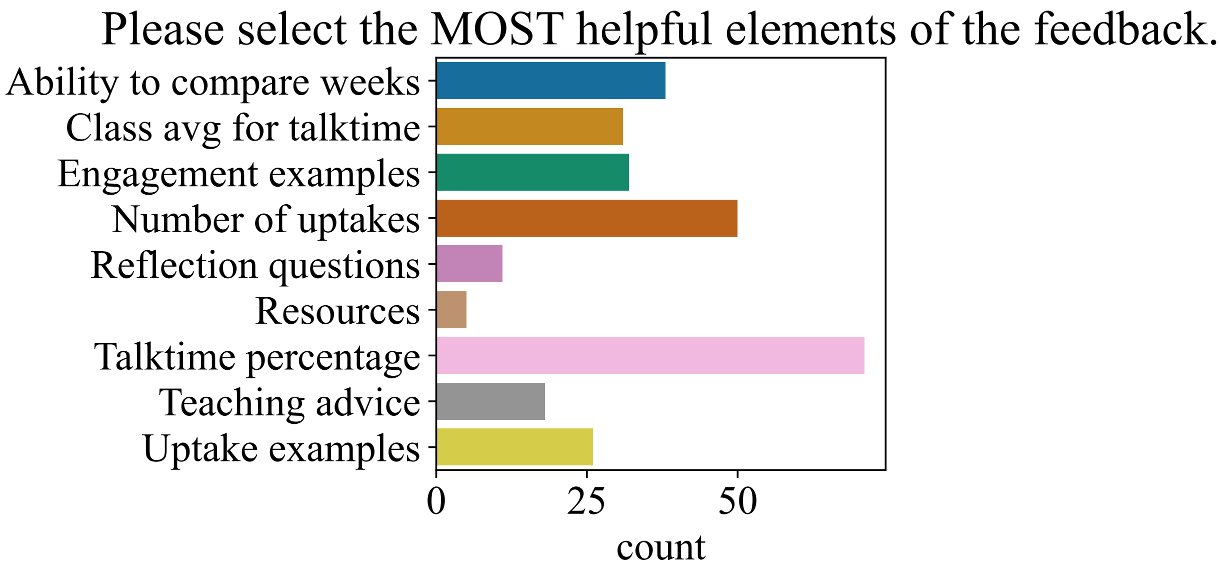


Figure A16

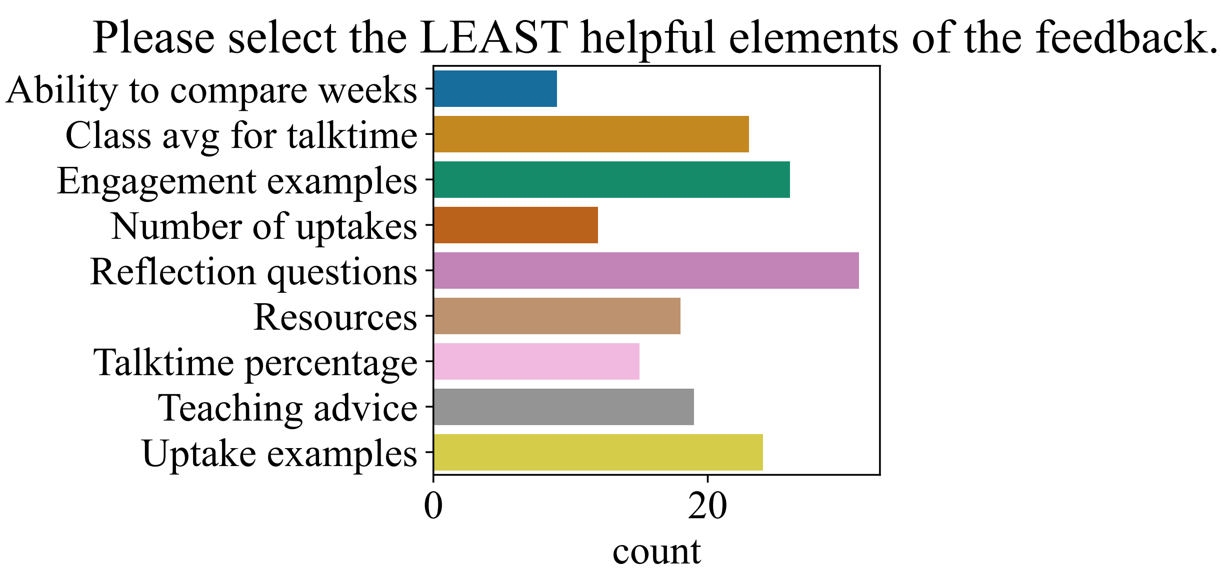


Figure A17