ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-875

Improving Teachers' Questioning Quality through Automated Feedback: A Mixed-Methods Randomized Controlled Trial in Brick-and-Mortar Classrooms

Dorottya Demszky
Stanford University

Jing Liu
University of Maryland, College Park

Heather C. Hill
Harvard University

Shyamoli Sanghi
Stanford University

Ariel Chung
University of Maryland, College Park

While recent studies have demonstrated the potential of automated feedback to enhance teacher instruction in virtual settings, its efficacy in traditional classrooms remains unexplored. In collaboration with TeachFX, we conducted a pre-registered randomized controlled trial involving 523 Utah mathematics and science teachers to assess the impact of automated feedback in K-12 classrooms. This feedback targeted "focusing questions" – questions that probe students' thinking by pressing for explanations and reflection. Our findings indicate that automated feedback increased teachers' use of focusing questions by 20%. However, there was no discernible effect on other teaching practices. Qualitative interviews revealed mixed engagement with the automated feedback: some teachers noticed and appreciated the reflective insights from the feedback, while others had no knowledge of it. Teachers also expressed skepticism about the accuracy of feedback, concerns about data security, and/or noted that time constraints prevented their engagement with the feedback. Our findings highlight avenues for future work, including integrating this feedback into existing professional development activities to maximize its effect.

VERSION: November 2023

Suggested citation: Demszky, Dorottya, Jing Liu, Heather C. Hill, Shyamoli Sanghi, and Ariel Chung. (2023). Improving Teachers' Questioning Quality through Automated Feedback: A Mixed-Methods Randomized Controlled Trial in Brick-and-Mortar Classrooms. (EdWorkingPaper: 23-875). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/8pnw-5q67

Improving Teachers' Questioning Quality through Automated Feedback: A Mixed-Methods Randomized Controlled Trial in Brick-and-Mortar Classrooms[*]

Dorottya Demszky

Jing Liu

Heather C. Hill

Shyamoli Sanghi

Ariel Chung

Abstract:

While recent studies have demonstrated the potential of automated feedback to enhance teacher instruction in virtual settings, its efficacy in traditional classrooms remains unexplored. In collaboration with TeachFX, we conducted a pre-registered randomized controlled trial involving 523 Utah mathematics and science teachers to assess the impact of automated feedback in K-12 classrooms. This feedback targeted "focusing questions" – questions that probe students' thinking by pressing for explanations and reflection. Our findings indicate that automated feedback increased teachers' use of focusing questions by 20%. However, there was no discernible effect on other teaching practices. Qualitative interviews revealed mixed engagement with the automated feedback: some teachers noticed and appreciated the reflective insights from the feedback, while others had no knowledge of it. Teachers also expressed skepticism about the accuracy of feedback, concerns about data security, and/or noted that time constraints prevented their engagement with the feedback. Our findings highlight avenues for future work, including integrating this feedback into existing professional development activities to maximize its effect.

Keywords: computer-assisted instruction, natural language processing, automated teacher feedback, randomized controlled trial, focusing questions

[*] Demszky and Liu contributed equally to this research. Dorottya (Dora) Demszky (ddemszky@stanford.edu) is an Assistant Professor in Educational Data Science at Stanford University. Jing Liu (jliu28@umd.edu) is an Assistant Professor in Education Policy at the University of Maryland, College Park and a Research Associate at the IZA Institute of Labor Economics. Heather Hill (heather_hill@gse.harvard.edu) is a Hazen-Nicoli Professor in Education at the Harvard Graduate School of Education. Shyamoli Sanghi (shyamoli@cs.stanford.edu) was a masters student in Computer Science at Stanford University at the time of the study, and currently a machine learning engineer at TeachFX. Ariel Chung (tchung13@umd.edu) is a Ph.D. student in Education Policy at the University of Maryland, College Park. The authors are grateful for financial support from the Learning Agency through its Tools Competition Award. Youngsun Lee provided excellent research assistance. This work would not have been possible without the partnership of the TeachFX team — special thanks to Jamie Poskin, Berk Coker and Alyssa Van Camp for their close collaboration. We also thank teachers and coaches who volunteered to be interviewed for the study.

1 Introduction

In recent years, automated feedback to teachers using natural language processing (NLP) techniques has emerged as a promising tool for teacher professional learning. Such automated feedback tools take a recording of a teacher's lesson as input, transcribe and analyze the recording, and deliver insights to the teacher to facilitate reflection and instructional improvement. Compared to human classroom observation and feedback, automated feedback is cost-efficient, easier to scale, and can be delivered to teachers privately, in a timely and frequent manner. Thus, such tools have received heightened interest from both educational technology firms and scholars alike.

To date, two randomized controlled trials in online environments have provided evidence on how automated feedback tools might support instructors, enhance instructional practice, and improve student outcomes. In an online computer science course offered by Stanford University, researchers randomly assigned half of the instructors to receive an email reminder to check automated feedback on their uptake of student ideas. They found that instructors in the treatment group improved their frequency of taking up student ideas by 13 percent compared with the control group (Demszky et al., 2023a). Similarly, in an online, one-on-one tutoring program that aimed to improve high school students' research skills, tutors who were offered automated feedback improved their uptake of student contributions by 10 percent as compared with tutors who did not have access to the feedback (Demszky & Liu, 2023). Findings from both studies also suggest that students taught by instructors who received feedback had more favorable perceptions of their learning experience compared with instructors who did not receive such feedback.

Despite the initial encouraging findings in online teaching settings, to our knowledge, there exists no rigorous experimental evaluation of whether or how automated feedback might work in K-12 in-person learning contexts. The lack of studies in this field is likely due to at least two reasons. First, it is technically challenging for teachers to record lessons with high audio quality in in-person K-12 environments, which can feature substantial background noise and multiple students talking over each other. Second, K-12 teachers today are overwhelmed by their daily work and the influx of technological tools, leaving little time and mental space to try an additional tool that requires their active participation to reap potential benefits.

The challenges to implementing automated feedback in K-12 contexts also underscore the need to understand how teachers engage with and perceive the utility of such feedback before further progress can be made in this area. Prior research on feedback targeted to students sheds some light on factors that might affect teachers' acceptance of feedback, including its provider (human vs. algorithm) (Wilson et al., 2021), content (quantitative summaries vs. subjective comments) (Saviano et al., 2023), and teachers' technology-related utility beliefs (Backfisch et al., 2021). We are, however, only aware of one study that touches on how teachers perceive automated feedback on their instruction. Specifically, Jacobs et al. (2022) find that teachers see automated feedback as a valuable vehicle for self-reflection, but that perceptions of accuracy can impact their engagement with such feedback.

This mixed-method study is the first that combines a randomized controlled trial and qualitative interviews to *test* the impact of and *describe* teacher engagement with automated feedback on instruction in K-12 in-person classrooms. We chose to focus on one specific high-leverage teaching practice — asking focusing questions, which refers to how a teacher "press[es] [students] to communicate their thoughts clearly, and expect[s] them to reflect on their thoughts and those of their classmates" (Leinwand, 2014; Demszky & Hill, 2023). In addition, as measuring focusing questions only relies on transcripts of teacher voices, which are often more accurate than those of student voices, centering teachers' questioning helps improve the precision of the feedback.

We conducted a randomized controlled trial in partnership with TeachFX, a company that delivers feedback to teachers based on classroom recordings via a phone application. We leveraged TeachFX's newly established partnership with the state of Utah, where 523 mathematics or science teacher users participated in the study. All teachers had access to TeachFX's platform, standard feedback that TeachFX offers, and a class report email that nudged teachers to view the platform once they made a recording. We randomly assigned half to receive additional targeted feedback on their use of focusing questions via a weekly email and via a link to this feedback on the TeachFX app. The control group did not receive such feedback on focusing questions. We are thus testing the effectiveness of sending teachers one additional distinct piece of automated feedback on a high-leverage teaching practice among teachers who already have access to TeachFX's automated feedback. While not a direct test of the effects of automated feedback vs. no such feedback, our approach has the benefits of not denying any participating teacher access to TeachFX services.

In this paper, we seek to address the following research questions:

1. To what extent do teachers engage with the automated feedback on focusing questions?

2. Does the automated feedback on focusing questions impact instruction, including teachers' use of focusing questions, student talk time, and student reasoning? [1]

We augmented these questions with a third question, which we seek to answer with our qualitative interviews in this mixed-methods study.

3. How do teachers perceive the automated feedback on both focusing questions and other teaching practices provided by the TeachFX platform? What are the barriers for them to engage with the feedback?

We find that the additional feedback significantly increases treatment teachers' use of focusing questions by about 20% compared to control group teachers ($p < 0.01$). However, we did not find evidence that this improvement translates into measurable impacts on other

[1] The first two research questions are pre-registered (https://www.socialscienceregistry.org/trials/11258). The pre-registration also included an additional research question about how the feedback changes teachers' perception of their own instruction. It also included heterogeneity analyses (e.g. how the impact of the intervention varies by teacher characteristics). However, due to a low survey response rate (n=74, 20%), we had to exclude these research questions from our study.

teaching practices such as building on student contributions; nor did it increase student talk or mathematical reasoning. We also observe stronger treatment effects among teachers who consistently recorded their lesson for a longer period of time, although we cannot tease out unobserved factors that motivate certain teachers to persistently use the feedback tool.

To further probe teacher engagement with and perceptions of the feedback, we interviewed 13 teacher participants from both the treatment and control groups. This qualitative data helps illuminate the factors that either help or hinder teachers' engagement with the automated feedback tool. Our findings suggest that while only a portion of teachers were aware of the feedback emails, those who read it perceived the feedback a valuable tool for reflecting on their questioning practices. We also found imprecise transcripts, concerns about data privacy, and lack of motivation and time to be among the most salient factors that prevent teachers from engaging with automated feedback.

2 Related Work

2.1 Automated Feedback to Teachers on Classroom Discourse

Providing teachers with formative feedback grounded in their practices can improve both their instruction (Shute, 2008) and student outcomes (Kraft et al., 2018). Such feedback is often provided through instructional coaches, who observe teachers' classroom, scaffold teachers' reflection on their practice and provide them with feedback and suggestions to improve. While expert coaching can be highly beneficial, it is expensive and time-consuming, and is thus challenging to scale. Recent efforts have sought to complement expert coaching by leveraging technology to facilitate teachers' self-guided improvement at scale, by allowing them to revisit (Sherin & Dyer, 2017) and receive automated feedback on their recorded lessons (e.g., Jacobs et al., 2022). Automated feedback is generated by automatically transcribing classroom recordings, computationally analyzing the transcripts and surfacing insights from these analysis to the teacher. In addition to measuring the *quantity* of talk (e.g. teacher talk time) (Wang et al., 2013), researchers have developed several natural language processing (NLP) measures that can analyze the *quality* of teacher and student talk in classroom transcripts. Such NLP measures tend to focus on detecting 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 key aspects of classroom discourse (Samei et al., 2014; Donnelly et al., 2017; Kelly et al., 2018; Jensen et al., 2020; Demszky et al., 2021). For example, Hunkins et al. (2022) introduces measures that identify growth mindset- and autonomy supportive talk moves in teacher utterances derived from recordings of middle school mathematics classrooms. Moving beyond measurement to teacher feedback, Suresh et al. (2021) introduces 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. Similarly, Demszky et al. (2023a) introduces the M-Powering Teachers application that provides feedback to teachers on their talk time and uptake of student ideas.

While new methods and tools for automated teacher feedback are emerging, the field still lacks data and rigorous evidence about whether such tools indeed improve teaching and student outcomes. For the limited number of tools studied by scholars, the results vary. Jacobs et al. (2022) found that the TalkMoves application was perceived positively by K-12 mathematics teachers. Using a pre/post design, the authors observed a positive but not statistically significant trend for the impact of the feedback on teacher practice; the lack of significance is potentially due to its small sample size (n=21). Another tool, M-Powering Teachers, has demonstrated success in improving instructional practice, specifically teachers' uptake of student ideas, as well as student satisfaction in virtual small group (Demszky et al., 2023a) and 1:1 settings (Demszky & Liu, 2023). This study fills the gap by being — to our knowledge — the first randomized controlled trial to test the impact of automated discourse-based teacher feedback in in-person K-12 instruction settings.

2.2 Qualitative Work on Teachers' Technology Integration

While research on teachers' perceptions and use of automated feedback in classrooms is limited, a rich literature describes factors influencing teachers' integration of technology in teaching. Many studies use the Theory of Planned Behavior (TPB) proposed by Ajzen (1991) as a framework to explain how teachers' attitudes toward technology, perceived social norms, and perceived degree of control shape their intention and actual behavior. For example, Teo (2011) analyzed self-report data from about 600 teachers to test hypotheses based on TPB. He found that teachers' perceived usefulness, perceived ease of use, facilitating conditions, and attitudes towards use significantly influenced teachers' intention to use technology in teaching. On the flip side, research has also described the barriers teachers face when trying to integrate technology, including lack of technology access and support (Fletcher, 2006), the fact that new technology is often less reliable (Butler & Sellbom, 2002), and not enough time to learn new technology and incorporate it during class planning (Bauer & Kenton, 2005).

More recently, Spiteri & Chang Rundgren (2020) conducted a systematic literature review of 27 studies focused on pre-service and in-service primary school teachers' use of technology. They identified four factors influencing teachers' use of technology: teachers' knowledge, teachers' skills, teachers' attitudes, and school culture. In terms of knowledge, effective technology integration occurs when teachers have content knowledge about the subject matter, knowledge of how to use the tool, and pedagogical knowledge about how to advance learning goals through the technology (Mishra & Koehler, 2006). Developing teachers' technology skills and multimodal teaching methods (Wake & Whittingham, 2013) and cultivating teachers' trust and self-efficacy toward technology (Kwon et al., 2019) is crucial for facilitating such effective integration. Support from the school and district and school leadership, including incentives and professional development provided to teachers, can also influence teachers' integration of technology in their classrooms (Kafyulilo et al., 2016; O'Dwyer et al., 2004).

3 Background

We conducted the study in partnership with TeachFX[2], an education technology platform designed to help teachers improve their instruction by providing them with automated feedback. Teachers use a mobile application to record their instruction. TeachFX then automatically transcribes and analyzes the recording using natural language processing tools. Within a day of the recording, all teachers receive an email to view their class report on the TeachFX platform. The class report includes the entire transcript of the class as well as insights related to student and teacher talk in the class, including teacher talk time, wait time, long student contributions, and a word cloud representing the frequency of terms used by the students and the teacher. Appendix D contains all insights available to teachers at the time of the study. We deployed the insight on focusing questions to a random subgroup of teachers participating in the study (Figure 2) — see more details in the Section 4.

The study ran for **five months** between October 10, 2022 and March 10, 2023. We ended the study in March because of the start of the standardized testing season, which interfered with teachers' bandwidth to use the tool. The study was approved under Stanford IRB 66094.

3.1 Participants

The study involved 523 TeachFX users from school districts in Utah. TeachFX had recently formed a new research partnership with the state of Utah, as part of which teachers were encouraged to use TeachFX and received professional development opportunities related to automated feedback. Because these new users were not biased by exposure to automated feedback beforehand, they were ideal participants for the study. Our experiment involved users who made their first recording with TeachFX after the beginning of the study.

[2] https://teachfx.com/

Information about participants. The only information TeachFX collects about users is their role in the school (Administrator, Instructional Coach, Teacher). Thus, the information we have about teacher demographics is limited to teachers who filled out a survey administered by TeachFX at the end of the study (n=74, response rate=20%, see questions in Appendix B). All teachers were offered a $10 Starbucks gift card for completing the survey. Teachers' assigned condition did not have an impact on survey completion (see Section 4.4). The survey asked teachers to self-identify their gender, race/ethnicity, subjects taught, grade levels taught, years of teaching experience, and whether they teach their own classrooms or instead offer special types of instruction (e.g. small groups, tutoring).

Analytical sample. Following our pre-registered filtering process, we excluded participants from the analytical sample who did not meet our study criteria. First, we excluded teachers who recorded lessons for fewer than two weeks. Since our feedback on focusing questions is designed for mathematics/science instruction, we excluded teachers who indicated that they did not teach mathematics/science in the endline survey. If they did not respond to the survey, we observed their recordings and excluded teachers for whom most recordings were in subjects other than mathematics/science. We also focused on teachers in self-contained classrooms and thus excluded participants who indicated in the endline survey that they do not have their own classroom and/or if they indicated via TeachFX that they were recording on behalf of another teacher (e.g. as an instructional coach). All of these filtering steps were performed blind to teachers' assigned conditions in the study. Our final analytic sample includes 369 mathematics/science teachers.

Demographic characteristics. Among teachers who filled out the survey, 84% are female, 14% are male, and 2% preferred not to report their gender. 82% are White or Caucasian, 4% are Hispanic or Latinx, 4% are Asian, and the rest of the teachers identified themselves as multiracial. In terms of teaching experience, the majority of teachers have 8 or more years of experience (54%); 15% have 3-4 years of experience; the rest vary. 38%

of teachers teach in elementary grades, 28% of teachers teach in middle grades, and 28% of teachers teach in high school grades. About 86% of teachers have a regular (non-special education) classroom.

3.2 Incentives for Teachers to Record Lessons

Because of declining participation as the school year progressed, TeachFX incentivized teachers to record via a raffle. These raffles involved the opportunity to win Amazon gift cards worth $250. TeachFX offered four raffles during the study period to all users in Utah, regardless of study condition. The raffles incentivized teachers to record at least once per week during each of the four different periods (Oct 10-Nov 10, Dec 5-Dec 23, Feb 6-Feb 17, Feb 27-March 10) for a chance to win a $250 gift card during that period.

TeachFX also sent out a "Words you used the most" email at the end of October to all Utah teachers who had recorded that month. For each teacher, this email displayed the most frequent words spoken by the teacher and their students (through teacher and student word clouds for the month) that month. It also included a reminder about the October-November raffle time period that would end mid-November. It encouraged teachers that they were on track to be considered for the raffle prize if they continued to record every week. This email also demonstrated to teachers the fact that they could receive similar insights if they continued to record consistently.

3.3 Statistics of Recorded Lessons

The analytic sample consists of 1,450 recorded lessons. This sample was obtained after our pre-registered filtering processing, as part of which we removed participants who did not meet study criteria (see Section 3.1), removed recordings shorter than 10 minutes, and removed recordings that teachers made past completing five weeks. The sample contains 4.31 unique recordings per teacher on average (SD=4.68, range=1-37) and 2.46 weeks of recordings per teacher on average (SD=1.53, range=1-5). The average duration of recordings is 38 minutes (SD=18, range=10-120). TeachFX computes the percentage of student talk transcribed, which can be an indication of recording quality (e.g., noisiness), which affects downstream transcription, especially for student speech that tends to generally cause performance issues for transcription systems. In the sample, 40% of student talk is transcribed on average (SD=19%, range=0-94%). Given the variation in the data in terms of recording quality, we control for this variable in our analyses.

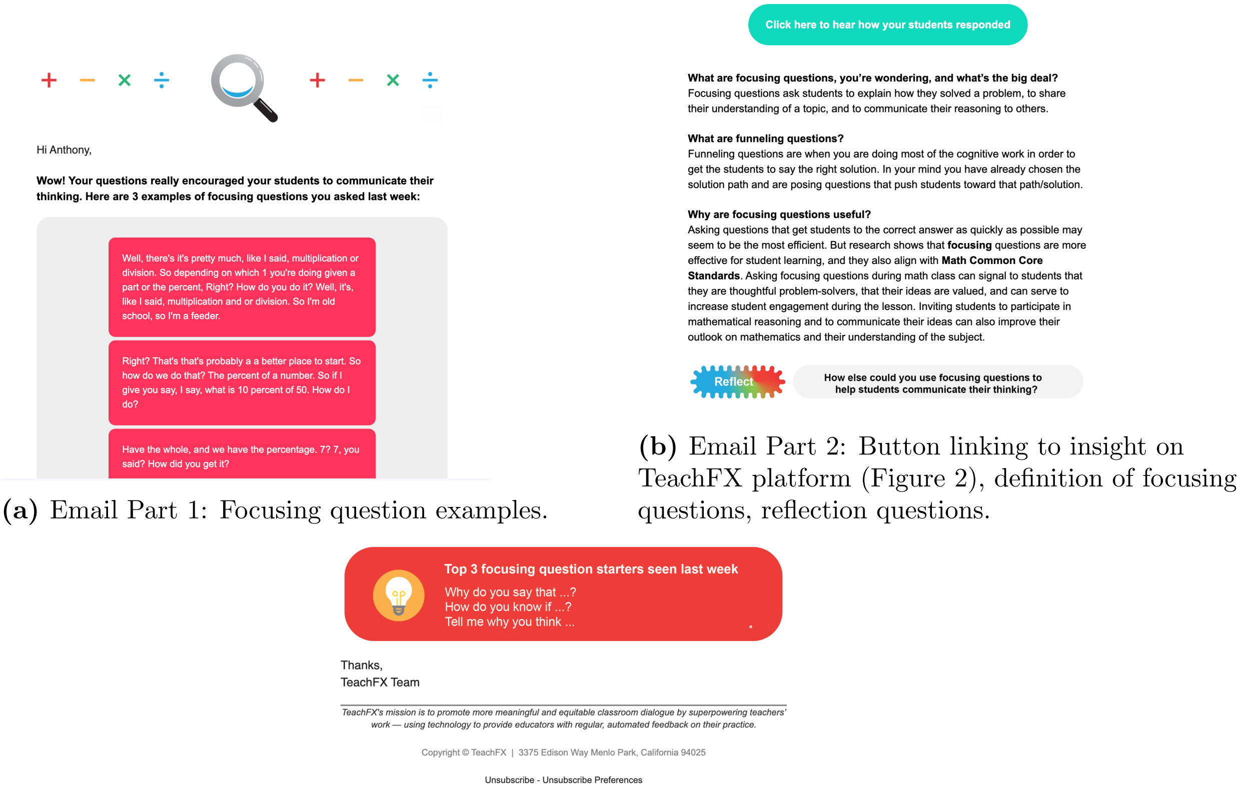
4 Randomized Controlled Trial

We conducted a randomized controlled trial to evaluate the effectiveness of providing feedback to mathematics and science teachers on focusing questions. Teachers were randomly assigned to the control or the treatment group after they made their first recording during the study. The random assignment was made via a hashing function within the TeachFX platform, which is similar to a coin flip.

4.1 Detecting Focusing Questions

Focusing questions were identified in classroom transcripts by our machine learning model. Given a transcript of a class recording, we extract teacher utterances and feed them to a binary classification model, which tells us whether or not each utterance is a focusing question. We obtain this model by fine-tuning Bert-base (Devlin et al., 2018)[3], a pre-trained language model, on labeled data from the NCTE elementary mathematics classroom dataset (Alic et al., 2022), which we augment with 694 annotated examples from TeachFX transcripts to facilitate adaptation to the target domain. To obtain labels for TeachFX data, we recruited two experienced mathematics instructional coaches to annotate teacher utterances for the presence of focusing questions using the annotation guide described in Alic et al. (2022).

[3] We experimented with RoBERTa-base(Liu et al., 2019) as well but found that Bert performed better.



Hi Anthony,

Wow! Your questions really encouraged your students to communicate their thinking. Here are 3 examples of focusing questions you asked last week:

Well, there's it's pretty much, like I said, multiplication or division. So depending on which 1 you're doing given a part or the percent, Right? How do you do it? Well, it's, like I said, multiplication and or division. So I'm old school, so I'm a feeder.

Right? That's that's probably a a better place to start. So how do we do that? The percent of a number. So if I give you say, I say, what is 10 percent of 50. How do I do?

Have the whole, and we have the percentage. 7? 7, you said? How did you get it?

(a) Email Part 1: Focusing question examples.

What are focusing questions, you're wondering, and what's the big deal?
Focusing questions ask students to explain how they solved a problem, to share their understanding of a topic, and to communicate their reasoning to others.

What are funneling questions?
Funneling questions are when you are doing most of the cognitive work in order to get the students to say the right solution. In your mind you have already chosen the solution path and are posing questions that push students toward that path/solution.

Why are focusing questions useful?
Asking questions that get students to the correct answer as quickly as possible may seem to be the most efficient. But research shows that **focusing** questions are more effective for student learning, and they also align with **Math Common Core Standards**. Asking focusing questions during math class can signal to students that they are thoughtful problem-solvers, that their ideas are valued, and can serve to increase student engagement during the lesson. Inviting students to participate in mathematical reasoning and to communicate their ideas can also improve their outlook on mathematics and their understanding of the subject.

How else could you use focusing questions to help students communicate their thinking?

(b) Email Part 2: Button linking to insight on TeachFX platform (Figure 2), definition of focusing questions, reflection questions.

Top 3 focusing question starters seen last week
Why do you say that ...?
How do you know if ...?
Tell me why you think ...

Thanks,
TeachFX Team

TeachFX's mission is to promote more meaningful and equitable classroom dialogue by superpowering teachers' work — using technology to provide educators with regular, automated feedback on their practice.

Copyright © TeachFX | 3375 Edison Way Menlo Park, California 94025

Unsubscribe - Unsubscribe Preferences

(c) Email Part 3: Example focusing question starters derived from transcripts of study participants.

Figure 1: Email about focusing questions, sent once a week to treatment group teachers.

The fine-tuned model achieves an 84% accuracy on a held-out set of TeachFX transcripts. We include more details about the model in Appendix A.

4.2 Treatment Email & Feedback

Teachers in the treatment group received an email early every Tuesday morning which contained both the number of focusing questions they asked in all class recordings in the previous week as well as a display of, at most, the top 3 chosen focusing questions.[4] Figure 1 shows an example email that was sent to treatment group teachers. The top questions (Figure 1a) were identified by two expert annotators, mathematics instructional coaches with decades of

[4] We decided to send the email early mornings to maximize the chances of teachers reading it before their workday; we sent it on Tuesday rather than Monday as we expected teachers to be the busiest on Mondays.

experience to ensure that we select the best questions to reinforce and extend this teaching move.[5] The email also contained an explanation of focusing questions (Figure 1b) as well as a link to the focusing questions insight page on the TeachFX app and to a more detailed blog post[6] explaining what focusing questions are and how to ask more focusing questions of students. Further, it included the top 3 focusing question starters seen in the week across treatment group teachers (Figure 1c), identified by the same expert annotators. The control group did not receive the email, nor did it have access to the focusing question insights through the TeachFX app.

After a teacher completed 5 weeks of recordings, we considered their participation in the research study complete and stopped sending emails to that teacher. At the conclusion of the study, TeachFX sent the teachers a survey that asked about their demographic information and perception of their instruction (Section 3.1). The full survey and the text of the email with the survey link are included in Appendix B.

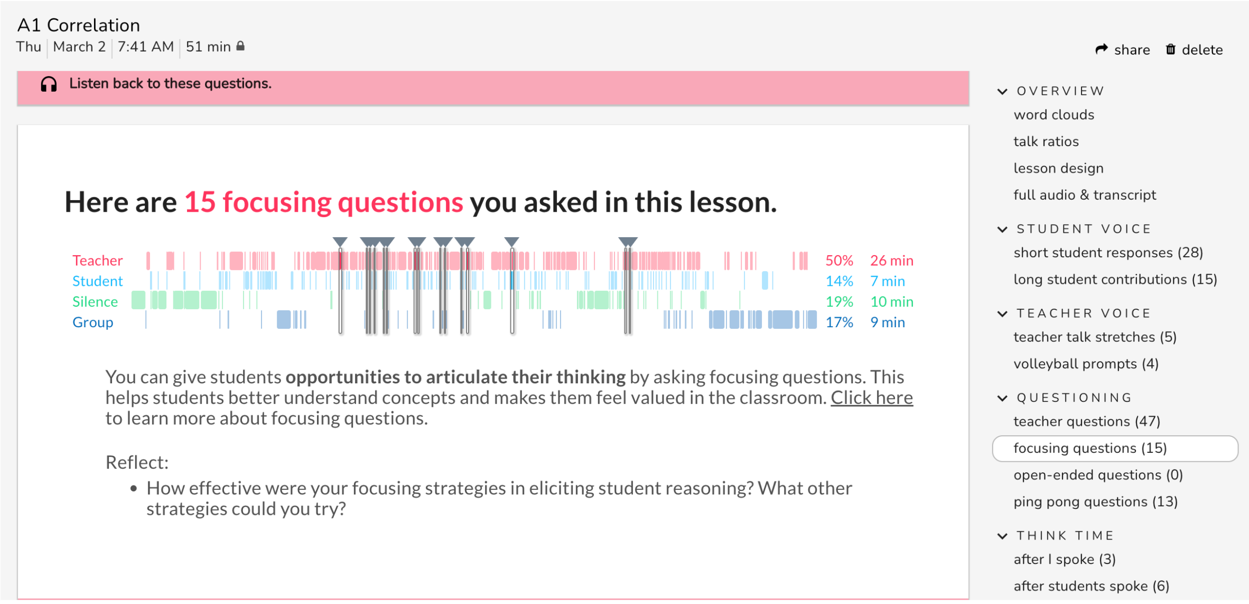
4.3 Measures of Outcomes

To understand how often teachers engage with the automated feedback (RQ1), we quantify their interactions with the email and the feedback page. To understand the impact of our intervention on instructional practice (RQ2), we quantify several discourse features.

Engagement with automated feedback. We quantify engagement with the feedback on focusing questions by tracking whether a teacher opened the focusing questions email and whether they viewed the focusing question insight on the TeachFX platform. We additionally measure if a teacher viewed any part of their class report on the TeachFX platform to compare

[5] We conducted the selection step manually because we do not yet have an automated way of ranking focusing questions, and we wanted to ensure that indeed the best examples are picked. Although the selection of the top 3 questions was manual, we conducted a comparison that shows that annotators performed this selection from the automatically identified questions 7 times faster than if they had to select examples from all teacher questions. Future work can explore effective approaches to automate the ranking process entirely.

[6] https://medium.com/dorademszky/resources-for-focusing-questions-47bc6cdd9953



Here are 15 focusing questions you asked in this lesson.

You can give students **opportunities to articulate their thinking** by asking focusing questions. This helps students better understand concepts and makes them feel valued in the classroom. Click here to learn more about focusing questions.

Reflect:

- How effective were your focusing strategies in eliciting student reasoning? What other strategies could you try?

Figure 2: Screenshot of the Focusing Questions insight within the TeachFX app.

engagement with automated feedback between control and treatment group teachers.

Discourse features. Our primary discourse-related outcome is the rate of focusing questions per hour. In addition, we quantify the percentage of student talk time, the rate of teachers' uptake of student ideas (Demszky et al., 2021), and the rate of student reasoning per hour (Demszky & Hill, 2023). We use these features because we observed positive correlations with student outcomes in prior work (Demszky & Hill, 2023; Demszky et al., 2021, 2023b; Demszky & Liu, 2023), and because we hypothesized that an increase in focusing questions would lead to an improvement along these discourse features. In fact, in our pre-intervention transcripts, too, the rate of focusing questions correlates significantly with student talk percentage ($\rho = 0.2$, $p < 0.001$), teachers' uptake of student ideas ($\rho = 0.18$, $p < 0.001$) and student reasoning ($\rho = 0.61$, $p < 0.001$). Thus, we expected that an improvement in focusing questions would increase student talk time, teachers' uptake of student ideas, and student reasoning during the intervention.

	Control Mean	**Treatment Mean**	**P Value**	**N**
Female	0.82	0.82	0.98	95
White	0.8	0.88	0.29	95
Teaches Mathematics	0.76	0.84	0.31	95
Teaches Elementary	0.42	0.46	0.71	95
Teaches Middle School	0.31	0.2	0.22	95
Teaches High School	0.31	0.22	0.32	95
Duration (minutes)	27.03	30.36	0.07	523
Focusing rate	28.15	26.88	0.58	523
Uptake rate	4.81	5.04	0.78	520
Student reasoning rate	3.35	3.32	0.96	520
Student talk percentage	21.91	21.78	0.94	523
Percentage of student talk transcribed	0.47	0.46	0.51	501
Week of first recording	7.53	7.79	0.62	523
Opened class report	0.13	0.12	0.57	523
Attrition				
Number of weeks teacher recorded	2.48	2.62	0.32	523
Number of unique recordings	1.69	1.89	0.26	523
Survey completed	0.17	0.2	0.4	523
Invalid recording	0.36	0.33	0.55	523

Table 1: Data used here include all participants before data filtering. Randomization check shows that the treatment and control group characteristics do not differ significantly. Rates for different discourse moves describe frequencies per minute. The week of the first recording represents the week during the RCT period when the teacher made their first recording. The attrition values show that attrition in the data (due to validity, lack of recording, or survey completion) are not affected by the randomization.

4.4 Validating Randomization

To verify whether our randomization created groups that were balanced on observable variables, 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 recorded lesson prior to receiving feedback. We use all participants in the study regardless of the number of weeks of recordings or whether their recordings are valid (n=523). As Table 1 shows, other than that the duration of the first recording is marginally significant, we do not find statistically significant differences between conditions in any of the teacher demographics and discourse features of the first section. This analysis suggests that any differences we observe later in the study are likely driven by the effects of the intervention.

4.5 Attrition Analyses

We also conducted an attrition analysis to examine whether the treatment and control conditions suffered from differential attrition. The results are presented in the bottom panel of Table 1. Attrition in our data occurred when teachers recorded for fewer than five weeks (and thus did not complete the study), made invalid recordings, or did not fill out the final survey. We find no differential attrition in the sample, suggesting that the intervention did not have a significant impact on teachers' likelihood of recording valid lessons and completing the survey.

4.6 Regression Analyses

To understand what impact the intervention had on teachers' practice (RQ2), we conducted a preregistered intent-to-treat analysis with an ordinary least squares regression. This analysis compares teachers' discourse features regardless of whether they chose to engage with the automated feedback. The models are specified as below:

$$Y_{it} = \beta_1 T_i + \beta_2 \boldsymbol{X}_i + \beta_3 \boldsymbol{M}_{it} + \varepsilon_{it} \tag{1}$$

where Y_{it} refers to a particular dependent variable for teacher $i's$ transcript t; T is a binary variable that indicates the treatment status, with a value of 1 indicating the treatment condition and 0 otherwise; $\boldsymbol{X}$ is a vector of teacher-level covariates, $\boldsymbol{M}$ is a vector of transcript metadata; β_1 is the parameter of interest, which measures the treatment effects of our intervention on teacher outcomes; and ϵ indicates the residuals. We conduct analyses at the transcript level and cluster standard errors at the teacher level to account for repeated observations within a teacher.

We fit this model to estimate the effects of the treatment on each dependent variable described in Section 4.3 above. Specifically, the outcomes we consider include the number of times a teacher asked a focusing question per hour, the number of times they took up student ideas per hour, the number of student reasoning instances per hour, and the student talk percentage. We use hourly rates instead of raw counts to account for differences in recording duration.

We use the following binary variables as teacher covariates $\boldsymbol{X}$ across all models derived from the final survey: female, white, having had at least 5 years of teaching experience, teaching mathematics, and teaching in elementary/middle/high school. Missing survey data was assigned a value of zero and we include a binary indicator for whether the data was imputed. We also include teachers' baseline discourse features in their first recording as a covariate: the rate of focusing questions, rate of teachers' uptake of student ideas, rate of student reasoning, and student talk percentage and percentage of student talk transcribed. We also include two variables as transcript covariates $\boldsymbol{M}$: indicators for the week of recording for the teacher during the RCT and percentage of student talk transcribed (as an indication of recording quality).

To verify that our estimates are not significantly influenced by the choice of demographic control variables, especially given the low survey response rates, we also estimate a version of our models without those variables.

5 Interviews

To better understand teachers' perception of the automated feedback, barriers to access, and suggestions for improvement (RQ3), we conducted qualitative interviews that asked teachers about their experiences engaging with the TeachFX platform and feedback and, for teachers in our treatment group, engagement and perceptions of the feedback on focusing questions. Virtual semi-structured interviews were conducted shortly after the experiment finished. In terms of data analysis, we adopted a convergent approach (Fetters et al., 2013), analyzing quantitative and qualitative simultaneously and comparing the results from both analyses to see how the interview data confirms or helps explain quantitative findings.

5.1 Semi-structured Interview

To further probe participants' perceptions of the automated feedback and identify the factors that promote or hinder their engagement, we conducted semi-structured interviews with 13 teachers. We recruited interviewees by working with TeachFX to email all participants who completed the post-treatment survey. Among all teachers who expressed willingness to participate in the interview, we reached out to a diverse group of study participants based on their reported demographics, years of teaching experience, and treatment status. We interviewed teachers in two waves between March and May of 2023. A total of fifteen teachers responded to the recruitment email. We interviewed six during the first wave and seven during the second wave. Among the thirteen interviewees, seven were from the treatment group, and the other six were from the control group. Eight were working in elementary schools, four in secondary schools, and one in both. All teachers have more than eight years of experience in teaching. In terms of teaching roles, eight interviewees are general education teachers, three are special education teachers, and another two are former teachers who now serve as instruction coaches at the district.

Interviews were conducted virtually over Zoom. Each interview lasted 1-1.5 hours. The interviews asked teachers about their teaching background, previous experience incorporating technology in their teaching, experiences setting up the environment for recording, experiences with the feedback tools, and any feedback they might have for the feedback tool. All interviews were recorded and transcribed using Zoom. Two graduate assistants and a Ph.D. student checked and corrected the auto-generated transcript for accuracy. Participants were compensated $50 per hour for the interview.

5.2 Qualitative Coding and Analysis

We conducted a thematic analysis of the resulting interview data. We started with inductive and deductive coding of the interview transcripts using the NVivo software. Based on the literature in Section 2, the research team developed a codebook to capture the barriers to engagement, ways to enhance engagement, treatment group experience, and feedback specific to TeachFX. Two Ph.D. students worked on coding all interview transcripts. The coders used one interview from the treatment group and one from the control group to conduct intercoder alignment training and ensure that intercoder agreement reached 100% before starting coding individually. After the first round of coding, coders reviewed each other's coded interviews to check for confusion or disagreement. All coding differences were discussed until an agreement was reached. To answer RQ3, the research team came together to discuss the coded data and summarize major themes.

6 Results

6.1 RQ1: To what extent do teachers engage with the automated feedback on focusing questions?

We tracked two key metrics of engagement with the automated feedback: email opens and views of the focusing question insight on the TeachFX platform. Overall, treatment group teachers opened the emails with the feedback at a much higher rate than viewing the insight on the platform. Between 55-61% of teachers opened their emails across weeks (61% for 1st email, 53% for 2nd, 61% for 3rd, 55% for 4th and 56% for 5th), but only 17-22% of them viewed the focusing insight page (22% for 1st week, 17% for 2nd, 20% for 3rd, 17% for 4th, 17% for 5th week). Similarly, while 67% of teachers in the treatment group opened the email at least once, only 40% of them viewed the insight page at least once. On average, teachers opened 1.8 emails (SD=1.9) out of 5 throughout the RCT. These results suggest that email is a more effective delivery method for the feedback than the platform, likely because accessing the platform requires an extra step (teachers clicking the link in the email or visiting the TeachFX platform separately). However, there is significant room for improving teachers'

	(1) **Focusing rate**	(2) **Uptake rate**	(3) **Student reasoning rate**	(4) **Student talk percentage**
Treatment	4.612**	0.274	0.655	0.001
	(1.741)	(0.523)	(0.485)	(0.012)
Control Mean	22.565	3.772	2.896	0.160
R^2	0.346	0.319	0.226	0.244
Observations	533	533	533	533

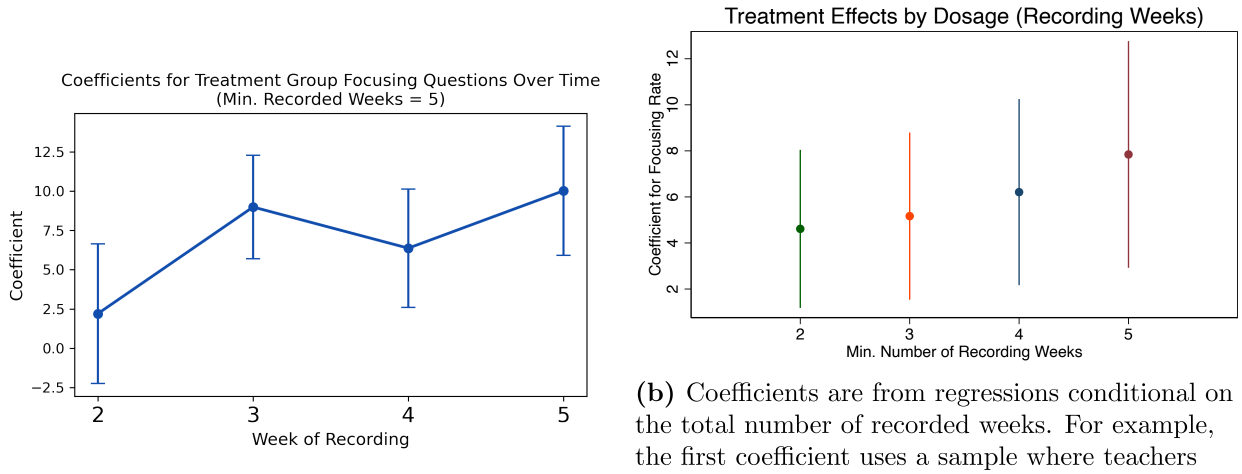
Table 2: Standard errors are in parentheses. ** p<0.01. These models estimate the effect of the automated feedback on focusing questions (treatment) on teachers' discourse features. All models include covariates listed in Section 4.6. We observe a statistically significant impact on focusing questions but not the other discourse features.

consistent engagement with both the email and the platform.

We also observe evidence that the intervention increases views of the TeachFX platform. While teachers in the control group viewed their class reports on the TeachFX platform 15% of the time, teachers in the treatment group opened their class reports 21% of the time, indicating a statistically significant difference ($p < 0.05$). This suggests that the additional email beyond the class report, targeted to a specific teaching practice (focusing questions), increases overall engagement with feedback provided by TeachFX.

6.2 RQ2: What impact does automated feedback on focusing questions have on teachers' instruction?

As shown in Table 2, the treatment significantly increased teachers' use of focusing questions. On average, treatment group teachers asked 4.48 more focusing questions per hour, indicating a 20% increase compared to the control group ($p < 0.01$). This effect size is greater than observed in prior interventions related to automated feedback on teachers' uptake of student ideas (Demszky et al., 2023b; Demszky & Liu, 2023). In contrast, the intervention did not have an impact on other discourse features besides focusing questions. Specifically, we do not see any statistically significant effect on student talk time, student reasoning, or teachers' uptake of student ideas, rejecting our hypothesis that improving teachers' use of focusing questions would lead to an improvement in other aspects of instruction.



(a) Coefficients for the rate of focusing questions in the treatment group, plotted over week of recording. This sample only includes teachers who recorded for at least five weeks.

(b) Coefficients are from regressions conditional on the total number of recorded weeks. For example, the first coefficient uses a sample where teachers recorded at least 2 and up to 5 weeks. The plot shows that the intervention had a greater impact when teachers used the feedback on a consistent basis for a longer period of time.

Figure 3: Treatment effects displayed over recorded weeks (a) and by dosage (b).

To explore how effect sizes for focusing questions vary over time, we consider the sample of teachers who completed the RCT — i.e. those who recorded for at least five weeks (n=83). Given that teachers' persistence to record depends on their motivation and perception of the feedback, our estimates are only suggestive rather than indicating a causal relationship between number of recordings and impact on instruction. As Figure 3a shows, the effect size increases from week 2 to week 3 and then generally stagnates. The greatest effect size hovers around 10 additional focusing questions per hour in week 5, indicating a 59% increase compared to the control group. This temporal pattern is consistent with previous studies (Demszky et al., 2023b; Demszky & Liu, 2023) that show that it takes about 2-3 weeks for the feedback to show its effect and then this effect stagnates or sometimes, diminishes, suggesting promise for varying the feedback after two weeks.

To shed light on how teachers' engagement with the tool might impact treatment effects, we analyze the relationship between effect sizes and the minimum number of weeks a teacher recorded. As Figure 3b shows, when we include teachers who recorded as few as two weeks in our regression (i.e., our main analytic sample in Table 2), the treatment effect is 4.48 focusing questions per minute. When we gradually increase the minimum of recorded weeks, we see a linear growth pattern of treatment effects. Specifically, when we only include teachers who recorded their lessons for a total of five weeks in our analytic sample, the treatment effect reaches 7.78, which is a 73% increase compared to our reported effect. This finding suggests that teachers who decide to record more consistently (e.g. because they like the tool or are more strongly influenced by the incentives to record) experience greater benefits from the feedback.

6.3 RQ3: How do teachers perceive the automated feedback on focusing questions?

After using quantitative data to document teacher engagement with our automated feedback on focusing questions and its impact on teaching practice, we now turn to using qualitative data from our interviews to illuminate teachers' subjective experiences and perceptions of automated feedback. We attend to not only treatment group teachers' perceptions of the feedback on focusing questions, but also teachers' overall experience with the TeachFX platform.

Interviews revealed two themes that help explain how teachers' perceptions of the automated feedback shaped their engagement with it. The first theme, which describes variations in awareness of the feedback on focusing questions, suggests that this feedback did not reach all members of the treatment group. The second theme, general barriers to engagement, references the various challenges teachers face when trying to engage with automated feedback tools on a consistent basis.

Varied Awareness of Feedback on Focusing Questions. Our interview data suggest that using emails to deliver automated feedback on focusing questions was only successful for a portion of teacher participants. Among the seven treatment group teachers we interviewed, only three were clearly aware of the emails and saw them as distinct from class reports they had already been receiving from TeachFX. This finding helps explain the moderate email open rates described in RQ1. However, teachers who did notice the feedback said they found the treatment emails and feedback helpful in improving their use of focusing questions and inviting more student input in classes. For example, one teacher shared how using the transcript to locate and learn from her own focusing questions helped her build more student-centered discussions, "I've been looking a lot at the focusing questions and I like to not just see the number [of focusing questions], but kind of click on it, and see the transcript and skip to the next one and the next one and read my focusing questions and see what were the questions that I asked. And how did the students respond to my questions?" This example illustrates how teachers might benefit from the feedback by using it as a tool to revisit their practice.

In contrast, some of the treatment teachers we interviewed were completely unaware of the the emails containing information about focusing questions. For example, when the interviewer shared screenshots from the treatment email, one teacher noted, "I have not seen an email like that." Others only had vague memories of the treatment emails and were unable to differentiate them from the more general TeachFX class reports they received. Without seeing the treatment emails that clarify and reinforce information about focusing questions, some teachers ended up not understanding what focusing questions were or confusing focusing questions with other types of questions identified on the TeachFX platform (e.g., volleyball questions, ping-pong questions, open-ended questions, etc. that analyze talk time dynamics rather the language of the talk – see documentation in Appendix D). Such confusion may prevent teachers from utilizing feedback on focusing questions or misinterpreting the feedback in such a way that leads them to ask other kinds of questions. As one teacher noted when being asked about focusing questions, "... I never really figured out like what the underlying element was that was tying them [the focusing questions] all together. And so I had a hard time really gaining much valuable information from that...." This suggests that including more details about focusing questions on the platform, or suppressing feedback on other question types, could have helped draw attention to and clarify focusing questions for users who did not read the email.

In sum, teachers' varied engagement with our treatment emails limits the impact our automated feedback can have on them, with some teachers fully capitalizing on the information we provided and actively adjusting their teaching practice and some others not even aware of it. This finding points to the importance of the delivery mechanism of automated feedback in achieving desired benefits.

General Barriers to Engagement. Interview participants reported other factors that inhibited their feedback use. Specifically, some teachers expressed concerns about data privacy, which impacted their willingness to record their classrooms. In addition, teachers reported having to work to maximize recording — and therefore ASR — quality. Some teachers reported problems trusting the quality of ASR and feedback models. Finally, teachers shared that time constraints, both in terms of their overall bandwidth and timing of the feedback in relation to their daily schedule, played an important role in whether they were able to engage with the feedback. We discuss each below.

First, some teachers expressed hesitation to use TeachFX due to concerns about data privacy and association with teacher evaluation and accountability. While TeachFX emphasized that the feedback is private to the teachers and intended only for their professional development, one teacher noted, "Nobody likes listening to themselves and being observed and things like that, so like finding ways to be able to share things that we're happy about without feeling like... I don't know, like you're going to be criticized." One instruction coach also observed reservations from the teachers he worked with, "Can it be accessed by principals? Or can it be accessed by parents, or whose data it actually is? This has been on [teachers'] mind."

Second, some teachers reported facing difficulties in achieving clear recordings and thus receiving accurate transcriptions of the lesson. Given that classrooms are busy and dynamic environments, even teachers who can use their phones to do the recordings would still have to take time out of their already full schedules to go through trials and errors with different set-ups to find a configuration that adequately records classroom discourse. For example, one teacher shared her experience configuring the recording by saying “I had a hard time with the system reading my voice at the beginning. I had to do a lot more [recording], and it took me a long, long time for the system to work on my phone. . . .. There was an update that came out, I think, around November or October, and then, when I did that one, then my phone started working. But before that it wasn’t working on my phone.” Some teachers also noted that their automated transcripts contained errors, especially for student speech, rendering the feedback less precise than what teachers would have preferred. One teacher shared how both she and her colleagues were troubled by the imprecise transcripts, “It has some really obvious flaws in the recording. And so a lot of us are like, ‘Oh, I did not say that.’ . . . I know that that’s a hang-up for a lot of teachers.” Another teacher noted, “And there still are parts were, like, it would say “student voice detected” [instead of transcribing what they said]... You know what they said If I went in and listen to the recording most of the time. Because I know my students most of the time, I could figure out what they were saying.” These findings corroborate those of Jacobs et al. (2022) showing that teachers’ are less likely to engage with feedback they deem inaccurate.

Lastly, while most interviewees praised the platform as being very straightforward and user-friendly, particularly the data visualization and reflection questions embedded in the platform, time constraints appear to be a crucial factor that prevents teachers from taking up the feedback. Some teachers reported that they were too busy to read through all the feedback and information provided in the summary email or on the platform. One teacher shared, “I think for me the hard [thing] is like I didn’t have time to sit and read it when it would come in, and then I would forget about it.” Another teacher further elaborated on the timing issue, saying, “Sometimes it [the TeachFX class report email] wouldn’t come to me until the afternoon, and by then I was done and doing my planning for the day.”

7 Discussions

The fast development of natural language processing techniques for educational contexts provide an unprecedented opportunity to improve teaching and learning. Automated teacher feedback, in particular, has demonstrated promise in supporting teachers to learn about their strengths and identify areas of improvement in their instruction (Demszky et al., 2023a; Demszky & Liu, 2023; Jacobs et al., 2022). In partnership with TeachFX, this study is among the first to show the impact of automated feedback in brick-and-mortar classrooms using a randomized controlled trial with a focus on a high-leverage practice–teachers' use of focusing questions. The quantitative evidence is further augmented by qualitative data that helped uncover how teachers engaged with automated feedback on focusing questions, and more general potential barriers to using such tools.

With five weeks of intervention, the automated feedback improves treatment teachers' use of focusing questions by 20% but does not impact other related teacher practice or student discourse measures. The effect demonstrates a clear dose-dependence pattern, with teachers who recorded their lesson the most consistently reaping greater benefits.

Our qualitative interviews reveal a range of factors that prevent teachers from fully engaging and actively using the feedback, including different levels of awareness of the feedback delivered in emails, concerns about transcript precision and data privacy, and motivation and time commitment to read and use the feedback.

Overall, this study goes beyond prior studies that were primarily conducted in online settings and further confirms that automated feedback can be a promising tool to improve teaching in in-person K-12 classrooms. The fact that a simple email intervention can significantly improve the targeted teaching practice within five weeks is remarkable and suggests promise of this low-cost intervention. Our study also points to many areas that need further development to fulfill the promise of automated feedback tools.

Strategies to enhance engagement with automated feedback. Teachers we interviewed shared important principles as well as diverse strategies to enhance trust, uptake, and use of automated feedback. As a fundamental step, it is important to improve transcription that serves as a bottleneck to accurate feedback. Second, is many teachers reflected that having a human component (e.g. a coach or a peer-learning group) during the TeachFX onboarding process was very useful for them to better trust, understand, and utilize automated feedback. Moreover, given that teachers have busy schedules that require them to constantly multi-task, it is crucial to develop innovative strategies that fit technology seamlessly into their current routines. For example, integrating several tech platforms that teachers use (e.g. for administrative tasks, curriculum development, teacher feedback) into a single platform could reduce teachers' cognitive burden of navigating across platforms. Sending quick weekly reminders before class and after feedback is ready can help teachers consistently record lessons, reflect on and remember feedback. Last but not least, feedback is most powerful when it includes concrete next steps that align with teachers' personal teaching goals. Teachers who set personal goals based on automated feedback were generally more motivated to engage with the feedback and found it more helpful. In addition to reflection questions that help translate descriptive feedback into actionable goals, adding concrete suggestions (such as question templates) and reminding teachers of these suggestions before each class can support teachers in incorporating feedback for improved teaching practices.

Limitations & future work. One principal limitation of this work is the absence of robust metadata, including information about students' and teachers' background, learning outcome data and data on students' beliefs and motivation. In future work, we hope to collect robust metadata to understand the impact of the feedback on downstream student outcomes, and how the impact might vary across different teacher and student populations. A second limitation is that we are not able to disentangle the effect of our automated feedback from feedback that is already provided by TeachFX. A promising area of future work would be to include a control condition that does not receive automated feedback, or another experimental design that allows us to disentangle the impact of different types of automated feedback.

Overall, wee see the integration of the automated feedback into existing professional learning programs as the most promising approach to improving its effectiveness. Instructional coaches can provide scaffolding around the use of automated feedback tools, by reviewing the automated feedback with the teacher and complementing it with their expert suggestions. Such integration into existing coaching sessions can thus remove the perception of automated feedback as “one more thing” on teachers’ plate. Furthermore, instructional coaches can help complement inaccuracies in the transcript or the feedback by listening to pertinent segments of the recording and providing their expert interpretation. Finally, instructional coaches can enhance the reflection process by providing teachers with actionable suggestions grounded in evidence teachers’ transcripts and their automated feedback.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, *50*(2), 179–211.

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.

Backfisch, I., Lachner, A., Stürmer, K., & Scheiter, K. (2021). Variability of teachers' technology integration in the classroom: A matter of utility! *Computers & Education*, *166*, 104159.

Bauer, J., & Kenton, J. (2005). Toward technology integration in the schools: Why it isn't happening. *Journal of technology and teacher education*, *13*(4), 519–546.

Butler, D. L., & Sellbom, M. (2002). Barriers to adopting technology. *Educause quarterly*, *2*(1), 22–28.

Demszky, D., & Hill, H. (2023). The ncte transcripts: A dataset of elementary math classroom transcripts. In *Proceedings of the 18th workshop on innovative use of nlp for building educational applications.*

Demszky, D., & Liu, J. (2023, July). M-powering teachers: Natural language processing powered feedback improves 1:1 instruction and student outcomes. *L@S '23: Proceedings of the Tenth ACM Conference on Learning @ Scale*.

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2023a). Can automated feedback improve teachers' uptake of student ideas? evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*.

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2023b). Can automated feedback improve teachers' uptake of student ideas? evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*, 01623737231169270.

Demszky, D., Liu, J., Mancenido, Z., Cohen, J., Hill, H., Jurafsky, D., & Hashimoto, T. B. (2021). Measuring conversational uptake: A case study on student-teacher interactions. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1638–1653).

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*.

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.

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs—principles and practices. *Health services research*, *48*(6pt2), 2134–2156.

Fletcher, D. (2006). Technology integration: Do they or don't they? a self-report survey from prek through 5th grade professional educators. *AACE Review (formerly AACE Journal)*, *14*(3), 207–219.

Hunkins, N., Kelly, S., & D'Mello, S. (2022). "beautiful work, you're rock stars!": Teacher analytics to uncover discourse that supports or undermines student motivation, identity, and belonging in classrooms. In *Lak22: 12th international learning analytics and knowledge conference* (pp. 230–238).

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.

Kafyulilo, A., Fisser, P., & Voogt, J. (2016). Factors affecting teachers' continuation of technology use in teaching. *Education and Information Technologies*, *21*, 1535–1554.

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`

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`

Kwon, K., Ottenbreit-Leftwich, A. T., Sari, A. R., Khlaif, Z., Zhu, M., Nadir, H., & Gok, F. (2019). Teachers' self-efficacy matters: Exploring the integration of mobile computing device in middle schools. *TechTrends*, *63*, 682–692.

Leinwand, S. (2014). *Principles to actions: Ensuring mathematical success for all.* National Council of Teachers of Mathematics, Incorporated.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers college record*, *108*(6), 1017–1054.

O'Dwyer, L. M., Russell, M., & Bebell, D. J. (2004). Identifying teacher, school and district characteristics associated with elementary teachers' use of technology: A multilevel perspective. *Education policy analysis archives*, *12*, 48–48.

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`

Saviano, M., Del Prete, M., Mueller, J., & Caputo, F. (2023). The challenging meet between human and artificial knowledge. a systems-based view of its influences on firms-customers interaction. *Journal of Knowledge Management*, *27*(11), 101–111.

Sherin, M. G., & Dyer, E. B. (2017). Mathematics teachers' self-captured video and opportunities for learning. *Journal of Mathematics Teacher Education*, *20*, 477–495.

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

Spiteri, M., & Chang Rundgren, S.-N. (2020). Literature review on the factors affecting primary teachers' use of digital technology. *Technology, Knowledge and Learning*, *25*, 115–128.

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)

Teo, T. (2011). Factors influencing teachers' intention to use technology: Model development and test. *Computers & Education*, *57*(4), 2432–2440.

Wake, D., & Whittingham, J. (2013). Teacher candidates' perceptions of technology supported literacy practices. *Contemporary Issues in Technology and Teacher Education*, *13*(3), 175–206.

Wang, Z., Miller, K., & Cortina, K. (2013). *Using the lena in teacher training: Promoting student involement through automated feedback* (Vol. 4). na.

Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G., & MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers & Education*, *168*, 104208.

Appendix A Focusing Questions Model

Original Model: We fine-tuned the large language model BERTforSequenceClassification for a binary classification task of identifying whether or not a teacher utterance was a focusing question or not on the following labeled data. We used $\sim$ 2000 NCTE (National Council for Teachers of English) teacher utterances that which had annotations for whether the teacher utterance was a focusing question, a funnelling question or neither (We considered the funnelling questions and neither category in the category of non-focusing questions). This data had a 90:10 split of non-focusing: focusing questions. To handle the class imbalance, we over-sampled the minority class to enable the training set to have an approximately equal number focusing and non focusing questions. The final model hyperperameters we used were: Learning rate: 2e-5; Max embedding length: 256 tokens; Number of epochs: 3; Optimizer: Adam; epsilon for Adam (to avoid divide by zero error): 1e-8.

Re-trained Model: In order to address the domain shift that arose from our model being trained on NCTE data but used for inference on TeachFX transcripts, we re-trained the model (after the pilot week of the study); that is, we fine-tuned our model on both TeachFX and the aforementioned NCTE data. We describe our method of re-training below:

We collected 22 transcripts from the pilot week of our study and extracted all teacher utterances (where, for TeachFX transcripts, we considered one utterance as 3 consecutive sentences of teacher talk, computed using a sliding window) from these transcripts. Using our original model, we obtained 444 teacher utterances that were predicted as focusing questions and approximately 700 which were not. We asked an annotator to label which of the predicted 444 focusing questions were indeed focusing questions. From this, we got 276 labeled focusing questions and 168 labeled non-focusing questions. In order to ensure that we collected a sufficient number of non-focusing questions (so that the fine-tuned model does not see too many positive examples while training), we manually verified and extracted 250 non-focusing questions of the 700 non-focusing question predictions. Thus, we obtained 694 teacher utterances (the aforementioned 444 + 250 teacher utterances) from the TeachFX pilot data. We added these utterances to our original NCTE dataset. From this dataset, we randomly sampled and chose 200 validation utterances for the held out set; we ensured that these were in a ratio similar to the distribution seen in an average TeachFX transcripts. We used the remaining utterances for re-training the model. For this training set, we oversampled the minority class such that we got approximately a 50:50 focusing question : non-focusing question split.

Appendix B Teacher Survey

Appendix B.1 Survey Email Text

“Thank you for using TeachFX to reflect on your teaching practice! To learn how to better support teachers and improve our app insights, we are sending you a survey about your background and the teaching practices used in your math/science lessons. This survey will take no more than 5 minutes to complete, and your identity will remain strictly confidential. To show our appreciation for your time, we will send you a $10 Starbucks gift card for completing the survey.”

Appendix B.2 Survey Questions

Appendix B.2.1 1. Thinking about your mathematics/science teaching, please indicate your opinion about each of the statements below:

(Scale: “not at all” to “all of the time”)

a) My questions elicit students’ mathematical/scientific thinking and reasoning.

b) My students talk about their mathematical/scientific ideas.

c) I pose open-ended questions.

d) I engage my class(es) in discussion.

e) I require students to explain their reasoning when giving an answer.

Appendix B.2.2 2. This set of questions seeks your opinion on how K-12 mathematics/science should be taught. Please indicate the extent to which you agree with the following statements:

(1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree)

a) It is valuable for students to see and hear other students' mathematical/scientific explanations.

b) Having students talk about their thinking helps other students make sense of mathematical/scientific ideas.

c) Teachers should ask students to explain how they got an answer.

d) Having a student clarify their thinking can surface mathematical/scientific reasoning in a way that helps other students learn.

e) Teachers should listen to all students' ideas, even if they are unusual or incorrect.

Appendix B.2.3 3. Which subject(s) do you teach?

(Select all that apply)

a) Mathematics

b) Science

c) Other(s): (Please specify)

Appendix B.2.4 4. At which grade level(s) do you teach mathematics/science this year?

(Select all that apply)

a) Kindergarten

b) 1st grade

c) 2nd grade

d) 3rd grade

e) 4th grade

f) 5th grade

g) 6th grade

h) 7th grade

i) 8th grade

j) 9th grade

k) 10th grade

l) 11th grade

m) 12th grade

Appendix B.2.5 5. Please indicate your role in the school:

a) I am a regular classroom teacher and teach a general/mixed population of students.

b) I have my own classroom, but exclusively teach a specific population of students (e.g., special education, emergent bilingual students)

c) I do not have my own classroom, but instead offer small-group instruction, tutoring or other kinds of special assistance to students from other teachers' classrooms.

Appendix B.2.6 6. How many total years of experience do you have teaching mathematics?

(Select one)

a) less than 1 year

b) 1-2 years

c) 3-4 years

d) 5-6 years

e) 6-7 years

f) over 8 years

Appendix B.2.7 7. Which of the following best describes your race-ethnicity?

(Select all that apply)

a) Asian

b) Native Hawaiian or Other Pacific Islander or Pacific Islander

c) Black or African American

d) Hispanic or Latinx

e) Native American, Alaska Native, or Indigenous

f) White or Caucasian

g) Multiracial or Biracial

h) Other race-ethnicity:

Appendix B.2.8 8. With which of the following do you identify?

(Select one)

a) Male

b) Female

c) Nonbinary

d) Prefer to self-identify:

e) Prefer not to say

Appendix C Table 2 Without Demographic Controls

	(1) **Focusing rate**	(2) **Uptake rate**	(3) **Student reasoning rate**	(4) **Student talk percentage**
Treatment	4.517*	0.284	0.514	0.004
	(1.805)	(0.563)	(0.491)	(0.013)
Control Mean	22.565	3.772	2.896	0.160
R2	0.316	0.276	0.190	0.169
Observations	533	533	533	533

Table C1: Standard errors are in parentheses. * p<0.05. These models estimate the effect of the automated feedback on focusing questions (treatment) on teachers' discourse features. These models exclude demographic control features obtained from the survey. We still observe a significant impact on focusing questions, but not the other discourse features.

Appendix D Other TeachFX Insights

Teachers could see a range of insights about pedagogical moves and classroom observations on the TeachFX app during the study. These are listed below:

1. Word Clouds for teacher talk and student talk
2. Talk ratios (percentage of teacher talk, student talk, group talk and silence)
3. Short Student Responses
4. Long Student Contributions (talk stretches where at least one student spoke for at least 7 seconds)
5. Student Questions
6. Teacher Talk Stretches
7. Volleyball Prompts (teacher ‘passing the ball’ back to students)
8. Teacher Questions
9. Open Ended Questions

   ↪ Although this insight may seem similar to focusing questions, it analyzes talk dynamics rather than the content (language) of the teacher utterance. Specifically, the insight shows every teacher question that is closely followed by a long student talk, while allowing for some silence and short teacher talk in between, in the case the teacher calls on a student. The start time is the start of the teacher question, and the end time is the end of the long student talk. This insight ignores any teacher question that contains ”can __ read” with any number of words in the blank.

10. Ping Pong Questions (teacher playing “ping pong” with students)

11. Think Time After Teacher Spoke (Wait Time 1)

12. Think Time after Student Spoke (Wait Time 2)