

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

Dorottya Demszky 

Stanford University

Jing Liu 

The University of Maryland, College Park

Heather C. Hill 

Harvard University

Dan Jurafsky

Chris Piech

Stanford University

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: *artificial intelligence, teacher education/development, instructional technologies, textual analysis, measurements, regression analyses, experimental research, statistics, randomized controlled trial, natural language processing, teaching practices, online learning*

CAUSAL evidence suggests that providing teachers formative feedback can improve both their instruction (Kraft et al., 2018) and their students' outcomes (Steinberg & Sartain, 2015; Taylor & Tyler, 2012). 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 United States 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). Furthermore, 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; Gibbons & Cobb, 2017; Kane & Rosenquist, 2019).

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 and 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). Although 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; Demszky et al., 2021; Herbel-Eisenmann et al., 2009; Nystrand et al., 2000; O'Connor & Michaels, 1993; Wells & Arauz, 2006). 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 5-week, free online computer science course organized by Stanford University. This course teaches introduction to programming to ~12,000 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 with 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 at 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 with the control group. Treatment-on-the-treated (TOT) 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 United States 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.

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; Demszky et al., 2021; Herbel-Eisenmann et al., 2009; Nystrand et al., 2000; O'Connor & Michaels, 1993; Wells & Arauz, 2006), 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 United States 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 and 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 subcategories 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 the section “Step 3: Transcript Analysis,” 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's $\rho = .54, p < .001$). 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 ELA virtual classrooms, and a text-based math and science tutoring setting.

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 (Donnelly et al., 2017; Jensen et al., 2020; Kelly et al., 2018; Samei et al., 2014). 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).

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 (ASR) 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 nonevaluative. 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.

Finally, the specific context of an online, voluntary computer science course closely mimics many emerging teaching settings such as virtual tutoring¹ 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.

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 was 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 and 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 shows the basic demographics of our analytic sample.

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. In all, 65% of our instructor sample described themselves as male, 32% as female, and 1% as nonbinary. Instructors ranged in age from 18 to 81 years, with an average of roughly 30 years old. They were located in 82 unique countries with the majority (63%) being in the United States; 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 motivations 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

TABLE 1
Descriptive Statistics of Analytic Sample

Variable	<i>M</i>	<i>SD</i>
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	
No. 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	0.347	
In North America	0.347	
In South America	0.025	
No. 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. 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.

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 United States (30%) accounting for more than 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 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.

The M-Powering Teachers Tool

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.

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.

Step 2: Transcription and Anonymization. We transcribed and algorithmically anonymized recordings using Assembly.ai, a service we chose because of its accuracy, cost-effectiveness (us\$1 per 1 hour 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,” and so on. One important limitation of this step is that ASR is known to be less accurate for speakers whose native language is not Standard American English (Koencke et al., 2020), and we do find disparate accuracies in our data as well. However, we have found evidence that the tool does not affect instructors outside the United States more negatively—see Supplementary Appendix A in the online version of the journal 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.

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: (a) the National Center for Teacher Effectiveness (NCTE) transcript dataset of elementary math classrooms (Demszky & Hill, 2022), (b) the Switchboard dataset of phone conversations, widely used in NLP research on dialog (Godfrey et al., 1992), and (c) 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 pairs 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, as 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.'s (2021) 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. Supplementary Appendix B in the online version of the journal provides more details on these measures and their correlation with the uptake measure.

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.

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 nonjudgmental information about teachers' instruction in a way that respects their agency and authority over their practice (Oolbekkink-Marchand et al., 2017; Priestley et al., 2015; Wills & Haymore Sandholtz, 2009). 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 to 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.

User Interface of the Feedback Application

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.

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

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. 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 nonpersonalized 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.

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 and Wednesday of each week, and we sent the email reminders on the Sunday of each week.

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 Supplementary Appendix C in the online version of the journal 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 "Step 3: Transcript Analysis" 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

AI-Based Feedback on Your Section

Week 1 ▾

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: 1% 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.



Students talked **21%** of the time and you talked **79%** of the time.

Giving the floor to your students is a great way to motivate them and help them learn.



Students in your section talked 3% less than the students on average across all week 1 sections (N=961, mean=24%, std=14%). This could also be because you engaged students in breakout rooms as opposed to the main room.

Check out things you said that got students to talk:

post conditions, and I think control flow basically like loops and conditionals, right?

You: And what would be a good use of the while loop? Hide

Student: Like when you wanted to be repeated? Like, when the condition is true or when you don't know the exact number of times you wanted to be repeated? Yes.

You: Sorry. Oh, by the way, you guys can just type it for us. I think I heard move two spaces deeper, where are we a

Student: [PERSON_NAME] and I thought function. And when [PERSON_NAME] so

Ideas for encouraging student participation

- Ask **open-ended questions**, including
 - reflection questions, e.g. "what do you think?", "what did you do when...?", "can you tell me more?", "what else?"
 - clarification/probing questions, e.g. "can you tell me more?", "how come you did X and not Y?"
 - hypothetical questions, such as "what would you do if...?"
- Give your student time to think (**wait at least 8 seconds** after asking a question).
- If you have more than one student, you can invite them to **respond to each others' comments**.

Reflection question

- What did you do and what else will you do to encourage students to talk? (Here are some **ideas** from other section leaders.)

Write down strategies and examples. We'll use your ideas to improve our advice to future section leaders.

FIGURE 1. Components of the M-Powering Teachers web application (Part 1).
Note. AI = artificial intelligence.

Our algorithm identifies moments when you affirm student contributions by:

- **acknowledging**,
- **revoicing**,
- and/or **reformulating** their contributions.

Example:

Student: "I made a separate function for calculating the first term."
Teacher: "Great, so you are modularizing your code by creating separate functions."

Our algorithm identifies moments when you move the learning forward by:

- **clarifying** or asking students to clarify what they said,
- **asking** a follow-up question about what students have said,
- and/or **guiding** students' thinking process.

Example:

Student: "We need to first define the variable."
Teacher: "Great catch, so what would happen if we didn't define it?"

Our algorithm has identified 16 moments when you built on student contributions.

Research shows that building on students' contributions can make them feel valued, help form connections, and signal to students that they are essential to the learning of the classroom. This is most effective when teachers **affirm student contributions** and then build on them to **move the learning forward**.

heard move two spaces, the deeper. Cool. So after we move two spaces deeper, where are we at? What should we do next?

Student: [PERSON_NAME] and I thought we should have, like, build hospital function. And when [PERSON_NAME] sort of comes across a beeper in a [PERSON_NAME] executes the built hospital function. Hide

You: Awesome. So I guess pre condition would be on top of the deeper, I think. Right? Yeah. Then what would you I was to build a hospital once you're on top of deeper, I guess.

Student: I guess we move, I guess [PERSON_NAME] moves until it finds another deeper and it executes the function again. Hide

You: We move until we find next deeper. Let me build a hospital again. And then that takes care of the second one. What we do for the third one, wherever many comes after that, I guess we do the same, I guess. Move until we find the text. You can repeat for the rest. Alright, cool. We have, like, our little bit on code here. So first I guess if you're talking about control flow and design decisions

Reflection questions

- What strategies for building on student contributions do you see yourself using in this section? Can you think of any missed opportunities?
- Which of these strategies (or other strategies) will you use in your next section?

Write down strategies and examples. We'll use your ideas to improve our advice to future section leaders.

Resources

- [Tips for encouraging student participation](#)
- [Dialogue in the Classroom \(Gordon Wells, 2006\)](#)
- [Using the Tool-Kit of Discourse in the Activity of Learning and Teaching \(Gordon Wells, 2010\)](#)
- [Aligning Academic Task and Participation Status through Revoicing: Analysis of a Classroom Discourse Strategy \(O'Connor & Michaels, 1993\)](#)
- [Questions in Time: Investigating the Structure and Dynamics of Unfolding Classroom Discourse \(Nystrand et al., 2003\)](#)
- ["Teaching isn't for Rock Stars" \(blog post by Patrick Watson, 2020\)](#)

Review Full Transcript

You: [00:00:00] Hi. How are you? Hey, How's everyone doing? We're probably going to wait for a few more minutes and then start, but we can chat. In the meantime, if you would like, Hello. Hello. [00:05:52]

Student 1: [00:05:52] How are you? [00:05:54]

You: [00:05:54] Good. And you? [00:05:55]

FIGURE 2. Components of the M-Powering Teachers web application (Part 2).

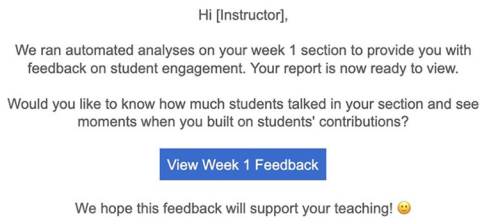


FIGURE 3. *Generic email encouraging instructors to check the feedback.*

instructors could not have yet affected the completion rate of 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 teaching, and suggestions for improving the tool. We include the survey in Supplementary Appendix D in the online version of the journal. Instructors were sampled irrespective of treatment status, received up to three reminders, and were incentivized with a chance to win 1 of 10 US\$40 Amazon gift cards. The survey achieved a 71% response rate ($n = 142$), which does not differ by treatment group ($p = .303$).

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, and so on. The lack of reminders and incentive explains the low response rate for the student survey. We include the survey in Supplementary Appendix E in the online version of the journal. 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 to 10 scale.⁶ All survey data were deidentified before analysis and linked through anonymous research IDs.

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 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 Supplementary Appendix F Table A2 in the online version of the journal 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.

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

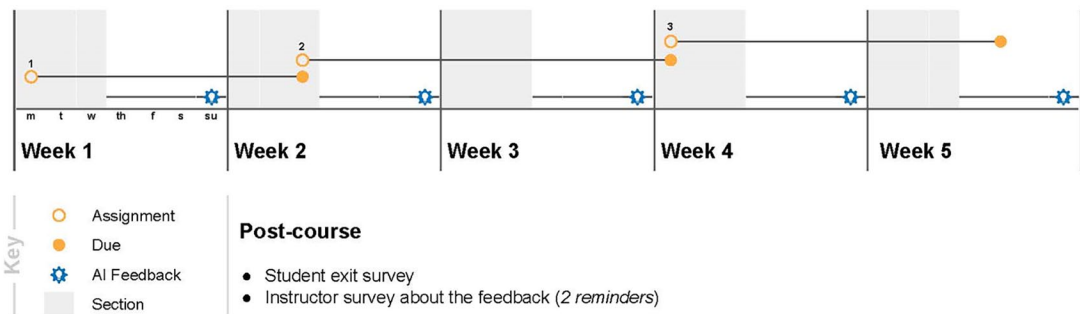


FIGURE 4. *Timeline of the study.*

Note. AI = artificial intelligence.

TABLE 2

Randomization Check

Variable	Control M	Treatment M	p value	n
Female	0.33	0.31	.52	918
Age	28.88	30.41	.04	917
First-time Code in Place instructor	0.8	0.78	.41	918
In Africa	0.02	0.02	.87	918
In Asia	0.16	0.18	.37	918
In Australia	0.01	0.02	.36	918
In Europe	0.12	0.11	.44	918
In North America	0.68	0.66	.54	918
In South America	0.01	0.01	.82	918
Offered Week 1 section	0.96	0.96	.63	918
Number of uptakes per hour (Week 1)	11.28	10.94	.41	880
Number of questions per hour (Week 1)	32.73	32.28	.66	880
Number of repetitions per hour (Week 1)	34.54	34.23	.77	880
Teacher talk time proportion (Week 1)	0.76	0.76	.96	880

Note. Joint F statistic 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.

a weekly basis and the course is 5 weeks long, we can observe how teaching practices evolve from Week 2 to Week 5. 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) estimator:

$$Feedback_{it} = \pi_0 + \pi_1 T_i + \pi_2 X_i + \epsilon_{it}, \quad (1)$$

$$Y_{it} = \beta_0 + \beta_1 Feedback_{it} + \beta_2 X_i + \mu_{it}, \quad (2)$$

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, age², in the United States, first-time Code in Place 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 United States, 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 United States, and whether they demonstrated high or low uptake in their first week of instruction. Finally, 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 5 weeks of teaching as the primary independent variable and conduct the analysis at the student level.

Results

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 four times more compared with 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 United States are more likely to interact with the feedback.

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.

The ITT results, which are reported in Panel A of Table 4, suggest that our intervention improved

TABLE 3
First Stages

Variable	Instructor ever checked feedback				
	(1)	(2)	(3)	(4)	(5)
	All weeks	Week 2	Week 3	Week 4	Week 5
Email reminder	0.536** (0.027)	0.490** (0.030)	0.537** (0.031)	0.555** (0.031)	0.570** (0.032)
Female	0.035 (0.029)	0.042 (0.032)	0.034 (0.033)	0.015 (0.034)	0.046 (0.034)
Age	0.030** (0.008)	0.027* (0.011)	0.031** (0.011)	0.036** (0.010)	0.024* (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 (0.032)	0.025 (0.037)	0.037 (0.038)	0.079* (0.039)	0.072 [†] (0.040)
In United States	-0.076* (0.030)	-0.094** (0.033)	-0.073* (0.034)	-0.064 [†] (0.035)	-0.071* (0.035)
Number of uptakes per hour (Week 1)	0.003 (0.004)	0.006 (0.005)	0.006 (0.005)	0.000 (0.005)	-0.002 (0.005)
Number of repetitions per hour (Week 1)	-0.001 (0.002)	-0.001 (0.002)	-0.001 (0.002)	-0.000 (0.002)	0.000 (0.002)
Number of questions per hour (Week 1)	-0.000 (0.002)	-0.002 (0.002)	-0.002 (0.002)	0.001 (0.002)	0.001 (0.002)
Teacher talk time proportion (Week 1)	-0.169 (0.152)	-0.220 (0.167)	-0.284 (0.180)	-0.118 (0.181)	-0.031 (0.180)
Week = 3	0.071** (0.012)				
Week = 4	0.113** (0.013)				
Week = 5	0.116** (0.014)				
Constant	-0.253 (0.209)	-0.116 (0.248)	-0.088 (0.264)	-0.298 (0.259)	-0.206 (0.255)
Control means	0.176	0.124	0.179	0.203	0.204
<i>F</i> statistics	34.151	17.991	19.837	20.697	21.482
<i>R</i> ²	.320	.282	.310	.337	.353
Observations	2,962	797	768	710	687

Note. Standard errors are in parentheses. 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 to 5, respectively. In addition to the covariates listed, all models include classroom demographics listed in the “Empirical Strategy” section.

[†] $p < .10$. * $p < .05$. ** $p < .01$.

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

TABLE 4

Effects of Automated Feedback on Teaching Practices

Variable	(1)	(2)	(3)	(4)
	Uptake	Question	Repetition	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	.275	.345	.279	.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	.273	.343	.278	.240
Observations	2,962	2,962	2,962	2,962

Note. Standard errors, clustered at the instructor level, in parentheses. 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.

* $p < .05$.

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 with 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%, $p < .05$). Similarly, we find that instructors who checked the feedback asked roughly 3.20 (11.4%) more questions per class ($p < .05$), 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 the binary weekly

indicators in Supplementary Appendix Table A3 in the online version of the journal. 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 with 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 ($p < .05$). While there is a marginally significant coefficient for repetition, we also observe a drop of instructors' talk time (4.9%, $p < .01$). In Weeks 4 and 5, the coefficients for the uptake outcome decrease while remaining statistically significant

TABLE 5

Treatment-on-the-Treated Effects on Teaching Practices by Week.

Variable	(1)	(2)	(3)	(4)
	Uptake	Question	Repetition	Talk time
Week 2 (<i>n</i> = 797)				
Ever checked feedback	0.622 (0.741)	2.233 (1.864)	0.460 (1.993)	-0.004 (0.015)
Control <i>M</i>	9.136	29.867	30.894	0.818
<i>R</i> ²	.290	.368	.346	.314
Week 3 (<i>n</i> = 768)				
Ever checked feedback	1.465* (0.677)	4.239* (1.849)	3.962 [†] (2.106)	-0.049** (0.018)
Control <i>M</i>	9.010	30.105	33.130	0.801
<i>R</i> ²	.260	.319	.269	.226
Week 4 (<i>n</i> = 710)				
Ever checked feedback	1.233 [†] (0.677)	3.366* (1.693)	1.607 (2.185)	0.014 (0.018)
Control <i>M</i>	8.174	25.532	31.579	0.806
<i>R</i> ²	.308	.346	.278	.233
Week 5 (<i>n</i> = 687)				
Ever checked feedback	1.132 [†] (0.676)	2.868 [†] (1.730)	1.762 (2.233)	-0.023 (0.018)
Control <i>M</i>	7.826	25.189	32.189	0.793
<i>R</i> ²	.240	.304	.241	.208

Note. Standard errors in parentheses. 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.

[†]*p* < .10. **p* < .05. ***p* < .01.

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 2 weeks.

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 United States, and whether they demonstrated high or low uptake in their first week of instruction. The results are shown in Table 6.

While female instructors increase the number of times they take up student ideas slightly more as a result of the feedback compared with 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 United States increased their uptake of student contributions by roughly two 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 United States 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

TABLE 6

Heterogeneous Treatment-on-the-Treated Effects on Teaching Practices

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable	Female	Male	First-time instructor	Returning instructor	In the United States	Not in the United States	High-Week 1 uptake	Low-Week 1 uptake
Uptake	1.450 [†] (0.856)	0.958 (0.597)	0.799 (0.556)	2.369* (1.108)	0.577 (0.648)	2.010** (0.706)	1.343 [†] (0.715)	0.930 (0.665)
Questions	3.586 (2.454)	2.958 [†] (1.608)	2.213 (1.525)	6.224* (2.958)	1.489 (1.697)	5.971** (2.057)	3.506 [†] (1.931)	2.938 (1.843)
Repetition	5.347* (2.592)	0.534 (1.989)	1.019 (1.833)	5.527 (3.465)	-0.496 (2.018)	5.836* (2.573)	3.131 (2.161)	0.259 (2.324)
Talk time	-0.034 (0.023)	-0.007 (0.016)	-0.013 (0.016)	-0.027 (0.025)	0.007 (0.017)	-0.052** (0.020)	-0.015 (0.018)	-0.019 (0.019)
<i>n</i>	952	2,010	2,350	612	1,919	1,043	1,467	1,495

Note. Standard errors in parentheses. 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.

[†] $p < .10$. * $p < .05$. ** $p < .01$.

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 was 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.

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. As 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 with the control mean ($p < .10$). 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 the student proportion of classes attended.

TABLE 7

Treatment-on-the-Treated Effects on Student Outcomes

Variable	(1)	(2)	(3)	(4)	(5)
	Assignment 2	Assignment 3	Proportion of classes attended	Responded to survey	Course rating
Ever checked feedback	0.035 [†] (0.021)	0.009 (0.019)	0.021 (0.024)	0.031* (0.015)	0.111 (0.155)
Control <i>M</i>	0.529	0.333	0.380	0.156	9.386
<i>R</i> ²	.019	.012	.029	.020	.018
Observations	9,658	9,658	9,704	9,704	1,623

Note. Standard errors in parentheses. 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 Weeks 2 to 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. As the data are aggregated across weeks, we also include controls capturing whether an instructor had a transcript for each week.

[†] $p < .10$. * $p < .05$.

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 with the control mean ($p < .05$). While we do not see statistically significant results 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 Supplementary Appendix Figure A5 in the online version of the journal 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 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.

Instructor Feedback

Because 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 Supplementary Appendix J in the online version of the journal). 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.

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 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., Gonzalez et al., 2022; Kraft et al., 2018).

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 (Georgia⁹ and Florida¹⁰), 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. Qualitative 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 with 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 US\$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. Furthermore, 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. Furthermore, teacher and student talk may not be audible if recorded on typical handheld devices (e.g., phones or tablets), and ASR 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 NLP methods that build upon it, to make our tool robust and fair (Kizilcec & Lee, 2022).

Despite its limitations, this study constitutes an important step toward 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.

Acknowledgments

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.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The project was partially funded by the HAI Hoffman-Yee grant (#203116) through Stanford University. Demszky acknowledges support of the Melvin and Joan Lane Stanford Graduate Fellowship.

ORCID iDs

Dorottya Demszky  <https://orcid.org/0000-0002-6759-9367>

Jing Liu  <https://orcid.org/0000-0002-9918-8642>

Heather C. Hill  <https://orcid.org/0000-0001-5181-1573>

Notes

1. <https://www.chalkbeat.org/2022/6/29/23186973/virtual-tutoring-schools-covid-relief-money>

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

3. In all, 3% in Canada; 2% each in Bangladesh, Germany, and the United Kingdom, 1% each in Nigeria, Turkey, Singapore, Australia, Pakistan, Brazil, Philippines, Japan, Nepal, Russia, Serbia, Kenya, Indonesia; and 16% total in other countries.

4. We do not have evidence for spillover effects. As 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.

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.

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.

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 instructor 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 were not able to find substitutes, they would share a recorded section by another instructor 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.

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, as instructors were not aware of the intervention and

most of them were also not aware of student end-line surveys. Thus, we can reasonably assume that these differences are due to an indirect effect of teaching practice on student satisfaction.

9. <https://www.legis.ga.gov/api/legislation/document/20112012/127888>

10. <https://www.fldoe.org/core/fileparse.php/5606/urlt/Virtual-Sept.pdf>

References

- Alic, S., Demszky, D., Mancenido, Z., Liu, J., Hill, H., & Jurafsky, D. (2022). Computationally identifying funneling and focusing questions in classroom discourse. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)* (pp. 224–233). Association for Computational Linguistics.
- 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* (pp. 148–160). Springer.
- Aslan, S., Alyuz, N., Tanriover, C., Mete, S. E., Okur, E., D'Mello, S. K., & Arslan Esme, A. (2019, May 4–9). Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms [Conference session]. *CHI'19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, UK.
- 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.
- Brophy, J. E. (1984). *Teacher behavior and student achievement* (No. 73). Institute for Research on Teaching, Michigan State University.
- 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*. Association for Supervision and Curriculum Development.
- 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)* (pp. 1638–1653). Association for Computational Linguistics.
- 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. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference —LAK '17* (pp. 218–227). Association for Computing Machinery.
- 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 517–520). IEEE.
- 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* (pp. 1–13). Association for Computing Machinery.
- 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*, 451–464. <https://doi.org/10.3102/0013189X18785613>
- Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In W. Holmes & K. Porayska-Pomsta (Eds.), *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. <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, 43*, 587–614.
- Madaio, M., Blodgett, S. L., Mayfield, E., & Dixon-Rom'an, E. (2022). Beyond “fairness”: Structural (in) justice lenses on AI for education. In W. Holmes & K. Porayska-Pomsta (Eds.), *The ethics of artificial intelligence in education* (pp. 203–239). Routledge.
- Madaio, M., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-designing check lists 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). Association for Computing Machinery.
- Nystrand, M., Gamoran, A., Kachur, R., & Prendergast, C. (1997). *Opening dialogue*. 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, 24*, 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* (pp. 973–979). Association for Computing Machinery.
- Priestley, M., Biesta, G. J. J., Philippou, S., & Robinson, S. (2015). The teacher and the curriculum: Exploring teacher agency. In D. Wyse, L. Hayward, & J. Pandya (Eds.), *The SAGE handbook of curriculum, pedagogy and assessment* (pp. 187–201). SAGE.
- 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*. <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.
- Steinberg, M.P., & Sartain, L. (2015). Doesteacherevaluation 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 [Preprint]. *arXiv*.
- 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*, 63, 574–605.
- Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, 102(7), 3628–3651.
- 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.

Authors

DOROTTYA DEMSZKY, PhD is an assistant professor in education data science at Stanford University. Her research focuses on developing and deploying natural language processing measures for improving instruction.

JING LIU, PhD is an assistant professor in education policy at the University of Maryland and a research affiliate at the IZA Institute of Labor Economics. He uses novel tools to understand the causes of education inequality and inform effective policy solutions that will combat inequality and improve educational effectiveness.

HEATHER C. HILL, PhD is the Hazen-Nicoli Professor in Education at the Harvard Graduate School of Education. Her primary work focuses on teacher and teaching quality and the effects of policies aimed at improving both.

DAN JURAFSKY, PhD is professor of linguistics and professor of computer science at Stanford University. He studies natural language processing and its application to the social and cognitive sciences.

CHRIS PIECH, PhD is assistant professor of computer science education at Stanford University. His research uses machine learning to understand human learning.

Manuscript received August 11, 2022

First revision received January 3, 2023

Second revision received March 7, 2023

Accepted March 23, 2023