M-Powering Teachers: Natural Language Processing Powered Feedback Improves 1:1 Instruction and Student Outcomes

Dorottya Demszky
Stanford University
ddemszky@stanford.edu

Jing Liu
University of Maryland
jliu28@umd.edu

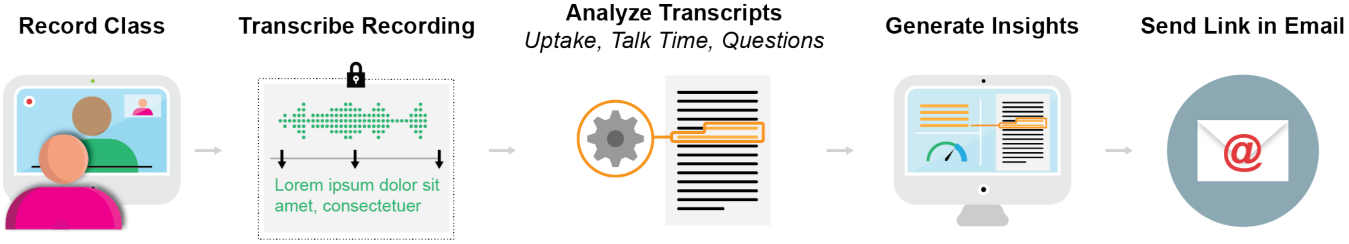


Figure 1: Pipeline for the automated teacher feedback.

ABSTRACT

Although learners are being connected 1:1 with instructors at an increasing scale, most of these instructors do not receive effective, consistent feedback to help them improve. We deployed M-Powering Teachers, an automated tool based on natural language processing to give instructors feedback on dialogic instructional practices —including their uptake of student contributions, talk time and questioning practices — in a 1:1 online learning context. We conducted a randomized controlled trial on Polygence, a research mentorship platform for high schoolers (n=414 mentors) to evaluate the effectiveness of the feedback tool. We find that the intervention improved mentors' uptake of student contributions by 10%, reduced their talk time by 5% and improved student's experience with the program as well as their relative optimism about their academic future. These results corroborate existing evidence that scalable and low-cost automated feedback can improve instruction and learning in online educational contexts.

CCS CONCEPTS

• **Applied computing** → **Computer-assisted instruction**.

KEYWORDS

natural language processing, automated teacher feedback, randomized controlled trial

ACM Reference Format:
Dorottya Demszky and Jing Liu. 2023. M-Powering Teachers: Natural Language Processing Powered Feedback Improves 1:1 Instruction and Student Outcomes. In *Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S '23), July 20–22, 2023, Copenhagen, Denmark.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3573051.3593379

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

L@S '23, July 20–22, 2023, Copenhagen, Denmark
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0025-5/23/07…$15.00
https://doi.org/10.1145/3573051.3593379

1 INTRODUCTION

Online 1:1 instruction is growing at an unprecedented speed, due to the well-documented effectiveness of 1:1 tutoring (Blooms' 2 sigma problem [3]), technological advances, funding resources, and demand for online educational opportunities generated by the Covid-19 pandemic[1]. Instructors on these platforms come from a wide range of backgrounds and are usually novices. However, apart from minimal training, most instructors do not receive consistent feedback on their instruction to help them improve. Studies in K-12 settings have consistently found that providing teachers with formative — nonevaluative, supportive, timely and specific — feedback through human-based classroom observations can improve both their instruction effectiveness [19, 24] and their students' outcomes [25, 28]. However, due to the large amount of resources required for conventional classroom observations, such methods for providing feedback cannot be easily scaled to these growing number of 1:1 teaching contexts.

Recent work has shown that consistent, automated feedback can improve instruction effectively and at scale in small group contexts. For example, Demszky et al. [7] introduced an automated tool powered by natural language processing (NLP), henceforth referred to as M-Powering Teachers. M-Powering Teachers provides 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[5, 21]. Their randomized study in Code in Place, an online programming course with a 1:10 instructor to student ratio showed that M-Powering Teachers can improve instructors' practice, including their uptake of student ideas and questioning, as well as students' satisfaction with the course. Building on Demszky et al. [7], we study the extent

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

to which the effectiveness of M-Powering Teachers translates to a 1:1 instructional setting.

We conducted a randomized controlled trial (n=414 instructors) on Polygence, a U.S.-based platform that pairs high schoolers with research mentors. Polygence complements Code in Place in several ways that help illuminate the generalizability of M-Powering Teachers. First, 1:1 instruction has different dynamics than small group instruction, allowing for more personalized instruction and the development of a closer personal relationship between the instructor and the student. These features may impact the effectiveness of the automated feedback in unforeseen ways. Second, unlike the five weekly sections of Code in Place, Polygence offers 10 sessions that spread across several months, which allows us to better understand how the impact of the feedback evolves over a longer period of time and with a bigger time gap between sessions. Third, the instructor and student population also differs from Code in Place: while the majority of Code in Place instructors are working professionals and the students are all adults, Polygence's mentors primarily consist of graduate students and the students are all high schoolers. Fourth, while Code in Place sections all follow pre-defined topics, Polygence's projects are chosen by the student and cover a diverse range of topics, such as neuroscience and art history, so we can see if the feedback applies to a diverse, student-driven curriculum.

We delivered automated feedback via M-Powering Teachers to a randomly chosen half of mentors, which we displayed on the Polygence platform within a day after each mentoring session. The feedback includes personalized insights on mentors' uptake of student contributions, talk time, actionable advice for eliciting and building on student ideas and reflection opportunities. We sent a reminder email to mentors once their feedback is released. This study addresses the following research questions:

(1) What percentage of mentors engage with the automated feedback?
(2) What is the impact of automated feedback on mentors' instruction?
(3) Does the automated feedback have a differential impact on different groups of mentors?
(4) What is the impact of automated feedback on project outcomes?

We use regression analyses while controlling for covariates and baseline, pre-intervention practices to answer these research questions. Our results indicate that 84% of treated mentors opened the feedback page at least once during their projects. The feedback improved treated mentors' uptake of student contributions by 9%, the number of times they ask questions and repeat students' substantive words by 6% and reduce mentors' talk time by 5%, compared to the control group. Heterogeneity analysis shows that the impact of the feedback is generally similar across observed demographic groups and over time, with a few exceptions. Specifically, female and non-STEM mentors were more likely to decrease their talk time as a result of the feedback, and STEM mentors were more likely to increase their uptake as a result of the feedback. We also find evidence that treated mentors' students enjoyed their Polygence projects 4% more and reported a 5% greater increase in optimism about their academic future.

The paper is structured as follows. Section 2 provides an overview of related work on automated tools for analyzing and providing feedback to teachers on their instruction. Section 3 introduces the background for the study, describing the statistics of the participant population and analytical sample. Section 4 describes M-Powering Teachers, our feedback tool. In Section 5, we introduce the setup for the randomized controlled trial and the analyses we used to answer each of the four research questions. In Section 6 we provide results for the research questions. Finally, in Section 7 we discuss implications and limitations of our study and provide a window into future work.

2 RELATED WORK

In the education literature, many scholars agree that providing formative feedback is critical for both learners and instructors [13]. With recent technological advances, there has been a growing number of efforts aimed at building 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 [19].

The majority of automated tools provide teachers with analytics on student engagement and progress that allow teachers to monitor student learning and intervene when needed [1, 2, among others]. For example, tools are designed to facilitate teachers to monitor student concentration on lesson activities [26], discover critical moments in group learning [23], and make decisions on what exercises to assign to students [4]. However, few tools provide teachers with feedback that can serve as a vehicle for self-reflection and instructional improvement based on teachers' own practice. 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 key aspects of classroom discourse [9, 10, 17, 18, 22]. For example, Kelly et al. [18] 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. [27] 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.

While new methods and tools for automated teacher feedback are emerging, the field still lacks data and rigorous evaluation on whether such tools indeed improve teaching and student outcomes. For the limited number of tools for which such studies exist, the results vary. Jacobs et al. [16] found that the TalkMoves application was perceived positively by K-12 math teachers. The authors observed a positive but not significant trend for the impact of the feedback on teacher practice; lack of significance is potentially due to its small sample size (n=21). The current study builds on prior research that showed positive impacts of the M-Powering Teachers tool on instructional practice [7], and fills a gap by being — to our knowledge — the first randomized controlled trial to test the impact of automated discourse-based teacher feedback in 1:1 instruction settings.

3 BACKGROUND

We conducted the study on Polygence[2], an online marketplace for project-based learning based in the U.S.. The platform pairs high school students with research mentors — most of whom are graduate students — based on high schoolers' interests and mentors' expertise. Mentors' responsibilities are providing guidance to high school students in each step of the research process: problem formulation, literature review, identifying methods, conducting analyses and showcasing their work. Mentors typically meet with students online through Zoom ten times for hour-long sessions over the course of 3-4 months; the timeline is decided by the mentors and students mutually.

The study ran between May, 2021 and September, 2022. Based on a priori power analysis using the Code in Place results, we would need four hundred mentors to achieve the minimal detectable effect size. Thus, we ended the recruitment of new mentors when we reached that targeted sample size. Participation in our study required mentors to consent to their sessions being recorded and their de-identified data to be analyzed by researchers.[3]

The analytic sample comprises of 622 completed projects and 5,037 sessions, representing 414 mentors and 624 students. Table 1 summarizes the demographics of mentors and students based on data collected by Polygence. Nearly all mentors (99%) are based in the U.S. and so are most students (84%). More than half of mentors (53%) are female, whereas about a third (34%) of the students are female. As the primary responsibility of the mentors is guiding research projects, it is not surprising that 99% of mentors have a college degree, 40% have a master's degree and 16% have obtained a PhD. The majority of mentors have expertise in STEM areas (85%), with the top 5 subjects being biology (43%), computer science (24%), neuroscience (20%) and psychology (18%). Unfortunately, we do not have data on mentors' race/ethnicity. As for students, 46% identified themselves as Asian, 11% as Caucasian, 2% as Hispanic, 1% as Black and 1% as Native American; the rest of students declined to report their race/ethnicity.

Table 1: Demographics of the Analytical Sample

Mentors		Students	
Num. Mentors	414	Num Students	624
In U.S.	99%	In U.S.	84%
In Europe	1%	In Asia	14%
Female	53%	In Europe	1%
College degree	99%	Female	34%
Masters degree	40%	*Race/Ethnicity*	
PhD degree	16%	Asian	46%
STEM	85%	Caucasian	11%
Humanities	44%	Hispanic	2%
Top 5 Subjects		Black	1%
Biology	43%	Native Am.	1%
Comp. Sci.	24%	Other	2%
Neuroscience	20%		
Social Science	19%		
Psychology	18%		

Notes: Subject, race and ethnicity categories are not mutually exclusive. Percentages may not add up to 100% due to missing values

4 THE M-POWERING TEACHERS TOOL

We adapt Demszky et al. [7]'s M-Powering Teachers tool to Polygence's infrastructure and domain. As Figure 1 shows, the workflow for generating feedback involves (1) recording sessions, (2) transcribing the recordings, (3) conducting NLP analyses and (4) displaying the results on a web page to mentors. Polygence uses Zoom to record classes, a service that provides automatically generated transcripts via Otter.ai. Below, we provide a brief overview of the back-end analyses and the feedback page. For more details, please refer to Demszky et al. [7].

4.1 Transcript Analysis

Similarly to [7], we algorithmically analyzed the transcripts to identify various discourse-related phenomena. Our primary focus is the **teachers' uptake of student contributions** measure, which was shown to be positively correlated with a range of important educational outcomes using secondary data [6, 9] and to improve practice via M-Powering Teachers [7]. We use [9]'s uptake model off-the-shelf, which was pre-trained on large corpora of educational interactions [6] and Switchboard [12] via self-supervision. Specifically, the model 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 student's 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's utterance. We use 0.8 as a cutoff for distinguishing high vs. low uptake [7].

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 proportion** using timestamps from the transcripts, dividing teachers' talk time by the sum of student and teacher talk time. 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 students' words** using Demszky et al. [8]'s method, who found repetition to be a core component of uptake. The repetition measure computes the percentage of student words that are repeated by the teacher in their subsequent utterances,

[2] www.polygence.org

[3] The study was approved by the Stanford IRB (#60435).

ignoring stopwords and punctuation. See [7] for more details on these measures and their correlation with the uptake measure.

4.2 User Interface

Our primary objective is to empower teachers by encouraging them to reflect on their practice and to draw their own interpretations from the statistics we provide. M-Powering Teachers was designed for this purpose, with positive and non-judgmental language, specific examples and reflection opportunities. We used the same interface as Demszky et al. [7], with minor adaptations to match Polygence's use case (e.g. changing "students" to "student" given the 1:1 setting).

Figure 2 shows the components of the one-page feedback application. Mentor name is masked with grey for anonymity. On the top of the page, a brief paragraph introduces the feedback to mentors, emphasizing that the feedback is private and its goal is to support their professional development. Then, mentors can see statistics about talk time and examples from their transcript when their questions elicited a long student utterance in the just finished session. Below the examples, mentors can see the number of uptakes they demonstrated (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 mentors' uptake examples that co-occur with them asking a question. We also provided an input box for users to reflect on these examples and plan for the next session. At the bottom of the page, we shared resources, including blog posts and papers on dialogic instructional practices. Finally, we provided the entire transcript to mentors for review.

5 METHODS

We describe the setup for the randomized controlled trial, as well as analytical methods we use to answer each of the four research questions.

5.1 Randomized Controlled Trial

We conducted a randomized controlled trial to evaluate the effectiveness of the M-Powering Teachers tool for Polygence mentors. Upon joining Polygence, we assigned each participant either to the treatment (n=192) or the control group (n=222) using a random number generator. Only treatment group mentors were able to access the feedback through the Polygence website. We also sent an email to each mentor in the treatment group when their feedback was ready, usually a day after their session. We designed the email to be generic, with a sentence "Below please find your AI-powered session feedback" and a button that leads to the feedback page (Figure 5), with the goal of ensuring any effect of the feedback is through mentors' direct interaction with the feedback page rather than the email.

To verify whether our randomization was successful, we evaluate whether the demographics of mentors in the treatment and control groups differ statistically. We also compare mentors' discourse features measured in their first Polygence session, prior to receiving any automated feedback. As Table 4 shows, 26 out of the 27 covariates do not show any statistically significant differences at the 5% level between the treatment and control groups. There is a statistically significant difference on mentors' uptake in the first session; interestingly, treated mentors' pre-intervention uptake is lower than those in the control group. We also conducted a joint significance test that considers all these baseline variables. The resulted F statistic is only 0.89, failing to reject balance between the two conditions. This analysis validates the success of our randomization and suggests that any differences we observe later in the course are likely due to the effects of the intervention.

5.2 RQ1: What percentage of mentors engage with the automated feedback?

We measure engagement with the feedback by considering whether a mentor opened the feedback page, either by clicking the link in the email or accessing it directly via the Polygence platform.[4] We consider the percentage of mentors that engaged with the feedback at least once over the course of their projects, and also conduct a per-session analysis to study engagement with the feedback over time.

5.3 RQ2: What is the impact of automated feedback on mentors' instruction?

We use measures generated using the NLP methods described in Section 4.1 as dependent variables for measuring changes in mentor's instructional practices. Concretely, we measure the number of times they took up student ideas per hour, the number of questions they raised per hour, the number of times they repeated student words per hour and proportion of mentor talk. We use hourly measures (rates) instead of raw counts to account for differences in section duration.

We run separate linear regression models to estimate the effect of the treatment on each dependent variable. The models are specified as below:

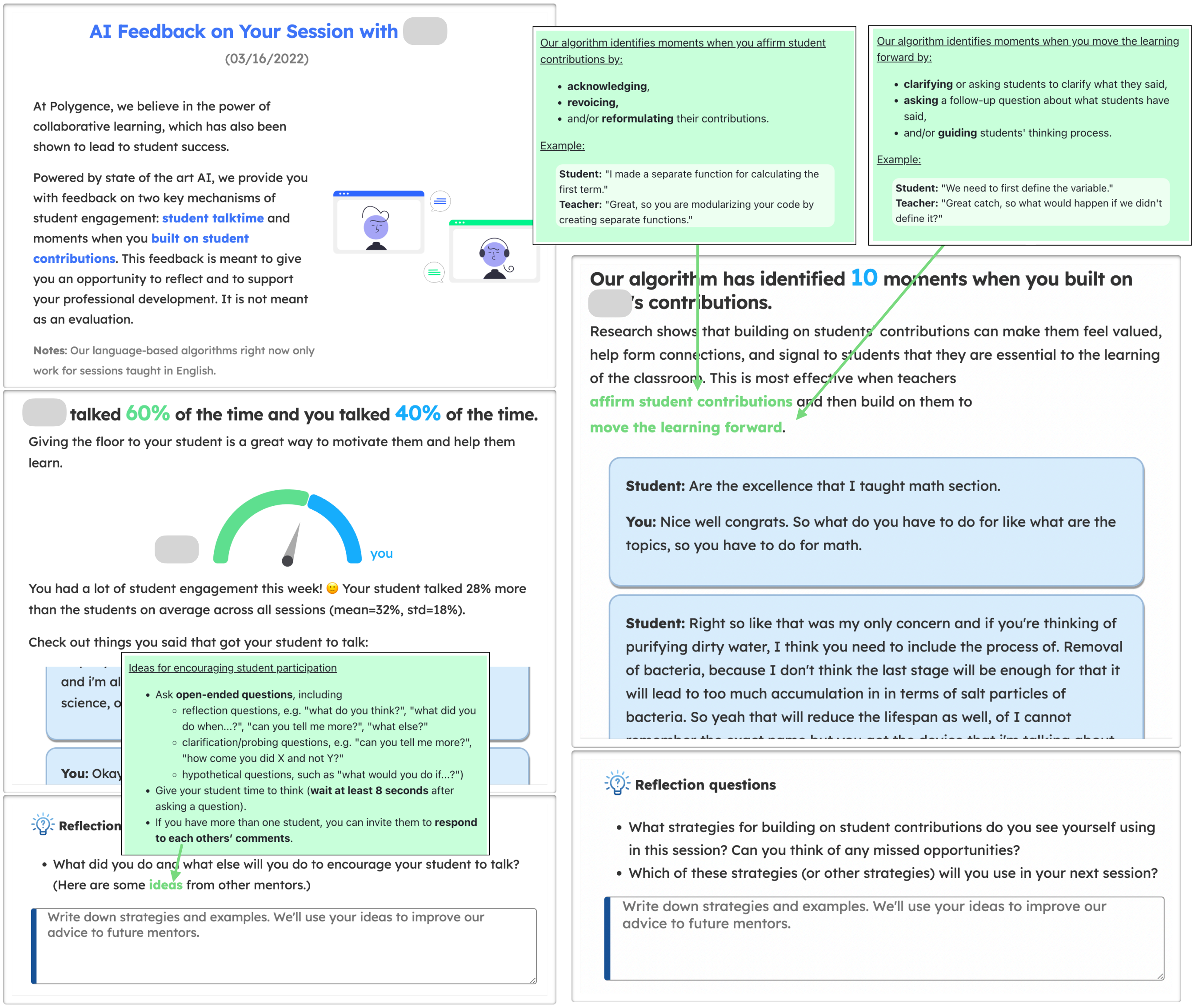
$$y_d = g\beta_1 + M\beta_2 + S\beta_3 + T\beta_4 + \varepsilon \quad (1)$$

where y refers to a dependent variable d; g is a binary variable that indicates the treatment status; M is a vector of mentor covariates; S is a vector of student covariates, T is a vector of transcript metadata; $\beta_1, \beta_2, \beta_3, \beta_4$ are parameters to be estimated and ϵ indicates the residuals. We conduct analyses at the transcript-level and cluster standard errors at the mentor level to account for repeated observations within a mentor.

We use the following binary variables as mentor covariates M across all models: self-identifies as female, has a college degree, has a master's degree, has a PhD, has coding skills, is located in Africa/America/Asia/Europe, and has mentoring experience in STEM subjects. We also include mentors' baseline discourse features in their first session as covariates: the number of times mentors took up student ideas per hour, the number of mentor questions per hour, the number of times mentor repeated student words per hour and the proportion of mentor talk per hour in their first, pre-feedback sessions.

As for student covariates S, we use the following binary variables across all models: self-identifies as female, was in the same timezone

[4] Unfortunately, we do not have data on how much time mentors spent on average on the feedback page which prevents us from analyzing to what extent mentors engaged with the feedback.



AI Feedback on Your Session with [redacted]

(03/16/2022)

At Polygence, 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: Our language-based algorithms right now only work for sessions taught in English.

[redacted] **talked 60% of the time and you talked 40% of the time.**

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

You had a lot of student engagement this week! 🙂 Your student talked 28% more than the students on average across all sessions (mean=32%, std=18%).

Check out things you said that got your student to talk:

and i'm al
science, o

You: Okay

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 explain 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

- What did you do and what else will you do to encourage your student to talk? (Here are some ideas from other mentors.)

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

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 10 moments when you built on [redacted]**'s 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**.

Student: Are the excellence that I taught math section.

You: Nice well congrats. So what do you have to do for like what are the topics, so you have to do for math.

Student: Right so like that was my only concern and if you're thinking of purifying dirty water, I think you need to include the process of. Removal of bacteria, because I don't think the last stage will be enough for that it will lead to too much accumulation in in terms of salt particles of bacteria. So yeah that will reduce the lifespan as well, of I cannot

Reflection questions

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

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

Figure 2: The user interface of the automated feedback on the Polygence platform.

as mentor, was located in Africa/America/Asia/Europe, and race or ethnicity (Asian, Black, Caucasian, Hispanic, Native American, and Other). We also include two variables as transcript covariates T to capture mentors' experiences with Polygence: the session number within the project (2-10) and the session number for the given mentor considering all of their Polygence projects. These numbers may be the same if the mentor only participated in one project throughout the study, but the latter can be higher if the mentor participated in multiple projects.

To verify that our estimates are not significantly influenced by the choice of demographic control variables, we also estimate a version of our models without them.

5.4 RQ3: Does the automated feedback have a differential impact on different groups of mentors?

Mentors from different backgrounds or with different characteristics may respond differently to the feedback. We conduct a heterogeneity analysis by gender, having obtained a PhD, mentoring in STEM, and whether they demonstrated below vs above average baseline uptake in the first session. These analyses use the formula in Equation 1, separately estimating coefficients for sub-populations (e.g. female vs non-female, STEM vs non-STEM mentors).

5.5 RQ4: What is the impact of automated feedback on project outcomes?

To address this question, we consider all available outcome variables from Polygence, including self-reported survey items and the publication status of students' work. We consider mentor and student survey items that are rated on a scale. The main item on these surveys is the *Net Promoter Score (NPS)*. We consider the NPS to be an estimate of mentors and students' overall experience with the project. For mentors, the question asks "On a scale from 0-10, how likely are you to recommend mentoring with Polygence to your friends / colleagues?". For students, the question asks "On a scale from 0-10 how likely is it that you would recommend Polygence to a friend?".

In addition to the NPS ratings, we also consider students' *Mentor Review Score*, on a 5-star scale ("Please leave a review of your mentor for future students who are matched with them,") and students' relative Optimism About their Academic Future as induced by Polygence, rated on a 0-10 scale ("Please rank how strongly you agree with this statement: 'My Polygence experience has made me feel more excited about my academic future' 0 = Strongly Disagree, 10 = Strongly Agree"). Appendix B lists all survey questions rated on a scale, which include the questions above and one additional question asking the mentor/student to rate their match. We exclude the match ratings in this analysis given space constraints.

Besides surveys to mentors and students, Polygence also tracks whether a project leads to a journal publication or a conference presentation based on information provided by the mentors and the students. We thus consider if a student's work was *accepted to a journal or a conference* by Dec 31, 2022 as an additional outcome. This measure has the benefits of capturing the success of the tutoring sessions in a more objective manner, but also has a couple of caveats. First, Polygence may not have publication information for all projects because such information relies on mentors and students providing that information. Another limitation is that peer reviews take time, and there may be undocumented projects that are published or presented beyond the deadline we set.

We use the formula in Equation 1 for analyzing the impact of the feedback on project outcomes. The main difference from RQ1 is that we conduct analyses at the project level, rather than the transcript level, as we only observe these outcomes at the end of each project. Given that T is at the transcript-level, we exclude those variables from project-level analyses.

6 RESULTS

We summarize results for each research question.

6.1 RQ1: What percentage of mentors engaged with the automated feedback?

We found that 84% of treated mentors opened the feedback page at least once during their projects, either by clicking on the link in the email or accessing it directly via the Polygence platform. This engagement started out high at the first session, when 74% of mentors checked the feedback, and it decreased over the course of the project, plateauing around 30%: session 2 (63%), session 3 (50%), session 4 (45%), session 5 (38%), session 6 (33%), session 7 (34%), session 8 (29%), and session 9 (31%).

Table 2: Impact of Treatment on Teaching Practices

	(1) Uptake	(2) Questions	(3) Repetitions	(4) Talk Ratio
Treatment	0.565*	1.043+	2.284*	-0.035**
	(0.250)	(0.618)	(1.075)	(0.011)
Control Mean	5.969	17.906	39.409	0.722
R^2	0.096	0.163	0.209	0.167
Observations	5037	5037	5037	5037

Notes: Standard errors in parentheses. + p<0.10 * p<0.05 ** p<0.01. 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 covariates for mentor and student demographics, session id and pre-intervention teaching practices — see Section 5.3.

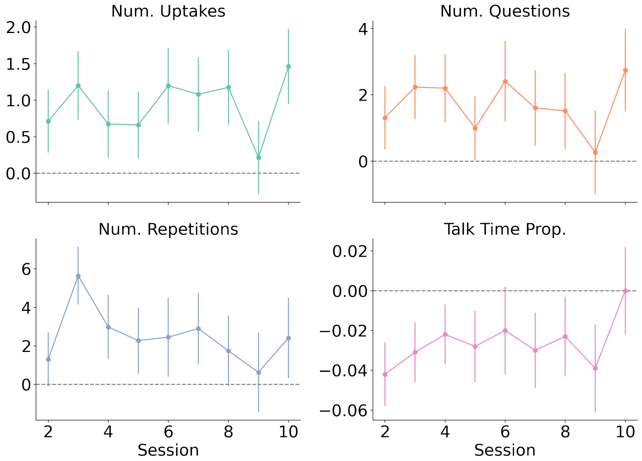


Figure 3: Impact of the Treatment On Teaching Practices Over Time

6.2 RQ2: What is the impact of automated feedback on mentors' instruction?

Table 2 shows that our intervention improved teaching practices across all discourse measures. On average, treated instructors took up student contributions 9% ($p < 0.05$) more times compared to the control group. Treated mentors also asked 6% more questions ($p < 0.1$) and repeated substantive words in student utterances in 6% more instances ($p < 0.05$). In contrast, treated mentors decreased their talk time to 69%, 5% less than control group mentors ($p < 0.01$). All the coefficients stay roughly the same size and significance levels when we exclude demographic controls from the models, confirming that the results are robust to the inclusion of covariates (Table 5).

To understand how the impact of the treatment evolved over time, we ran separate analyses for the first 10 sessions for each mentor (the average duration of a single project). Figure 3 shows the coefficients over time, with the horizontal line at zero indicates no effect. The coefficients suggest that while there is some variation over time but the positive impact of the feedback is consistent

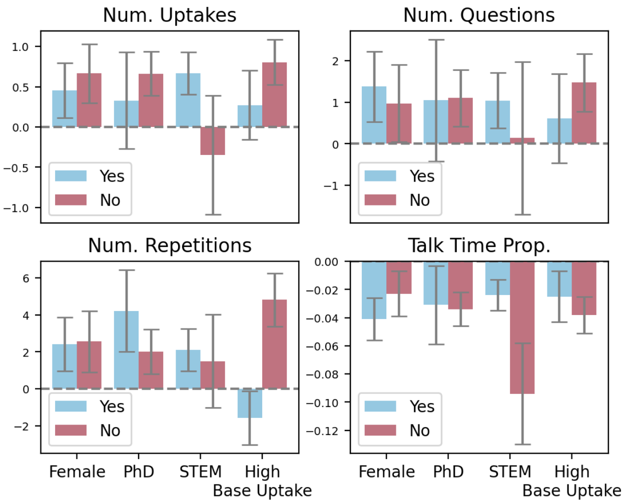


Figure 4: Heterogeneity Analysis by Mentor Demographics

across the span of the experiment. Note that while the coefficients on teachers' talk time are all negative, they carry a positive meaning as it suggests that the intervention increases students' talk time. Interestingly, the treatment effects seem to vary more in the middle of the intervention than in the beginning and at the end. For example, the impact of the feedback on uptake, questioning, and repetition all increases in the third session compared to the second session, while we also see a dip of the impact in the ninth session and a rebound in the last session. In contrast, the pattern of the impact of the feedback on talk time is different from others: it is highest for session 2 and 9, and the lowest in the last session when it is exactly zero.

6.3 RQ3: Does the automated feedback have a differential impact on different groups of mentors?

As shown in Figure 4, we find that the feedback decreased female mentors' talk time almost twice as much as that of their counterparts, but that there is no significant difference by gender for the other features. As a result of the feedback, mentors with a PhD repeated students' words twice as many times as non-PhD mentors. The difference for other features is not statistically significant by PhD status. The treatment decreased non-STEM mentors' talk time three times as much than that of STEM mentors, but that STEM mentors increased their uptake more than non-STEM mentors. There are no other observable differences between STEM and non-STEM mentors due to large standard errors, explained by the relatively small percentage of non-STEM mentors. Lastly, we find that generally the feedback had a stronger impact on teaching practices for mentors who demonstrated below average baseline uptake compared to their counterparts.

6.4 RQ4: What is the impact of automated feedback on project outcomes?

As Table 3 shows, we find evidence that the treatment improved students' NPS scores by 4% ($p < 0.05$) and that it improved students' relative optimism about their academic future by 5% ($p < 0.05$). The treatment also increased mentors' NPS ratings by 3% with marginal significance ($p < 0.1$). The treatment did not have an impact on students' mentor review scores nor on their publication status. These estimates are robust to the inclusion of covariates as well (Table 6).

7 DISCUSSION & FUTURE WORK

We set out to test the extent to which M-Powering Teachers can be effective in a 1:1 instructional context. In this section, we compare our findings to prior work on Code in Place, review implications and limitations, and highlight the avenues for future work.

7.1 Comparison with Code in Place

Our findings corroborate prior results from Code in Place [7] by showing that M-Powering Teachers can improve instructional practice and student outcomes. In both contexts, the feedback improved instructors' uptake of student contributions. Our analysis shows that in both contexts, the impact of the feedback generally increased as mentors progressed in their projects and peaked in the third session. Such positive effects then sustained in subsequent sessions throughout the intervention. This temporal pattern indicates that it takes a relatively short time for treated instructors to change their practice based on the feedback. Both studies also provide evidence that the feedback has a positive downstream impact on students' satisfaction with their learning experience. Finally, while we observe some heterogeneity in treatment effects based on instructor demographics and baseline characteristics, the effects are generally positive across different subgroups. One consistent finding across both teaching contexts is that M-Powering Teachers has a similar impact on instructors regardless of their gender, except that female instructors decrease their talk time more as a result of the feedback.

There are also interesting differences between the results from the two studies, likely explained by differences in the number of students in a session and the content of instruction, along with several other differences between the two platforms. In the Code in Place study, the feedback only reduced instructors' talk time in the third week but not the other weeks, whereas in Polygence the feedback decreased mentors' talk time across all sessions except the last. In the Polygence context, the feedback increased repetition, whereas in Code in Place we do not observe an effect on repetition. In contrast, treated instructors in Code in Place increased their use of questioning significantly, while in Polygence, treated mentors did so only marginally and relied more on repeating students' contributions as a strategy to improve their uptake. Finally, while in the Code in Place context, the feedback has a slightly greater impact on instructors with below average baseline uptake, the opposite is true for Polygence instructors, where those with below average baseline uptake were more likely to improve all of their practices as a result of the feedback. However, the coefficients in Code in Place are noisier, and only marginally significant, whereas for Polygence the differential positive impact on those with low baseline uptake

Table 3: Effect of Treatment on Project Outcomes

	(1) Mentor NPS	(2) Student NPS	(3) Student Mentor Review Score	(4) Student Optimism About Acad. Future	(5) Published Work
Treatment	0.230+	0.310*	0.020	0.391*	0.013
	(0.124)	(0.129)	(0.028)	(0.152)	(0.025)
Control Mean	9.144	8.093	4.871	8.155	0.107
R2	0.075	0.066	0.088	0.087	0.039
Observations	558	503	557	407	622

Notes: Standard errors in parentheses. + $p<0.10$ * $p<0.05$ ** $p<0.01$. 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 covariates for mentor and student demographics and pre-intervention teaching practices — see Section 5.3.

is more robust across all discourse features. The Polygence results indicate that instructors who have more room for improvement are more likely to benefit from feedback.

The differences we observe between the two learning contexts pose both open questions about why such differences exist and opportunities for future research. It is impossible to fully pinpoint the underlying reasons for the differences based on the data we currently have. Therefore, as a next step, we will conduct qualitative interviews with participants of our studies to decode the mechanisms behind M-Powering Teachers in different learning contexts. Once we accumulate more knowledge on how the tool works in a particular setting, we could also imagine adapting our tool based on the specific patterns of effects we observe. For example, we could strengthen the feedback we provide to mentors in Polygence during week 9 to prevent the "dip" in impact.

7.2 Addressing Limitations

Our study still leaves some questions open with respect to the generalizability of M-Powering Teachers to other 1:1 instructional contexts. Most Polygence mentors are graduate students, unlike tutors in other virtual settings who tend to be college students or retired professionals doing tutoring for a living. The Polygence student population is also not representative of the general U.S. student population, as most of them come from upper/upper middle class backgrounds, have college educated parents, and have access to many educational opportunities. That being said, there is some evidence suggesting that compared with their peers, the contribution of students from low-income backgrounds, underrepresented minorities, and English Language Learners, is less likely to be taken up in classrooms [11, 14, 20]. Therefore, it is possible that M-Powering Teachers may be more beneficial for students from disadvantaged backgrounds. It is our high priority to continue testing the effectiveness of this tool across a diverse range of instructional contexts and various teacher and student populations in a systematic manner. Our ultimate goal is to ensure that the tool can improve learning opportunities for marginalized student populations and enhance educational equity.

Although the results on student outcomes are promising, a comprehensive evaluation of the feedback's impact requires more reliable measures of student learning and growth. Similar to many informal learning platforms, Polygence does not systematically collect data on student learning outcomes, such as assignment completion, grades, or test scores. In this study, we primarily rely on students' self-reported satisfaction and the publication status of students' work as proxies for a project's success. As discussed in Section 5.5, both measures have limitations and might not accurately capture students' learning. Future work may evaluate best practices for collecting learning outcome measures on informal learning platforms.

There is also plenty of room for improving the tool, starting with addressing speech transcription errors raised by Demszky et al. [7]. Despite the fact that 1:1 online contexts provide much clearer audio than noisy in-person classrooms, the transcription quality is still far from perfect, especially for speakers of non-Standard American English. It is essential to address this issue before scaling the use of the tool to ensure that it does not propagate inequities in teacher professional development. The tool can also be further enhanced by making it more interactive. Only a small percentage of instructors used the reflection boxes, so we need to come up with more engaging ways of encouraging reflection. For example, we could ask instructors to choose the type of feedback they receive, track their practice over time and provide rubric-based reviews for themselves. Although we are cautious about using generative models before ensuring that they are reliable and safe, recent advances that involve guardrails (OpenAI's ChatGPT[5]) suggest that we may soon have a trustworthy solution, and such models can allow us not only to measure but also to provide adaptive suggestions to teachers.

7.3 Conclusion

In sum, we provide new evidence for the effectiveness of automated feedback on instruction in a 1:1 teaching context. We show that it is possible to improve instruction and student outcomes through consistent, individualized, automated feedback, *at minimal to no cost.* Our tool is particularly promising for settings that lack teacher professional development opportunities, and might also be used in resourceful settings to complement human-based feedback mechanisms that are proven to be less effective [15]. Upon further

[5] https://openai.com/blog/chatgpt/

evaluation and improvement, M-Powering Teachers and similar tools could help scale professional development opportunities for instructors and ensure that scale does not come at the expense of high-quality instructional support. We invite others to join our effort to test, improve and build upon M-Powering Teachers, with the ultimate goal of providing equitable access to becoming and to being taught by expert, effective instructors.

ACKNOWLEDGMENTS

We thank Heather Hill, Dan Jurafsky, Greg Walton, Gábor Orosz, and members of the Stanford HAI-Education seminar group for helpful feedback. We thank members of the Polygence team, especially János Perczel, Staci Hill, Veronica Edwards, Claire Jung, Jin Chow, Ádám Gyulavári, Levente Bélteky for implementing the feedback tool, sharing insights and de-identified data for analysis. D.D. is grateful for funding from the Melvin and Joan Lane Stanford Graduate Fellowship. D.D. has a financial relationship with Polygence.

REFERENCES

[1] L. Alrajhi, A. Alamri, F. D. Pereira, and A. I. Cristea. 2021. Urgency Analysis of Learners' Comments: An Automated Intervention Priority Model for MOOC. In *International Conference on Intelligent Tutoring Systems*. 148–160.

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

[3] Benjamin S Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher* 13, 6 (1984), 4–16.

[4] Nathalie Bonneton-Botté, Sylvain Fleury, Nathalie Girard, Maëlys Le Magadou, Anthony Cherbonnier, Mickaël Renault, Eric Anquetil, and Eric Jamet. 2020. Can tablet apps support the learning of handwriting? An investigation of learning outcomes in kindergarten classroom. *Computers & Education* 151 (2020), 103831.

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

[6] Dorottya Demszky and Heather Hill. 2022. The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts. *arXiv preprint arXiv:2211.11772* (2022).

[7] Dorottya Demszky, Jing Liu, Heather C Hill, Dan Jurafsky, and Chris Piech. 2023. 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* (2023).

[8] D. Demszky, J. Liu, Z. Mancenido, J. Cohen, H. Hill, D. Jurafsky, and T. Hashimoto. 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)*.

[9] Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori B Hashimoto. 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)*. 1638–1653.

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

[11] Noel Enyedy, Laurie Rubel, Viviana Castellón, Shiuli Mukhopadhyay, Indigo Esmonde, and Walter Secada. 2008. Revoicing in a multilingual classroom. *Mathematical Thinking and Learning* 10, 2 (2008), 134–162.

[12] John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, Vol. 1. IEEE Computer Society, 517–520.

[13] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.

[14] Beth Herbel-Eisenmann and Niral Shah. 2019. Detecting and reducing bias in questioning patterns. *Mathematics Teaching in the Middle School* 24, 5 (2019), 282–289.

[15] Heather C Hill. 2009. Fixing teacher professional development. *Phi Delta Kappan* 90, 7 (2009), 470–476.

[16] Jennifer Jacobs, Karla Scornavacco, Charis Harty, Abhijit Suresh, Vivian Lai, and Tamara Sumner. 2022. Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education* 112 (2022), 103631.

[17] E. Jensen, M. Dale, P. J. Donnelly, C. Stone, S. Kelly, A. Godley, and S. K. D'Mello. 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.

[18] S. Kelly, A. M. Olney, P. Donnelly, M. Nystrand, and S. K. D'Mello. 2018. Automatically Measuring Question Authenticity in Real-World Classrooms. *Educational Researcher* 47 (2018), 7. https://doi.org/10.3102/0013189X18785613

[19] M. A. Kraft, D. Blazar, and D. Hogan. 2018. The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence. *Review of Educational Research* 88(4) (2018), 547–588. https://doi.org/10.3102/0034654318759268

[20] Soyong Lee and Seungho Moon. 2013. Teacher Reflection in Literacy Education–Borrowing from Bakhtin. *International Journal of Higher Education* 2, 4 (2013), 157–164.

[21] M. C. O'Connor and S. Michaels. 1993. Aligning Academic Task and Participation Status through Revoicing: Analysis of a Classroom Discourse Strategy. *Anthropology & Education Quarterly* (1993), 318–335.

[22] B. Samei, A. M. Olney, S. Kelly, M. Nystrand, S. D'Mello, N. Blanchard, X. Sun, M. Glaus, and A. Graesser. 2014. Domain Independent Assessment of Dialogic Properties of Classroom Discourse. (2014). https://eric.ed.gov/?id=ED566380

[23] Baruch B Schwarz, Naomi Prusak, Osama Swidan, Adva Livny, Kobi Gal, and Avi Segal. 2018. Orchestrating the emergence of conceptual learning: A case study in a geometry class. *International Journal of Computer-Supported Collaborative Learning* 13 (2018), 189–211.

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

[25] M. P. Steinberg and L. Sartain. 2015. Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching project. *Education Finance and Policy* 10, 4 (2015), 535–572.

[26] Yen-Ning Su, Chia-Cheng Hsu, Hsin-Chin Chen, Kuo-Kuang Huang, and Yueh-Min Huang. 2014. Developing a sensor-based learning concentration detection system. *Engineering Computations* 31, 2 (2014), 216–230.

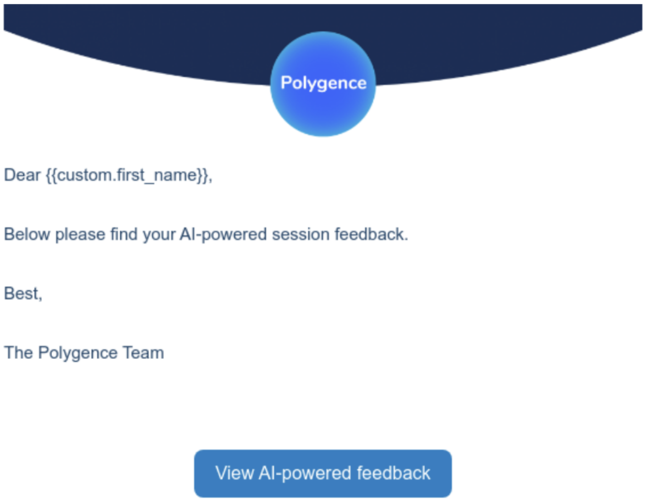
[27] A. Suresh, J. Jacobs, V. Lai, C. Tan, W. Ward, J. H. Martin, and T. Sumner. 2021. Using Transformers to Provide Teachers with Personalized Feedback on their Classroom Discourse: The TalkMoves Application. arXiv. (2021). arXiv:2105.07949 preprint.

[28] E. S. Taylor and J. H. Tyler. 2012. The effect of evaluation on teacher performance. *American Economic Review* 102, 7 (2012), 3628–51.

Table 4: Randomization Check

	Control Mean	Treatment Mean	P Value
Female	0.53	0.54	0.85
College degree	1	0.98	0.25
Masters degree	0.38	0.41	0.55
PhD degree	0.15	0.16	0.82
In America	0.99	0.99	0.39
In Europe	0.01	0.01	0.65
STEM	0.87	0.84	0.38
Humanities	0.43	0.47	0.41
Physics	0.06	0.1	0.13
Chemistry	0.09	0.08	0.81
Engineering	0.18	0.15	0.29
History	0.04	0.05	0.75
Social Science	0.15	0.23	0.05
Comp Sci	0.23	0.24	0.8
Business	0.12	0.13	0.79
Biology	0.45	0.39	0.19
Psychology	0.19	0.17	0.48
Medicine	0.16	0.1	0.09
Neuroscience	0.23	0.17	0.14
Literature	0.05	0.06	0.57
Mathematics	0.06	0.07	0.69
Arts	0.08	0.08	0.93
Languages	0.08	0.09	0.65
Uptake (Session 1)	8.95	7.54	0.01
Questions (Session 1)	41.79	40.71	0.51
Repetitions (Session 1)	21.07	18.88	0.06
Talk Ratio (Session 1)	0.73	0.71	0.11

A EMAIL



Dear {{custom.first_name}},

Below please find your AI-powered session feedback.

Best,

The Polygence Team

Figure 5

B SURVEY QUESTIONS

The final surveys had the following questions rated on a scale; we exclude open-ended questions here, as they could not be easily converted into numerical outcomes. We use the bolded items in our analyses.

B.1 Mentor Survey

(1) Now that you've completed the program with this student, how would you rate this match on a scale from 0-10?
(2) How likely are you to recommend mentoring with Polygence to your friends / colleagues on a scale from 0-10?

B.2 Student Survey

(1) On a scale from 0-10 how likely is it that you would recommend Polygence to a friend? 0=Not Likely, 10=Very Likely
(2) Please rank how strongly you agree with this statement: "My Polygence experience has made me feel more excited about my academic future" 0 = Strongly Disagree, 10 = Strongly Agree
(3) Now that you've completed the program with your mentor, how would you rate mentor match on a scale from 0-10?
(4) Please leave a review of your mentor for future students who are matched with them.
    - 1 Star
    - 2 Stars
    - 3 Stars
    - 4 Stars
    - 5 Stars

Table 5: Table 2 Without Controls

	(1)	(2)	(3)	(4)
	Uptake	**Questions**	**Repetitions**	**Talk Ratio**
Treatment	0.534*	1.034	2.270*	-0.034**
	(0.257)	(0.645)	(1.107)	(0.011)
Control Mean	5.969	17.906	39.409	0.722
R^2	0.075	0.141	0.178	0.147
Observations	5037	5037	5037	5037

Notes: Standard errors in parentheses. + $p<0.10$ * $p<0.05$ ** $p<0.01$. 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). We exclude demographic covariates and only include controls for session id and pre-intervention teaching practices — see Section 5.3.

Received 3 February 2023

Table 6: Table 3 Without Controls

	(1) **Mentor NPS**	(2) **Student NPS**	(3) **Student Mentor Review Score**	(4) **Student Optimism About Acad. Future**	(5) **Published Work**
Treatment	0.226+	0.344**	0.029	0.379*	0.011
	(0.122)	(0.128)	(0.028)	(0.152)	(0.024)
Control Mean	9.144	8.093	4.871	8.155	0.107
R2	0.009	0.013	0.002	0.015	0.000
Observations	558	503	557	407	622

Notes: Standard errors in parentheses. + $p<0.10$ * $p<0.05$ ** $p<0.01$. 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). The models do not include any controls.