Does Feedback on Talk Time Increase Student Engagement? Evidence from a Randomized Controlled Trial on a Math Tutoring Platform

Dorottya Demszky
ddemszky@stanford.edu
Stanford University
United States

Rose E. Wang
Stanford University
United States
rewang@stanford.edu

Sean Geraghty
CueMath
India
sean.geraghty@cuemath.com

Carol Yu
CueMath
India
carol.yu@cuemath.com

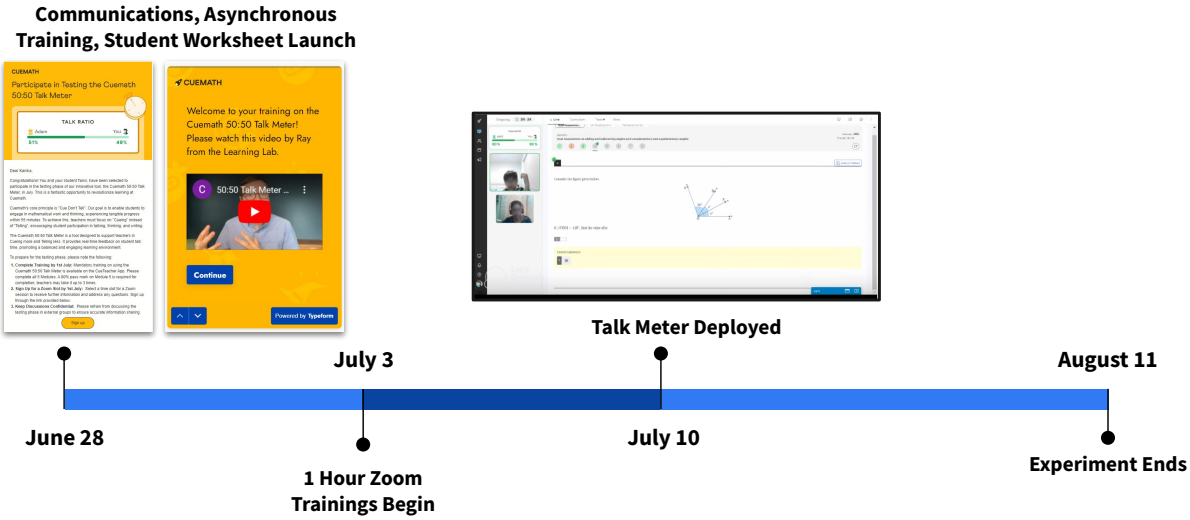


Figure 1: Timeline of our intervention.

ABSTRACT

Providing ample opportunities for students to express their thinking is pivotal to their learning of mathematical concepts. We introduce the Talk Meter, which provides in-the-moment automated feedback on student-teacher talk ratios. We conduct a randomized controlled trial on a virtual math tutoring platform (n=742 tutors) to evaluate the effectiveness of the Talk Meter at increasing student talk. In one treatment arm, we show the Talk Meter only to the tutor, while in the other arm we show it to both the student and the tutor. We find that the Talk Meter increases student talk ratios in both treatment conditions by 13-14%; this trend is driven by the tutor talking less in the tutor-facing condition, whereas in the student-facing condition it is driven by the student expressing significantly more mathematical thinking. Through interviews with tutors, we find the student-facing Talk Meter was more motivating to students, especially those with introverted personalities, and was effective at encouraging joint effort towards balanced talk time. These results demonstrate the promise of in-the-moment joint talk time feedback to both teachers and students as a low cost, engaging, and scalable way to increase students' mathematical reasoning.

LAK '24, March 18–22, 2024, Kyoto, Japan

ACM ISBN 979-8-4007-1618-8/24/03.
https://doi.org/10.1145/3636555.3636924

CCS CONCEPTS

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

KEYWORDS

student engagement, automated feedback, talk time, randomized controlled trial, math tutoring, joint feedback to students and teachers

ACM Reference Format:
Dorottya Demszky, Rose E. Wang, Sean Geraghty, and Carol Yu. 2024. Does Feedback on Talk Time Increase Student Engagement? Evidence from a Randomized Controlled Trial on a Math Tutoring Platform. In *The 14th Learning Analytics and Knowledge Conference (LAK '24), March 18–22, 2024, Kyoto, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3636555.3636924

1 INTRODUCTION

Talking about math is central to learning math [11]. In the U.S., both the National Council of Teachers in Mathematics (NCTM) and the Common Core [4] emphasize the importance of student discourse in representing, understanding and connecting math concepts, and encourage teachers to provide students with opportunities to express mathematical thinking in the classroom. However, most learning contexts still present a lot of room for improving student talk time and their engagement in STEM discussion, with teacher talk time ranging between 72-88% in whole classroom, small group and 1:1 learning contexts [14, 16, 17]. Typically, increasing student talk falls on the shoulders of teachers. To elicit student engagement, teachers have to use the right talk moves in the right moments, adapting their practice to the students' background, personality and learning style [11]. Such high-quality teaching practice takes a lot of coaching to master, which is unavailable to most teachers on a regular basis, especially in informal contexts [34]. And even for expert teachers, monitoring and effectively increasing student talk is challenging among numerous concurrent tasks they juggle during their teaching session.

Technological advancements have created novel opportunities to improve the quality of student-teacher interactions, via automated feedback to teachers. A recent line of work showed that teachers who receive automated feedback on student talk time and teacher talk moves *after* their teaching session improves their teaching practice as well as student engagement and satisfaction [16–18]. For example, Demszky et al. [18] conducted a randomized controlled trial in an online 1:1 mentoring context that showed that providing mentors with feedback on talk time and uptake of student ideas based on their session increased student talk time in subsequent sessions and improved students' experience with the program and optimism about their academic future [16]. Automated feedback to teachers thus seems to be an effective way to facilitate reflection and professional learning for teachers. However, such post-session feedback does not address the issue of the teacher being fully responsible for monitoring and increasing student talk real-time.

A parallel line of work indicates that gamified representations of student activities via points, badges and leaderboards can tap into students' intrinsic motivation and facilitate student engagement [13, 29, 39, 45, 48]. The gamification of learning activities helps distribute the cognitive load between the teacher and the student as students become active participants of their learning experiences [26, 36]. Could automated language-based feedback be shown real-time to students as well, to help encourage active learning in mathematics?

To answer this question, we study the effectiveness of *real-time* talk time feedback to *both* teachers and students in increasing students' engagement in mathematical discussion. We conduct a randomized controlled trial on the CueMath platform (n=742), which offers 1:1 virtual math tutoring to students worldwide. We introduce the Talk Meter, which provides intermittent feedback (every 20 minutes) to tutors and students on their talk ratio during the 55 minute tutoring session. We thus extend prior work by testing the effectiveness of in-the-moment—rather than post-teaching—feedback and by creating two treatment arms to compare the effectiveness of providing the feedback to *both the student and the tutor* to providing the feedback to the tutor alone.

Our study seeks to answer the following three key research questions:

(1) What is the impact of the Talk Meter on the tutor-student interaction, as measured by talk ratio, talk time, use of various language features such as focusing questions and mathematical terms?
(2) How did tutors perceive the Talk Meter and the impact it had on their instruction and student engagement?
(3) How did students perceive the Talk Meter?

We answer these questions through a mixed-methods approach: We use quantitative analyses to answer the first question, and qualitative interviews to answer the second and third question. We find that the Talk Meter increases student talk ratios in both treatment conditions by 13-14%; this trend is driven by the tutor talking less in the tutor-facing condition, whereas in the student-facing condition it is driven by the student expressing significantly more mathematical thinking. Through interviews, we find the student-facing Talk Meter was more motivating to students, especially those with introverted personalities, and was effective at encouraging joint effort towards balanced talk time. These results demonstrate the promise of in-the-moment joint talk time feedback to both teachers and students as a low cost, engaging, and scalable way to increase students' mathematical reasoning. This work also supports the hypothesis that joint feedback can be an effective way to lift the burden from the teachers' shoulders and help foster students' feeling of ownership over their learning.

2 RELATED WORK

2.1 Measuring Student Engagement with Talk Time

Student engagement in their learning environments, such as tutoring programs, often predict their learning achievement [22, 37, 50, 54, 59]. A simple measure of this is the talk time split across students and teachers [51, 52]. Increasing a student's talk time leads to learning opportunities for students to express mathematical thinking, fill in gaps in their understanding, and seek new information—all of which align with the Common Core State Standards for Mathematical Practices [4]. Additionally, increased student talk time can indicate that the student is motivated to actively learn [49, 57], engage in productive struggle [53] or build stronger relationships with their instructors [32]. Prior work in gamification for learning new languages focus on measuring student talk for capturing student engagement [31, 46].

2.2 Automating Feedback for Educators

Providing feedback to learners and instructors is critical for their growth [24]. With recent technological advances, there has been a growing number of efforts aimed at building automated feedback tools and analytical dashboards for educators, including information on educator and student talk time as well as other pedagogically relevant aspects of the discourse [TeachFX, 3, 6, 30, 56]. Such scalable and consistent feedback provides complementary advantages of expert human feedback, which is challenging to scale due to resource constraints. For example, Demszky and Liu [16] provides evidence through a randomized control trial in a 1:1 virtual mentoring context that automated feedback delivered to mentors after they complete their teaching session decreases mentor talk time by 6% and improves students' experience. We extend this work to evaluate the effectiveness of in-the-moment automated feedback on talk time at improving math tutors' instruction.

2.3 Sharing Feedback across Educators and Students

While a lot of prior education work has focused on providing feedback to either students or educators, less work has explored providing the same feedback to *both* students and educators. Previous works note the importance of feedback on literacy, for example studying how students and educators respond to the feedback they receive [9, 40]. Richardson [47] notes how feedback does not typically change how teachers instruct because they do not seriously respond to student evaluations. Another example is Chamberlin et al. [10], where they show how feedback for students—particularly negative feedback—enhanced anxiety and and demotivated students. These works study the lack of engagement with *asymmetrical* feedback, where feedback is written by one party and received by another. Our work explores the effectiveness of *symmetrical* feedback, where the same type of feedback like talk time is shared across both parties.

3 STUDY BACKGROUND

We conducted the study on CueMath[1], an education technology platform that offers 1:1 online math tutoring to 37,000+ students worldwide. Headquartered in India, an emerging economy with a 24% female labor force participation rate (World Bank, 2022), CueMath employs more than 3,000 tutors, 95% of whom are women, many with backgrounds in STEM fields. Sessions are conducted on Cuemath's proprietary platform, featuring video calls, a digital whiteboard, and curriculum-aligned materials. The study was approved under institutional IRB.

3.1 Tutor Training and Professional Learning

CueMath focuses on active learning, encouraging productive struggle as set forth by the National Council of Teachers in of Mathematics (NCTM): "Effective teaching of mathematics consistently provides students, individually and collectively, with opportunities and supports to engage in productive struggle as they grapple with mathematical ideas and relationships" [35]. The platform additionally focuses on ensuring that each session is in the zone of the student's proximal development [55]. This means that tutors need to forgo the temptation to lecture or explain for the majority of the session. Instead, they are expected to guide, prompt or "cue" the student, so that more of the cognitive work is done by the student. This provides the student with more opportunities to practice, perform and master a skill independently. CueMath onboards and trains all tutors. Coinciding with the experiment[2], CueMath retrained all of its tutors through in-person regional trainings to re-establish not just professional but pedagogical expectations regarding the aforementioned principles (see details in Appendix A).

Table 1: Demographics of our participant sample.

Tutors	Mean/%	SD	**Students**	Mean/%	SD
Total number	742		Total number	1,266	
Female	94%		Female	55%	
Age	40.40	8.17	*Grade level*	4.95	2.24
Years w/ CueMath	3.56	1.68	Elementary (GR1-6)	74%	
Baseline Talk Ratio	56%	18%	Middle (GR7-8)	21%	
			High school (GR9-12)	5%	
			Region		
			India	9%	
			UK	11%	
			US	58%	
			Rest of the world	23%	

3.2 Participants

A month before the intervention, we randomly selected 780 tutors for baseline data collection. For each tutor, we randomly selected up to two students that were assigned to the tutor, resulting in 1350 tutor-student pairs (some tutors only work with one student). Since 38 participants attrited from the sample during the baseline data collection period (due to inactivity, leaving the platform, or their students transferring to an out-of-sample tutor), the **final analytic sample includes 742 tutors and 1,266 students.** Table 1 summarizes the characteristics of the participant sample using available demographic information on CueMath. While most of the tutors are female (94%), genders are roughly balanced among students (55%). The average tutor age is 40 (SD=8.18) and they have about 3.6 years of experience at CueMath (SD=1.68). Their average talk percentage prior to the intervention is 56% (SD=18%), which is relatively low compared to 70-80% talk time observed in many other educational contexts [15, 16]. The majority of students are in elementary school (74%) and are located in the US (58%).

4 RANDOMIZED CONTROLLED TRIAL

We conducted a randomized controlled trial to evaluate the effectiveness of giving feedback to tutors and students on their talk ratios during their session. The study had three experimental arms: Control, TutorTM, TutorStudentTM. Participating tutors were randomly assigned to one of the arms. The Control group conducted "business as usual", without receiving feedback on their talk time. Below we describe the intervention for the two treatment groups, TutorTM and TutorStudentTM.

[1] https://www.cuemath.com

[2] The concurrence of the retraining with the experiment was accidental. Since it was offered to all tutors, it did not interfere with the randomization, but it did help ensure that all tutors were aware of the importance of encouraging student participation in the mathematical discourse.

4.1 Timeline & Trainings

The study was conducted for about six weeks between June 28 and August 11, 2023. Figure 1 includes the timeline with three relevant dates that indicate launches for trainings and communication about the TalkMeter. Only the two treatment groups (TutorTM and TutorStudentTM) received these trainings. As mentioned in Section 3.2, a month prior the experiment (May 22), we started to collect baseline data for the study to observe instructional practices prior to the randomized intervention.

On June 28, treatment group tutors received an email that explained that 1-2 students of theirs were selected to be part of a pilot for a new product feature on the tutoring platform. The email included brief pedagogical rationale behind the tool (see online supplement). Tutors were also told to complete an asynchronous training, and join a live Zoom training before the deployment of the feature on July 10. Five asynchronous training modules were released. The talk meter was referred to as the "50:50 talk meter" to encourage an average student and teacher talk ratio of 50:50 in classes. Tutors in the TutorStudentTM group were given additional messaging and resources to brief their students before the new feature launched. They were asked to complete a Student Worksheet with participating students, which was designed to help students understand the learning impact of them talking out loud and explaining their thinking to their tutors (see excerpts in online supplement).

From July 3 to 10, several one hour Zoom sessions were held in groups of 20-50 to go over additional content on strategies to increase student talk, and elicit student thinking during tutorial. Tutors watched video of tutorials that had high student talk and low student talk, and discussed them together. Finally, on July 10, the Talk Meter was deployed to tutor-student pairs, according to their treatment group assignments.

4.2 The Talk Meter

The first treatment group (TutorTM) received a tutor-facing TalkMeter, as part of which, every 20 minutes during the class session, a frame appeared within the video calling session that showed tutor their talk ratio (Figure 2a). The talk meter appeared 20 minutes into class, then 40 minutes into class, and at the end of class (Figure 2b).[3] Its appearance during class lasted for 1 minute. Results were shaded as red (student talk <= 25%), yellow (student talk between 25-50%) or green (student talk >= 50%), to indicate improvement required. In the second treatment group (TutorStudentTM), the TalkMeter was also visible to the student, to encourage participation via joint reflection on talk ratios.

CueMath calculated talk times for the student and the teacher by aggregating periods of continuous sound captured by their microphones. Talk ratios were calculated by dividing the duration of student speech by the total duration of student and teacher speech. For example, in an hour-long class where the student spoke for 20 minutes and the tutor spoke for 25 minutes and the rest was silence, the talk ratio would be 44:56 (student talk:teacher talk)%. We did not measure the duration of silence as it can happen for many reasons that we do not have a way to disentangle (e.g., the student working on a problem, the recording staying on before or after class).

[3]We considered a continuous version for the Talk Meter but decided on intermittent instead because a) talk time did not change dramatically minute to minute, and b) based on feedback we received from tutors in a small pilot with a different group of tutors showed that tutors preferred the intermittent one and found that the continuous one can more easily be ignored or become a distraction.

4.3 Recordings & Transcripts Collected

We collected 22,845 session recordings throughout the study, out of which 10,811 were collected during the baseline period and 12,034 were collected during the experimental phase. Each tutoring session is scheduled for 55 minutes. We transcribe a random subset of 4436 recordings for each tutor given the high costs of transcribing the entire dataset. We selected the earliest baseline recording available, and 2 of the most recent recordings from the experimental phase for each teacher. We used DeepGram[4] to transcribe these recordings and used the transcripts for the language analysis described below.

4.4 Measures of Outcomes

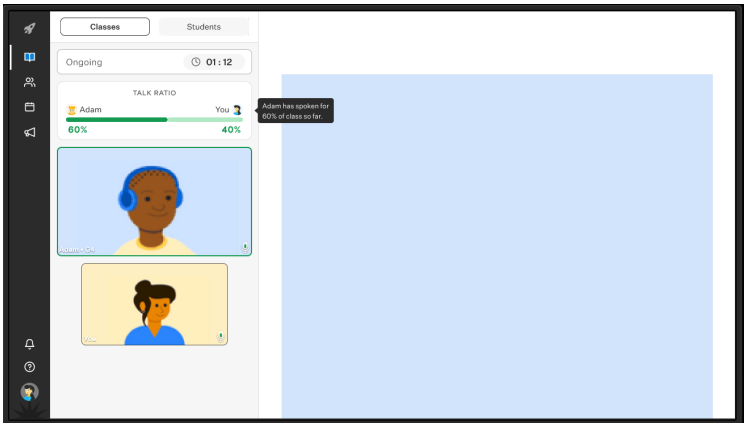
Since our primary research question focuses on understanding the impact of the Talk Meter on the student-teacher interaction, we use talk ratio, talk time and several language-based measures that capture changes in the tutoring discourse. CueMath does not track learning outcomes that are standardized across regions, and its students also enroll in CueMath at different points throughout the year. Thus, we are unable to measure the impact of the intervention on students' performance. We explain each of the outcomes below.

4.4.1 Talk Ratio and Talk Time. Student talk ratio and student talk time are key outcomes, being primary intervention targets. We compute talk ratios as defined in Section 4.2, as the ratio of student talk to the total amount of student and teacher talk. To better understand the amount of change in student and teacher talk, we also use calculate their talk time in minutes.

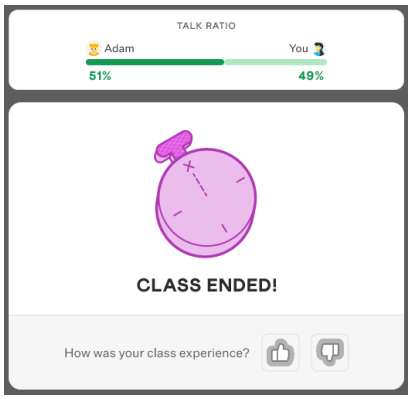
4.4.2 Language Measures. We use natural language processing (NLP) to identify several language-based features that estimate presence of high-leverage mathematics instructional practices. We use four open-source measures developed and validated by prior work [2, 15, 28] on a dataset of elementary math classroom transcripts. We chose these measures as they were readily available to the research team and because they had been correlated positively with expert observation scores of instruction quality and students' academic outcomes in math instructional datasets. These four language-based measures capture teachers' use of focusing questions, teachers' uptake of student ideas, student reasoning as well as students' and tutors' use of mathematical terms. The models receive a transcript of a tutoring session as input, and output binary or continuous predictions for each utterance in the transcript, as described below. We aggregate these predictions to the transcript-level to generate outcomes.

Focusing questions. Students are more engaged and learn more when teachers pose focusing questions — defined as questions that attend to what the students are thinking, pressing them to communicate their thoughts clearly, and expecting them to reflect on their thoughts and those of their classmates [2, 7, 25, 44]. The use of focusing questioning patterns has been linked to better student learning outcomes and confidence in mathematics [21, 23]. Prior

[4]https://deepgram.com



(a) Wireframe for Talk Meter, shown every 20 minutes during class.



(b) Talk Meter shown once class ends.

Figure 2: Components of the TalkMeter.

work developed models for computationally identifying focusing questions, training on math classroom data [2, 15, 18]. Our work uses the fine-tuned RoBERTa model [38] from [15] to identify focusing questions in the tutor's utterances (binary variable).

Teachers' uptake of student ideas. Teachers' uptake of student ideas, e.g. via revoicing or elaboration, promotes dialogic instruction by amplifying student voices and giving them agency in the learning process [5, 12, 41, 58]. Such uptake can be an indicator of responsive teaching and has been linked to higher student achievement [8, 15, 19, 42, 43]. Prior work has developed and validated a measure of uptake [19] and has shown that this measure can be can provide successful feedback to instructors in group and 1:1 settings [16, 17]. Our work uses [19]'s fine-tuned Bert model [20] to identify uptake in tutors' utterances (binary variable).

Student reasoning. Student reasoning is a strong indicator of dialogic instruction where students are active participants of the learning process [1, 56]. We use a fine-tuned RoBERTa model [38] from prior work [15] that was trained on an elementary math classroom dataset annotated by expert educators with a definition of student mathematical reasoning adapted from the widely used Mathematical Quality of Instruction (MQI) observation protocol's "Student Provide Explanations" [27] item. We apply this model to student utterances to detect mathematical reasoning (binary variable).

Use of mathematical terms. The use of mathematical terms is one indication of students' engagement in mathematical thinking. Educators play a critical role in exposing students to mathematical terms, be it through connecting these terms to mathematical content or representations in their instruction. They also play an important role in encouraging students to practice using the terms Prior work collected a dictionary of mathematical terms and, in the setting of elementary school mathematics classrooms, found that students whose teachers use more mathematical language are more likely to use it themselves [28]. Additionally, these students of higher mathematical term use perform better on standardized tests. We use this dictionary of mathematical terms to identify the total number and the unique number of mathematical terms used by students and tutors.

4.5 Post-Study Interviews and Video Observations

For qualitative insights, we randomly selected 10 tutors total from TutorTM and TutorStudentTM to participate in a 15 minute interview. We also randomly selected 19 students total from TutorStudentTM to participate in a 15 minute interview. To avoid bias, all interviews were conducted by a member of CueMath's Learning Lab who was not involved in the experiment. For tutors, the interviewer asked three questions: 1) "How was your overall experience in using the Talk Meter?", 2) "You received your talk ratio results for each class with [student name] for about 1 month. How did this change the way you taught?", 3) "Is there anything else you want to share with us?".

For students, the interview asked: 1) "In the past month, did you see something called a Talk Meter?" 2) "What do you think about it? Does it help you?" and 3) "Which would you prefer, a class with the Talk Meter or without?" All classes from Control, TutorTM and TutorStudentTM were recorded. Members of CueMath's Learning Lab also randomly watched video recordings to observe how students and teachers reacted to the Talk Meter.

4.6 Regression Analyses

We model the impact the intervention had on tutors' practice via an ordinary least squares regression. We run a separate regression to estimate the effect of the treatment on each dependent variable described in Section 4.4 above. Concretely, we measure the impact of the intervention on student talk ratio and talk time and frequency of each language feature (Section 4.4.2). The models are specified as $Y_{it} = \beta_1 T_i + \beta_2 \boldsymbol{X}_i + \beta_3 \boldsymbol{m}_{it} + \varepsilon_{it}$ where Y_i refers to a particular dependent variable for tutor $i's$ transcript t; T is a factor variable that indicates the treatment status, with a value of 0 indicating Control, 1 indicating TutorTM and 2 TutorStudentTM; $\boldsymbol{X}$ is a vector of tutor and student-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 and student level to account for repeated observations within a teacher and student.

We use the following binary variables as tutor and student covariates $\boldsymbol{X}$ across all models: tutor is female, tutor age, tutor CueMath years, student is female, student grade and student region. We also include baseline baseline language features from tutors' first recording as covariates. For analyses using student talk ratio and talk minutes, we include students' baseline talk ratio, students' baseline talk minutes and tutor baseline talk minutes as covariates. For analyses using language features as dependent variables, we additionally include baseline values for all language features as covariates. The reason why do not include these baseline language features as covariates for the other models is because we only have them available for a subset of the data, and hence including them would restrict the analytic sample. In all models, we additionally include the session count for the given tutor-student pair as the transcript covariate $\boldsymbol{m}$.

We also conduct **heterogeneity analyses** to understand how the impact of the treatment on talk ratio and talk time might vary across participants, especially as it relates to their compliance with trainings. We study heterogeneity based on binary indicators of whether the tutor had an above average or below average baseline talk ratio, whether the student completed the talk time worksheet, and whether the tutor completed relevant trainings. For these analyses, we use the same model as described above, but instead of representing T as a factor variable with three levels, we use a binary indicator for treatment status. We include an interaction term between T and the heterogenous variable of interest. Since the student worksheets were only available to TutorStudentTM, we exclude TutorTM from the analysis that uses student worksheet completions as a dependent variable.

Since training and worksheet completion is affected by selection bias, we cannot draw causal relationships between the heterogeneous variables and the outcome. What these analyses do help us understand is what characteristics may be predictive of intervention success for participants. For example, while we can't determine if worksheet completion *causes* greater improvement in student talk ratios, we can understand if a tutor's decision to have their student complete the worksheet *is correlated with* a greater improvement in their talk ratios.

4.7 Validating Randomization

To verify whether our randomization was successful, we evaluate whether the characteristics of each group differ significantly via a three-way ANOVA. We compare tutor and student demographics, the validity of the recording and discourse features measured in tutors' first recorded baseline lesson. As the p values in Appendix B Table 5 show, we do not find statistically significant differences among conditions in any of the characteristics. This suggests that any differences we observe later in the course are likely due to the effects of the intervention.

5 RESULTS

In this section, we summarize both the quantitative and qualitative results of the Talk Meter intervention. For the quantiative analyses (Sections 5.1-5.2), we provide a breakdown of results for each outcome variable introduced in Section 4.4. As for qualitative findings, we provide a summary of post-study interviews.

5.1 Impact on Talk Ratios and Talk Time (Research Question 1)

Table 2 summarizes the main results. The results show that the TalkMeter significantly increases student talk both overall and in relation to teacher talk. In both treatment conditions, we observe a similar increase in students talk ratios: in the TutorTM group, the student talk ratio increases by 5.67% ($p < 0.01$), showing a 13% increase compared to the Control group mean (43%), and in the TutorStudentTM group, the talk ratio increases by 6.10% (14% more than Control, $p < 0.01$). However, the increase in student talk ratio is explained by different patterns across the two conditions. In TutorTM, the tutor decreases their talk time more, talking -1.744 minutes less on average (14% less than Control, $p < 0.01$), while the the student is only talking .73 more minutes on average (7% more than Control, $p < 0.01$). In contrast, students in the TutorStudentTM condition increase their talk time by 1.83 minutes (18% more than Control, $p < 0.01$) while the tutor talking only 0.92 minutes less (7% less than Control, $p < 0.01$). Thus, the similar improvement in student talk ratios between the two conditions is driven primarily by the tutor striving to talk less in TutorTM and the student striving to talk more in TutorStudentTM.

To better understand how treatment effects change over time, we computed regressions separately for each session, using the same covariates as shown in Table 2. The results are plotted in Figure 3, with the left figure showing treatment effects for student talk ratios and the right plot showing treatment effects for student talk in minutes over time, separated by condition. These plots offer three primary takeaways. First, we can see that treatment effects generally increase in the first three sessions, after which they plateau (with some variance, e.g. an unexplained dip for session 5 for student talk minutes). Second, while the coefficients for student talk ratios is only significantly greater for TutorStudentTM compared to TutorTM in session 1 and 7, the coefficients are consistently much greater TutorStudentTM compared two TutorTM for student talk minutes. This trend demonstrates that the results from the analysis in Table 2 represent a consistent pattern in the student-facing Talk Meter being more successful at increasing the amount of student talk than the tutor-facing Talk Meter alone. Third, zooming into session 1, the TutorStudentTM shows an immediate increase in student talk while in TutorTM it takes one additional session until we can observe a significant increase in student talk compared to the treatment group. This suggest that it takes more time for the tutor to increase student engagement when they are the only recipients of the talk time feedback.

Finally, we study the how different student and tutor characteristics — with a focus on compliance with trainings — correlate with treatment effects. Following the approach described in Section 4.6, we conduct binary comparisons across student-tutor pairs with above vs below average baseline student talk ratio, whether the student completed the worksheet and whether the tutor completed the asynchronous training, the Zoom training (Section 4.1) or the company-wide re-training (Section 3.1). Figure 4 shows the results

Table 2: Impact of the TalkMeter on student talk ratio and talk time in minutes. Standard errors are in parentheses. + p<0.10 * p<0.05 **. Each column displays the results of a separate regression. We omit covariates (as described in Section 4.6) from this table for readability – the full table is included in Appendix C. The results show a significant increase in student talk, both overall (student talk minutes) and in relation to teacher talk (talk ratio, teacher talk minutes).

Independent variable	(1) **Student talk ratio**	(2) **Student talk mins**	(3) **Teacher talk mins**
Group=TutorTM	**5.669****	**0.731****	**-1.744****
	(0.655)	(0.201)	(0.258)
Group=TutorStudentTM	**6.100****	**1.830****	**-0.924****
	(0.676)	(0.232)	(0.271)
Control Mean	43.014	10.093	12.531
R^2	0.385	0.461	0.409
Observations	8972	8972	8972

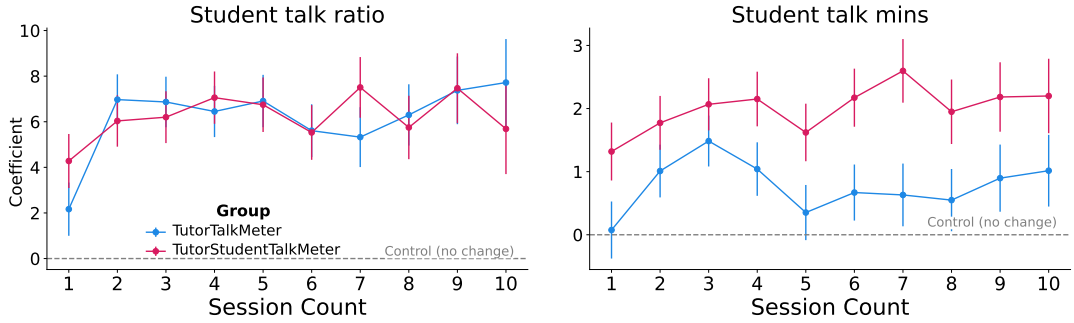


Figure 3: The impact of the TalkMeter on student talk ratio and student talk in minutes, plotted over time by session count. Each dot in the plot represents a separate regression, with the same covariates as those in Table 2. The error bars represent standard errors obtained in the regressions. The colors represent the condition (TutorTM or TutorStudentTM). The trend shows that while student talk ratios differ significantly only for the first session, the treatment effects on student talk minutes are consistently different across condition over time.

of these analyses. Perhaps the most noticable finding is that students who completed worksheets showed a three times greater increase in talk ratios, and a six times greater increase in talk minutes compared to those who did not complete the worksheets. This finding indicates that tutors' encouragement and students' willingness to complete the worksheet relate to a much larger impact of the Talk Meter. Similarly, although with smaller effect sizes, we see that tutors' compliance with all three trainings, especially the ones specifically designed for the Talk Meter (async and Zoom), correlate with approximately a 1.5 greater treatment effect in student talk ratios and a two times greater impact in student talk minutes. And finally, we find that the Talk Meter had a ~1.2 greater impact on tutor-student pairs with a below average student talk ratio compared to those with an above average talk ratio. This indicates that the intervention is more successful for participants who have more room for improvement.

5.2 Language Features (Research Question 1)

Our final quantitative analyses focus on the impact of the intervention on tutor and student discourse features. Table 3 summarizes the results. We find that tutors in both treatment conditions significantly ask more **focusing questions**; by 13% for TutorTM ($p < 0.05$) and 14% for TutorStudentTM ($p < 0.01$) compared to the Control group. This indicates that although tutors decrease their talk time, they do increase their use of questions that probe the students' thinking. Tutors also marginally increase their uptake of student ideas in TutorStudentTM (by 6%, $p < 0.1$), but not in TutorTM. Finally, along with a decreased talk time we see fewer math terms by tutors in both conditions ($p < 0.01$ for TutorTM and $p < 0.05$ for TutorStudentTM). In contrast, we find that students increase their **overall use of math terms** in both treatment groups (by 13% for TutorTM ($p < 0.05$) and 14% for TutorStudentTM ($p < 0.01$) compared to the Control group). These results suggests that teacher math talk is being "replaced" by student math talk during the tutoring session. And importantly, we find that in TutorStudentTM, but not in TutorTM, students also use 18% more **unique math terms** ($p < 0.01$) and 24% more **student reasoning** ($p < 0.01$) compared to the Control group. These findings indicate that the student-facing talk meter elicited more diverse use of terms and an increased talk out loud reasoning in students compared to the Control and TutorTM conditions.

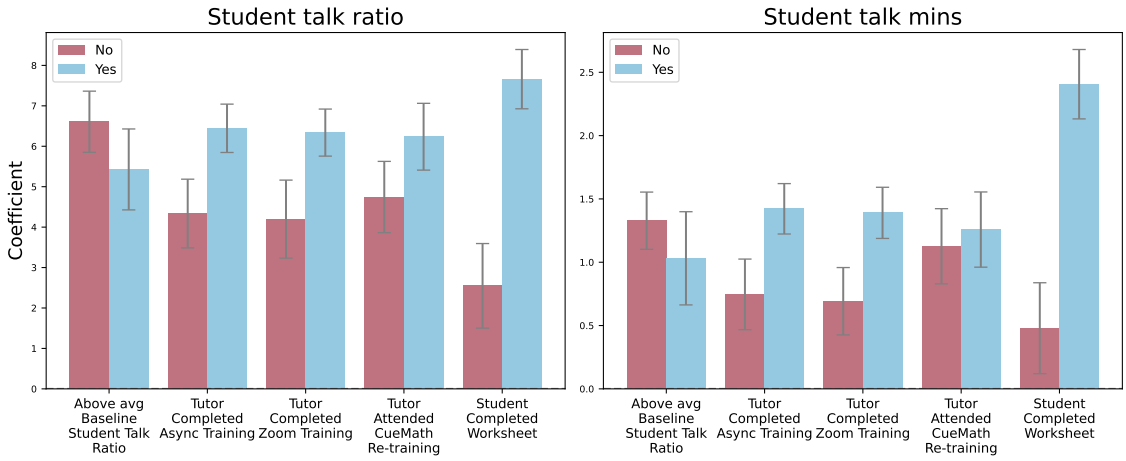


Figure 4: The impact of the TalkMeter on student talk ratio and student talk in minutes, plotted separately based on whether the tutor-student pair had an above or below average baseline student talk ratio, whether the student completed the talk time worksheet (TutorStudentTM only), whether the tutor completed the asynchronous and Zoom trainings (offered to both TutorTM and TutorStudentTM), and whether tutors completed the company-wide re-training. Each pair of barplots represents a separate regression, with the same covariates as those in Table 2 but with an added interaction term between the heterogeneous variable and the treatment condition. The error bars represent standard errors obtained in the regressions. The trends show that completing the talk time worksheet and trainings correlates with a greater treatment effect, and so does having a below average baseline student talk ratio.

Table 3: Impact of the Talk Meter on language features.

	(1) Num. tutor focusing questions	(2) Num. tutor uptakes	(3) Num. tutor math terms	(4) Num. tutor unique math terms	(5) Num. student math terms	(6) Num. student unique math terms	(7) Num. student reasoning
G=TutorTM	0.987*	-0.225	-11.768**	-1.107**	5.567*	0.522	0.208
	(0.390)	(0.818)	(3.859)	(0.426)	(2.748)	(0.356)	(0.138)
G=TutorStudentTM	1.072**	1.383+	-9.670*	-0.873*	15.684**	1.566**	0.396**
	(0.404)	(0.813)	(3.980)	(0.394)	(3.253)	(0.349)	(0.145)
Control Mean	7.523	22.213	75.571	13.519	37.540	8.581	1.651
R2	0.275	0.398	0.188	0.239	0.220	0.244	0.194
Observations	2318	2318	2313	2313	2315	2315	2318

5.3 Teacher Interviews (Research Question 2)

There are a couple of core themes that cut across interviews in both treatment groups. Tutors mentioned that the Talk Meter provided them with **more awareness** of what was actually happening during the session and reminded them to encourage the student to talk. One tutor (TutorTM) said, "…if there is no talk meter, we are not aware how much a teacher is talking in the class and how much the student is talking in the class.". Another tutor (TutorTM) admitted, "It wasn't something that I kept in my mind that I need to ensure that the child is speaking. But when the talk meter came in, I think it was like a reminder that I need to get the child to speak out. So there are questions that I came up with frequently…Now I try and give those prompts to make sure the child has interactions.". Tutors in both treatment groups also mentioned that their two **students were temperamentally different**, and that the impact of the Talk Meter varied by the student. One (TutorStudentTM) explained, "It's more impacted with [Student A] because [Student A] is one of my students who was really introvert. He hardly used to talk with me.... So once after this talk ratio, and still I'm struggling, but I think his participation has definitely increased.[…] [My other student S] is always excited. See, we have been keeping 50:50 ratio. And then sometimes he even said that see ma'am, I got the major ratio. I have been talking more. You're not letting me to talk."

Differences also emerged from the tutors in TutorTM and TutorStudentTM. TutorTM participants' feedback focused more on their increased awareness, and their efforts to reduce long explanations, or to hold back to let the student speak instead. TutorStudentTM participants' feedback focused on ways the tool **shared some of the teacher's burden in motivating the student** – especially introverted students – to speak more in class. In TutorStudentTM, a tutor also reflected on how the worksheet helped her student realize the importance of talking more: "[Student S] in fact, had struggles with math. She [...] was half a grade below her actual grade when she joined in.[...] So when [we] went through

that sheet [...] she herself could arrive at that, oh, okay, this is why I should talk, that changed it for her. And what's been happening is she notices the talk meter. She actually notices. She's really proud of herself, and she does, oh, I spoke 60% of the time, or I spoke 70%. So that's been happening.". Tutors brought up objectivity of the feedback ("it becomes easier for a kid to take something that's very factual rather than coming from a person.") and the gamification of obtaining a 50:50 ratio as factors that contribute to the effectiveness of the student-facing TalkMeter ("The kids are also excited to see. They themselves know now that after 20 minutes after 40 minutes it will come up and I have to maintain my talk time.").

Although most of the feedback was positive, tutors also shared challenges, such as feeling pressured to stick to a 50:50 ratio, or **feeling unnatural when holding back from speaking**. A tutor (TutorTM) argued that equal talk ratio may not be possible or desired in all context: "...every time it's not possible. Like when we are introducing a new concept to the child or we are doing more puzzle cards in the class, a teacher has to speak more because if I'm cueing the child but puzzle cards are really very struggling for a child, we need to speak more. And the talk meter flashes that you are speaking more. So sometime it hampers the learning but overall, in my overall experience it hampers only a few times, but the impact is good more in more classes.".

Finally, many of them shared the strategies they used, such as using prompts to get their student speaking, or to **ask open-ended questions** ("Firstly I have to ask open ended question to the student...what did they understand by the question or how should they go about the solution?"; TutorTM), or **shortening explanations** ("So what really changed, I think, was the long explanation...whereas the other bit, making them involved[...] might not have changed, but keeping explanation short, I think that is something I took away or that is something that I'm consciously doing a lot more after the whole thing. In other things, I think I was already doing it, but it just sort of got more reinforced."; TutorStudentTM).

5.4 Student Interviews & Video Observations (Research Question 3)

Two themes cut across the majority of student responses. Most students (14 of 19) had a **clear understanding** of what the Talk Meter was, what it was intended for, and many students (10 of 19) also noted that talking more during tutorial was desirable. One student noted "It has its uses as to encourage [the student] to talk a bit more". A notable subset of students (6 out of 19) also echoed **gamification and competition** themes we observed when directly watching classes ourselves (see below). These students viewed the Talk Meter as a fun, interactive tool. One student expressed, "It's like a competition. So if you talk more, it's like, I think you're better at it." Another noted "When I see that it's red, I get a little bit sad and then I keep on talking, then I see it yellow, and then I keep on talking more. Then I see it green and then I'm super happy".

The general feedback from students was **predominantly positive**, with 12 students expressing favorable views. One student said, "Well, it gets me involved with questions, and I have the courage to ask questions, so it's pretty helpful". However, 4 students had neutral responses, and 3 expressed negative views, citing the tool's **occasional intrusiveness** during focused activities, "It can get annoying because sometimes when I'm trying to look at a question, it just appears, and then sometimes I can't get rid of it.".

A random selection of video recordings revealed similar themes as the interviews, but also highlighted how students approached the Talk Meter. Many children approached it as a game, and as a welcome way to break up a 55 minute session. Below, we present 2 example exchanges, that are representative of many other video recordings. Student-Teacher Pair 1 has a more reserved and quiet student, whereas Student-Teacher Pair 2 has a more effusive, talkative student.

6 DISCUSSION

We deployed a Talk Meter on the CueMath platform to test the hypothesis that visually rewarding student talk in the moment would lead to more productive student talk and thinking during class. We also tested if the TalkMeters' impact and reception would vary if results were shown just to the tutor, or shown both to the student and tutor. Three key take-aways emerge from the study. First, the Talk Meter in **both treatment conditions increased students' math-related talk**, as shown by the significant increase in student talk ratios, student talk minutes and use of mathematical terms, observing similar effect sizes as a previous study on feedback in 1:1 online teaching contexts [16]. Given the one-time cost of building the feature and the added trainings, this intervention shows promise for scalable implementation [33].

A second take-away is that although the impact on student talk ratio was similar across TutorTM and TutorStudentTM, student and teacher experiences were different between the two groups. Overall, the student and teacher-facing Talk Meter generated more ownership from the student in an engaging and unpressed manner, facilitating joint effort between the student and teacher in creating a class where the student does more talking and thinking. While the **change in ratio for TutorTM was driven largely by the teacher talking less, the change in ratio for TutorStudentTM was driven by the student speaking more**. Further, whereas TutorTM not exhibit increased student reasoning, TutorStudentTM increased student reasoning and use of unique math terms by terms by as much as 24% and 18%, respectively, indicating a moderate effect size. This suggests that the increase in student talk is not just superficial, but that it reflects increase in substantive mathematical thinking. Qualitative interviews and video observations corroborate the quantitative results, indicating that the student-facing TalkMeter motivated students to talk more and led to positive, lighthearted interactions upon its appearance. This result sheds light onto a new area — automated, language-based feedback during instruction — where gamification can increase student engagement.

Third, both interventions resulted in some **"substitution" of cognitive work from the tutor to the student**. This is consistent with the objective of the study for "students to do more of the cognitive work of talking, thinking, and writing themselves." In TutorTM, this exchange largely occurred through math terms; students used more while teachers used less compared to the control. In TutorStudentTM, the same exchange happened with math terms, but was more pronounced - students used 42% more terms relative

Table 4: Transcript excerpts

Student-Tutor Pair 1	Student-Tutor Pair 2
	Student: Where did my talk ratio go? It's not here yet. **Tutor:** Yeah, it'll come. It came to me. 78% you and 22% me. **Student:** (Calls her sister.) My talk ratio is going to come soon, like in less than a minute.
Student: I have more than you in the talk ratio! **Tutor:** You're almost 50:50. (Smiles.) **Student:** Okay. If COA equals 110, find the value of x. . . (Returns to math.)	**Tutor:** Yeah. See! **Student:** So this is how much I talk during the class and this is how much she talks during the class. Basically even the first one. Last time, I think it was at like 13% for me and the rest for the teacher. **Tutor:** Yes. (Laughs.) **Student:** And that was very bad. **Tutor:** Awesome! Clap for you. **Student:** Yay. I'm awesome! (Sings "Everything Is Awesome.")

to control versus 14% more in TutorTM. In essence, the student-facing talk meter reconfigured the tutor-student exchange: students spoke more, used more math terms, and more frequently provided explanations. In response, tutors asked better questions and built on contributions more frequently through uptake. While there appears to be zero-sum trade-off on math terms in both treatment groups (students use more, teachers use less), in TutorStudentTM increased student reasoning seems "positive-sum," as it elicits better teacher questions and uptake of student ideas.

One primary limitation of this study is the absence of learning outcomes and measures on students' confidence and beliefs regarding math. In future work, we hope to collect outcome measures on students' performance, confidence and beliefs. A second limitation is that since trainings and the worksheet were only offered to the treatment groups, we cannot study their causal influence on the tutoring session. A future experiment could disentangle the impact of these trainings from the impact of the Talk Meter via a randomized design. A third limitation relates to the representativeness of the sample. In addition to demographic representation (with tutors being Indian women, and limited information on students), CueMath sessions also show, for example, a higher average student talk ratio (43%) in the control group than other contexts (online 1:1 mentoring: 28% [16], online small group: 20% [17]) – suggesting that CueMath sessions may not be representative of other teaching contexts.

Evaluating the Talk Meter in other learning contexts, such as regular classrooms, student group work and subjects beyond math, and with different teacher and student populations is a highly promising direction for future work. Doing so would help us understand the context-dependence of effect we observe, and would also help us adapt the Talk Meter to the needs of teachers and students in different learning context and from different cultural and demographic backgrounds. It is also crucial to conduct thorough fairness evaluation to ensure that the Talk Meter is not biased against certain tutor or student populations. For example, imprecise measurement due to the students speaking a certain dialect or in the presence of background noise that may correlate with socioeconomic factors, can create inequities in the quality of feedback received by students and tutors.

Finally, future research should explore how we can make the TutorStudentTM even more effective, and address some of the concerns (e.g. intrusiveness) mentioned by tutors and students in the interviews. Would adaptive feedback timing, additional gamification, or a different type of design or metric be even more successful at motivating student talk and thought? In future iterations, we would also like to study the effectiveness of providing feedback to tutors and students on the content of their speech, e.g. by using the language measures described in this paper. How can such language-based feedback be delivered to students and tutors in a way that is not overwhelming, and effective at facilitating active learning?

ACKNOWLEDGMENTS

We would like to thank Ranjith Babu, Rashmi Arora for their contributions to our analysis, Anushray Gupta, Akash Antil and Sandeep Gunduboyina for building the feature, and Andrew Ho for his insights on methodology.

REFERENCES

[1] Robin John Alexander. 2008. Towards dialogic teaching: Rethinking classroom talk. (2008).
[2] Sterling Alic, Dorottya Demszky, Zid Mancenido, Jing Liu, Heather Hill, and Dan Jurafsky. 2022. Computationally Identifying Funneling and Focusing Questions in Classroom Discourse. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*. 224–233.
[3] Sinem Aslan, Nese Alyuz, Cagri Tanriover, Sinem E Mete, Eda Okur, Sidney K D'Mello, and Asli Arslan Esme. 2019. Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.
[4] National Governors Association et al. 2010. Common core state standards. *Washington, DC* (2010).
[5] M. M. Bakhtin. 1981. *The dialogic imagination: four essays*. University of Texas Press.
[6] Roghayeh Barmaki and Charles E Hughes. 2015. Providing real-time feedback for student teachers in a virtual rehearsal environment. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 531–537.
[7] David Blazar. 2015. Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review* 48 (2015), 16–29. https://doi.org/10.1016/j.econedurev.2015.05.005
[8] Jere E Brophy. 1984. *Teacher behavior and student achievement*. Number 73. Institute for Research on Teaching, Michigan State University.
[9] David Carless and David Boud. 2018. The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education* 43, 8 (2018), 1315–1325.
[10] Kelsey Chamberlin, Maï Yasué, and I-Chant A Chiang. 2023. The impact of grades on student motivation. *Active Learning in Higher Education* 24, 2 (2023), 109–124.
[11] Suzanne H Chapin, Mary Catherine O'Connor, and Nancy Canavan Anderson. 2009. *Classroom discussions: Using math talk to help students learn, Grades K-6*. Math Solutions.
[12] James Collins. 1982. Discourse style, classroom interaction and differential treatment. *Journal of Reading Behavior* 14, 4 (1982), 429–437.
[13] Luma da Rocha Seixas, Alex Sandro Gomes, and Ivanildo José de Melo Filho. 2016. Effectiveness of gamification in the engagement of students. *Computers in*

Human Behavior 58 (2016), 48–63.
[14] D. Demszky and H. Hill. 2022. The NCTE Transcripts: A dataset of elementary math classroom transcripts. *arXiv preprint* (2022). arXiv:2211.11772.
[15] Dorottya Demszky and Heather Hill. 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 (BEA 2023)*. Association for Computational Linguistics, Toronto, Canada, 528–538. https://doi.org/10.18653/v1/2023.bea-1.44
[16] 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* (Copenhagen, Denmark) *(L@S '23)*. Association for Computing Machinery, New York, NY, USA, 59–69. https://doi.org/10.1145/3573051.3593379
[17] Dorottya Demszky, Jing Liu, Heather 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* (May 2023).
[18] Dorottya Demszky, 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. (2023).
[19] D. Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori 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)*.
[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.
[21] Megan Loef Franke and Elham Kazemi. 2001. Learning to teach mathematics: Focus on student thinking. *Theory into practice* 40, 2 (2001), 102–109.
[22] Guher Gorgun, Seyma Nur Yildirim-Erbasli, and Carrie Demmans Epp. 2022. Predicting Cognitive Engagement in Online Course Discussion Forums. In *Proceedings of the 15th International Conference on Educational Data Mining*, Antonija Mitrovic and Nigel Bosch (Eds.). International Educational Data Mining Society, Durham, United Kingdom, 276–289. https://doi.org/10.5281/zenodo.6853149
[23] Sara Hagenah, Carolyn Colley, and Jessica Thompson. 2018. Funneling versus Focusing: When Talk, Tasks, and Tools Work Together to Support Students' Collective Sensemaking. *Science Education International* 29, 4 (2018), 261–266.
[24] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.
[25] Beth Herbel-Eisenmann and M. Breyfogle. 2005. Questioning Our Patterns of Questioning. *Mathematics Teaching in the Middle School* 10 (05 2005), 484–489. https://doi.org/10.5951/MTMS.10.9.0484
[26] J. Hiebert and D. A. Grouws. 2007. *The effects of classroom mathematics teaching on students' learning*. Vol. 1. 371–404.
[27] Heather C Hill, Merrie L Blunk, Charalambos Y Charalambous, Jennifer M Lewis, Geoffrey C Phelps, Laurie Sleep, and Deborah Loewenberg Ball. 2008. Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and instruction* 26, 4 (2008), 430–511.
[28] Zachary Himmelsbach, Heather C. Hill, Jing Liu, and Dorottya Demszky. 2023. A Quantitative Study of Mathematical Language in Classrooms. *EdWorkingPapers* 855 (October 2023). http://www.edworkingpapers.com/ai23-855
[29] Caitlin Holman, Stephen Aguilar, and Barry Fishman. 2013. GradeCraft: What Can We Learn from a Game-Inspired Learning Management System?. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (Leuven, Belgium) *(LAK '13)*. Association for Computing Machinery, New York, NY, USA, 260–264. https://doi.org/10.1145/2460296.2460350
[30] Nicholas Hunkins, Sean Kelly, and Sidney D'Mello. 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*. 230–238.
[31] MIF Kareema. 2014. Increasing student talk time in the ESL classroom: An investigation of teacher talk time and student talk time. (2014).
[32] Min-Young Kim and Ian AG Wilkinson. 2019. What is dialogic teaching? Constructing, deconstructing, and reconstructing a pedagogy of classroom talk. *Learning, Culture and Social Interaction* 21 (2019), 70–86.
[33] Matthew A Kraft. 2020. Interpreting effect sizes of education interventions. *Educational Researcher* 49, 4 (2020), 241–253.
[34] 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
[35] Steve Leinwand. 2014. *Principles to actions: Ensuring mathematical success for all*. National Council of Teachers of Mathematics.
[36] D. Lemov. 2010. *Teach like a champion: 49 techniques that put students on the path to college (K-12)*. John Wiley and Sons.
[37] Mark R Lepper and Jennifer Henderlong. 2000. Turning "play" into "work" and "work" into "play": 25 years of research on intrinsic versus extrinsic motivation. *Intrinsic and extrinsic motivation* (2000), 257–307.
[38] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
[39] Pedro J. Muñoz Merino, José A. Ruipérez Valiente, and Carlos Delgado Kloos. 2013. Inferring Higher Level Learning Information from Low Level Data for the Khan Academy Platform. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (Leuven, Belgium) *(LAK '13)*. Association for Computing Machinery, New York, NY, USA, 112–116. https://doi.org/10.1145/2460296.2460318
[40] Mark Newman, Irene Kwan, Karen Schucan Bird, and Hui-Teng Hoo. 2021. The Impact of Feedback on Student Attainment: A Systematic Review. *Education Endowment Foundation* (2021).
[41] Martin Nystrand, Adam Gamoran, Robert Kachur, and Catherine Prendergast. 1997. *Opening dialogue*. New York: Teachers College Press.
[42] Martin Nystrand, Lawrence L Wu, Adam Gamoran, Susie Zeiser, and Daniel A Long. 2003. Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse processes* 35, 2 (2003), 135–198.
[43] Mary C O'Connor and Sarah Michaels. 1993. Aligning academic task and participation status through revoicing: Analysis of a classroom discourse strategy. *Anthropology & Education Quarterly* 24, 4 (1993), 318–335.
[44] National Council of Teachers of Mathematics. 2014. *Principles to actions: Ensuring mathematical success for all*. The National Council of Teachers of Mathematics, 35–41.
[45] Alejandro Ortega Arranz, Alejandra Martínez Monés, Juan Ignacio Asensio Pérez, Miguel Luis Bote Lorenzo, et al. 2021. GamiTool: Towards actionable learning analytics using gamification. (2021).
[46] Hayo Reinders and Sorada Wattana. 2012. Talk to me! Games and students' willingness to communicate. In *Digital games in language learning and teaching*. Springer, 156–188.
[47] John TE Richardson. 2005. Instruments for obtaining student feedback: A review of the literature. *Assessment & evaluation in higher education* 30, 4 (2005), 387–415.
[48] Errol Scott Rivera and Claire Louise Palmer Garden. 2021. Gamification for student engagement: a framework. *Journal of Further and Higher Education* 45, 7 (2021), 999–1012.
[49] Richard M Ryan and Edward L Deci. 2009. Promoting self-determined school engagement. *Handbook of motivation at school* (2009), 171–195.
[50] Sitwat Saeed and David Zyngier. 2012. How motivation influences student engagement: A qualitative case study. *Journal of Education and learning* 1, 2 (2012), 252–267.
[51] Danner Schlotterbeck, Pablo Uribe, Roberto Araya, Abelino Jimenez, and Daniela Caballero. 2021. What classroom audio tells about teaching: a cost-effective approach for detection of teaching practices using spectral audio features. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. 132–140.
[52] Klara Sedova, Martin Sedlacek, Roman Svaricek, Martin Majcik, Jana Navratilova, Anna Drexlerova, Jakub Kychler, and Zuzana Salamounova. 2019. Do those who talk more learn more? The relationship between student classroom talk and student achievement. *Learning and instruction* 63 (2019), 101217.
[53] Cynthia Townsend, David Slavit, and Amy Roth McDuffie. 2018. Supporting all learners in productive struggle. *Mathematics teaching in the middle school* 23, 4 (2018), 216–224.
[54] Vicki Trowler. 2010. Student engagement literature review. *The higher education academy* 11, 1 (2010), 1–15.
[55] Lev S Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.
[56] Rose Wang and Dorottya Demszky. 2023. Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, 626–667. https://doi.org/10.18653/v1/2023.bea-1.53
[57] Noreen M Webb, Megan L Franke, Marsha Ing, Jacqueline Wong, Cecilia H Fernandez, Nami Shin, and Angela C Turrou. 2014. Engaging with others' mathematical ideas: Interrelationships among student participation, teachers' instructional practices, and learning. *International Journal of Educational Research* 63 (2014), 79–93.
[58] Gordon Wells. 1999. *Dialogic inquiry: Towards a socio-cultural practice and theory of education*. Cambridge University Press.
[59] Kaylene C Williams and Caroline C Williams. 2011. Five key ingredients for improving student motivation. *Research in Higher Education Journal* 12 (2011), 1.

A TUTOR RE-TRAINING

Coinciding with the experiment[5], CueMath retrained all of its tutors through in-person regional trainings. The training re-established a shared understanding of the core goal of every tutoring session, and of the pedagogical expectations of every tutor. The core goal of each tutoring session is described as maximizing the "delta", i.e. the math skills a student is able to do at the end of a class versus what a student is able to do at the beginning of a class. The training additionally emphasized that productive struggle is critical to maximizing learning [53], and that the three ingredients leading to productive struggle are a) high ratio, b) the right "zone", and c) strong motivation. "High Ratio" refers to ensuring that cognitive work is done by the student, and that they are not passively listening to the tutor explaining a concept for the majority of the class. "Right Zone" refers to ensuring that the content is not too easy, and not too hard for the student, but at the zone of proximal development. "Strong Motivation" refers to ensuring that the tutor maintains an encouraging and positive relationship with the student, so that a student is able to persist moments when productive struggle is challenging. The contents of this training overlapped with themes in the asynchronous training and Zoom training offered to the treatment groups in the experiment (Section 4.1).

B RANDOMIZATION CHECK

Table 5: Randomization check using baseline data. The reason why we do not have N=742 is due to missing data: lack of baseline recordings, self-reported responses or for the language features, only having analyzed a random subset of recordings.

	Control Mean	TutorTM Mean	TutorStudentTM Mean	P Value	N
Tutor female	0.95	0.92	0.95	0.62	727
Tutor age	39.91	39.76	41.56	0.97	734
Tutor CueMath years	3.49	3.41	3.78	0.96	738
Student female	0.57	0.54	0.56	0.19	738
Student grade	4.95	4.74	5.13	0.85	738
Student region=IND	0.09	0.11	0.07	0.71	738
Student region=UK	0.1	0.1	0.11	0.05	738
Student region=US	0.52	0.57	0.59	0.75	738
Student region=ROW	0.29	0.23	0.23	0.77	738
Invalid recording	0.36	0.37	0.39	0.14	738
Baseline Discourse Features					
Student talk ratio	43.52	45.17	42.28	0.73	600
Student talk mins	10.9	10.98	9.73	0.89	600
Teacher talk mins	13.66	12.83	12.14	0.93	600
Num. tutor focusing questions	7.35	6.85	6.24	0.65	301
Num. tutor uptakes	22.07	22.88	19.78	0.7	301
Num. student math terms	34.29	37.55	31.8	0.49	299
Num. tutor math terms	77.26	75.04	66.74	0.61	300
Nu. tutor unique math terms	13.98	14.14	12.93	0.63	300
Num. student unique math terms	8.04	9.04	7.59	0.84	299
Num. student reasoning	1.6	1.49	1.29	0.46	301

[5]The concurrence of the retraining with the experiment was accidental rather than intentional. Since it was offered to all tutors, it did not interfere with the randomization, but it did help ensure that all tutors were aware of the importance of encouraging student participation in the mathematical discourse.

C TABLE 3 WITH ALL COVARIATES

Table 6: Impact of the TalkMeter on student talk ratio and talk time in minutes. Standard errors are in parentheses. + p<0.10 * p<0.05 **. Each column displays the results of a separate regression. The results show a significant increase in student talk, both overall (student talk minutes) and in relation to teacher talk (talk ratio, teacher talk minutes). The key variables pertaining to treatment group assignment are bolded, and all covariates are listed (as described in Section 4.6).

	(1)	(2)	(3)
Independent variable	**Student talk ratio**	**Student talk mins**	**Teacher talk mins**
Group=TutorTM	**5.669****	**0.731****	**-1.744****
	(0.655)	(0.201)	(0.258)
Group=TutorStudentTM	**6.100****	**1.830****	**-0.924****
	(0.676)	(0.232)	(0.271)
Tutor female	-1.051	-0.209	-0.299
	(1.354)	(0.366)	(0.489)
Tutor age	0.062+	0.024*	-0.014
	(0.037)	(0.012)	(0.015)
Tutor CueMath years	0.317+	0.019	-0.092
	(0.188)	(0.061)	(0.071)
Student grade	-0.313*	-0.102*	0.039
	(0.127)	(0.046)	(0.054)
Student female	1.314*	0.173	-0.315
	(0.517)	(0.192)	(0.197)
Student region=IND	-1.035	0.006	0.458
	(0.934)	(0.331)	(0.353)
Student region=US	0.753	0.087	0.214
	(0.674)	(0.243)	(0.245)
Student region=UK	-0.758	-0.188	0.638+
	(1.041)	(0.345)	(0.377)
Baseline student talk ratio	0.637**	0.035+	0.007
	(0.055)	(0.020)	(0.016)
Baseline teacher talk mins	-0.138	0.086*	0.654**
	(0.097)	(0.035)	(0.036)
Baseline student talk mins	0.213	0.684**	0.033
	(0.131)	(0.056)	(0.040)
	(0.427)	(0.157)	(0.126)
Session count	0.024	-0.315**	-0.430**
	(0.054)	(0.020)	(0.030)
Control Mean	43.014	10.093	12.531
R2	0.385	0.461	0.409
Observations	8972	8972	8972