Text Analysis IV: Teacher Learning and RCTs

ISEA Session 11

Jing Liu
University of Maryland
April 11, 2026

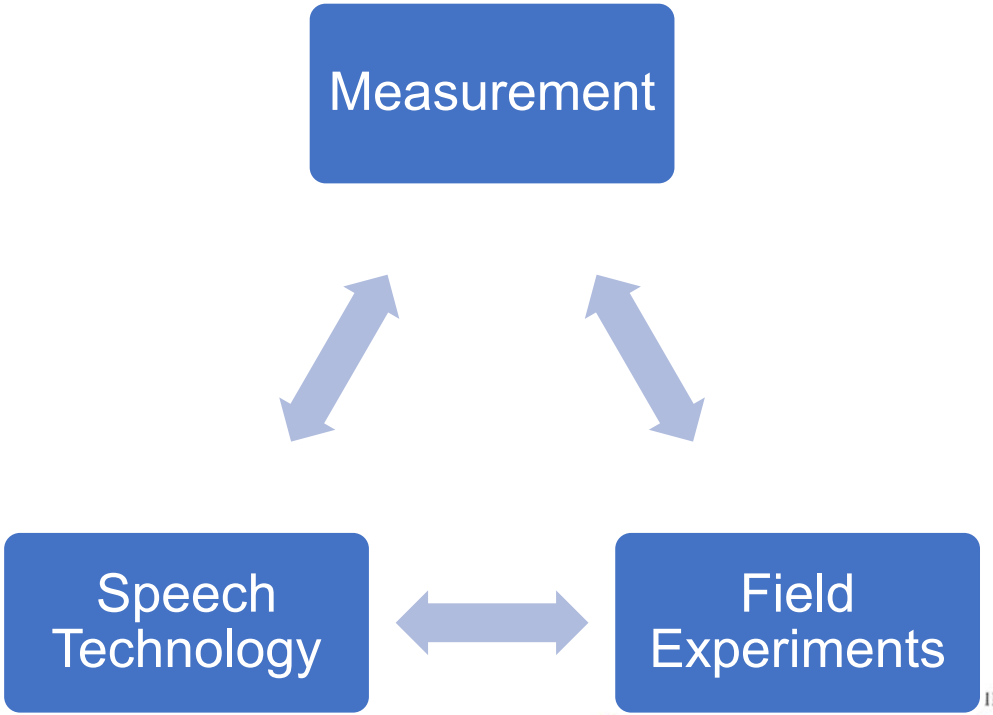
Institute of Education Sciences
AmplifyLearn.AI
University of Washington eScience Institute
Advancing data-intensive discovery in all fields
University of Oregon

Opening Questions

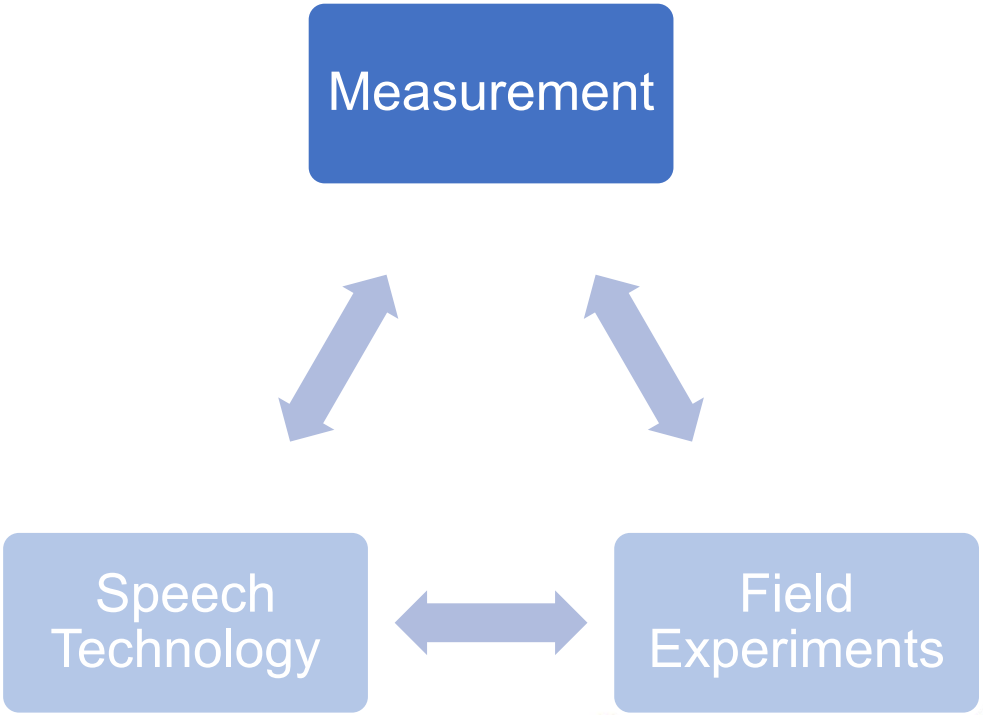
1. What are the unique challenges and opportunities we face for evaluating AI tools/applications in education?

2. What can be some solutions to these challenges you identify?

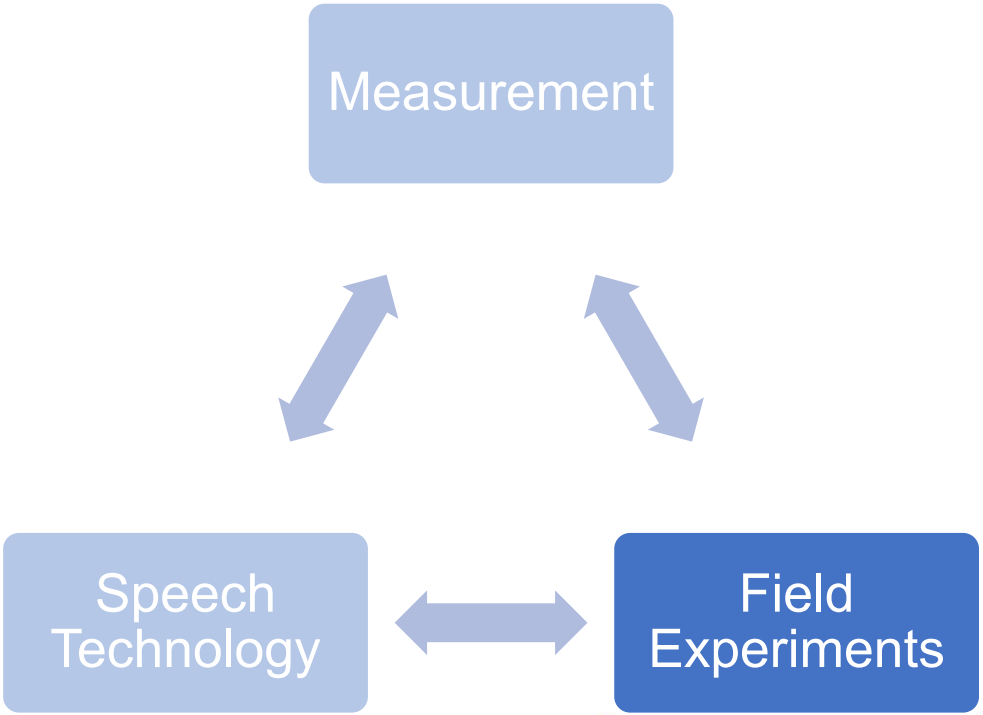
Using NLP to Measure and Improve Teaching: A Framework



Using NLP to Measure and Improve Teaching: A Framework



Using NLP to Measure and Improve Teaching: A Framework



The Importance of Formative Feedback

- Providing teachers with formative feedback can improve both their instruction and their students' outcomes (Taylor & Tyler, 2012; Steinberg & Sartain, 2015; Kraft et al., 2018).

- Formative feedback is nonevaluative, supportive, timely, and specific, with the intention to modify teachers' thinking or behavior to improve their teaching (Shute, 2008).

- Few educators experience such feedback on a regular basis.
  - An average public school teacher only receives formative feedback once or twice per year (Kraft & Gilmour, 2016)
  - Teachers report the feedback they get as low utility (Hellrung & Hartig, 2013)
  - Only 40% of schools provide teachers access to a math or reading coach AND limited coach time on instruction (Taie & Goldring, 2017, Bean et al., 2010; Gibbons & Cobb, 2016; Scott et al., 2012)

Providing Instructors with Automated Feedback: Three RCTs

- Online Computer Science Courses
- Online Tutoring
- Brick-and-Mortar Classrooms

Polygence

TeachFX

IES Education Sciences

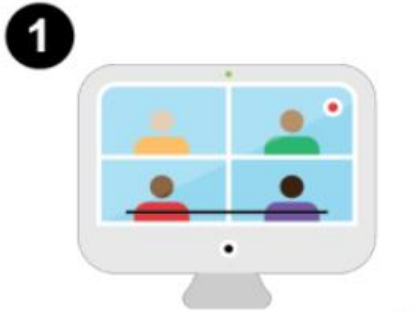
AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

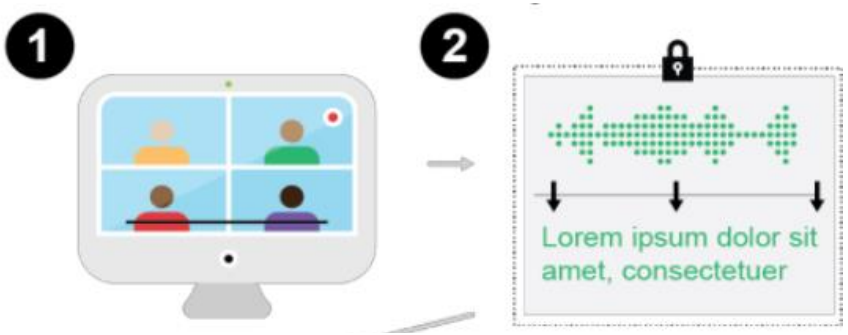
Pipeline of Giving Feedback

Recording



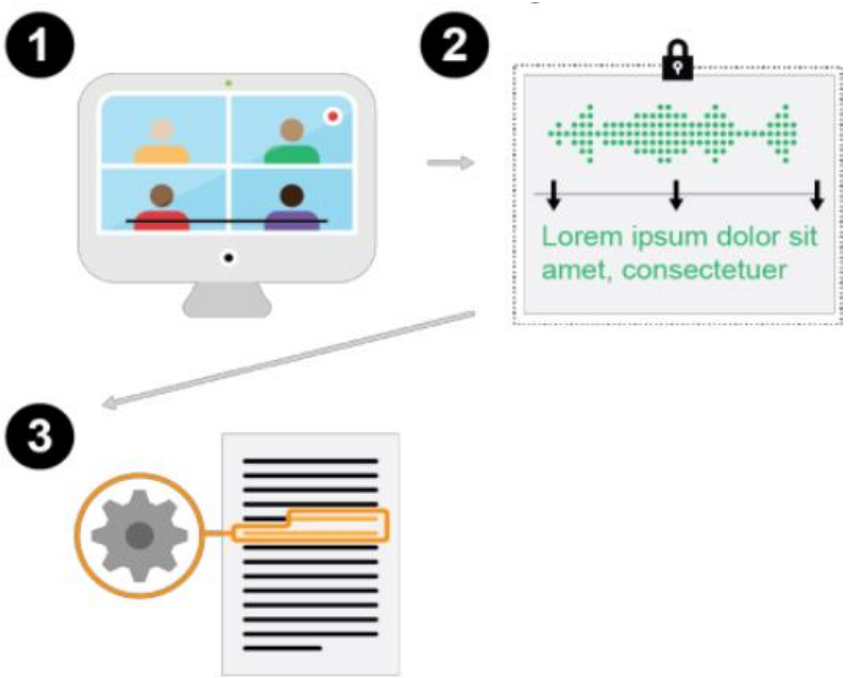
Pipeline of Giving Feedback

Recording

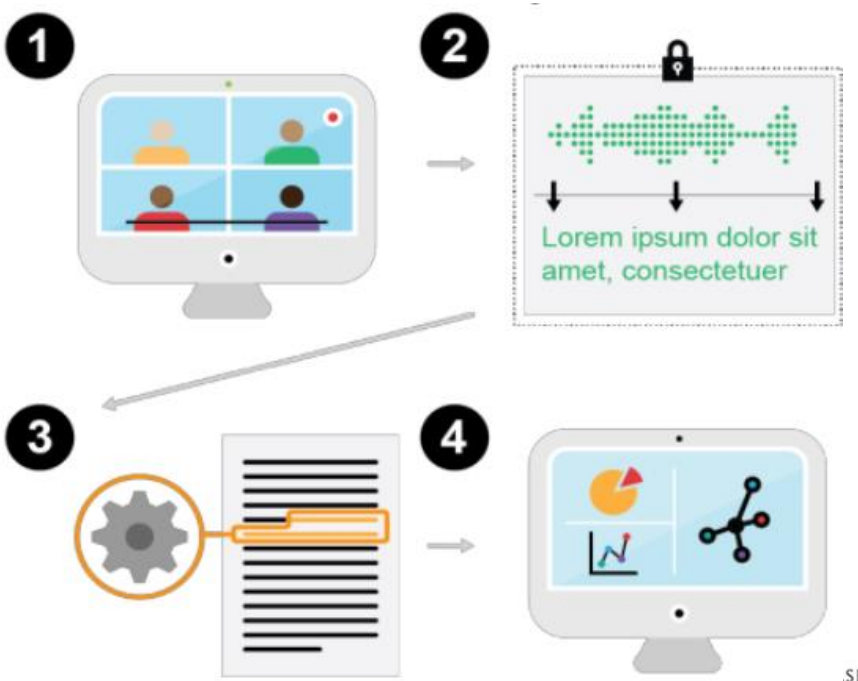


Transcription & Anonymization

Pipeline of Giving Feedback



Pipeline of Giving Feedback



Design principles for reflective feedback

1. Non-judgmental & private

2. Concise, specific & actionable

3. Timely & regular

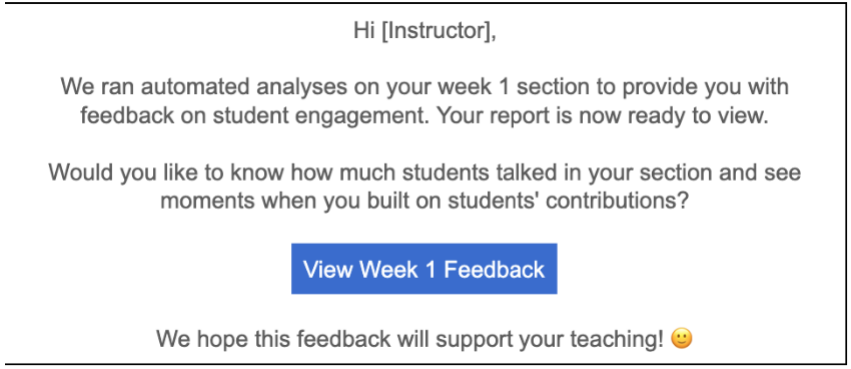
RCT with Code in Place

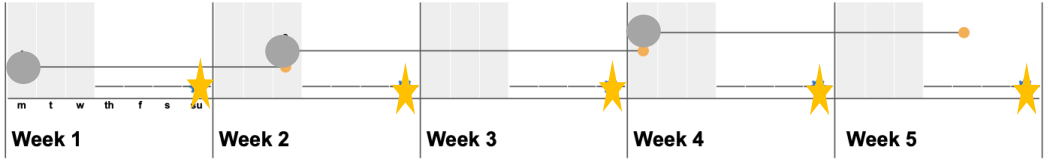
- Code in Place is a five-week free online computer science course organized by Stanford University.
- 12k students + 1.2k section leaders
- Provide automated feedback to instructors on a key teaching practice—**uptake of student contributions**, and evaluate how such feedback affects instruction and student outcomes
- Among the first to evaluate the impact of automated feedback on teacher instruction through a large-scale RCT.

RCT with Code in Place

RCT Design

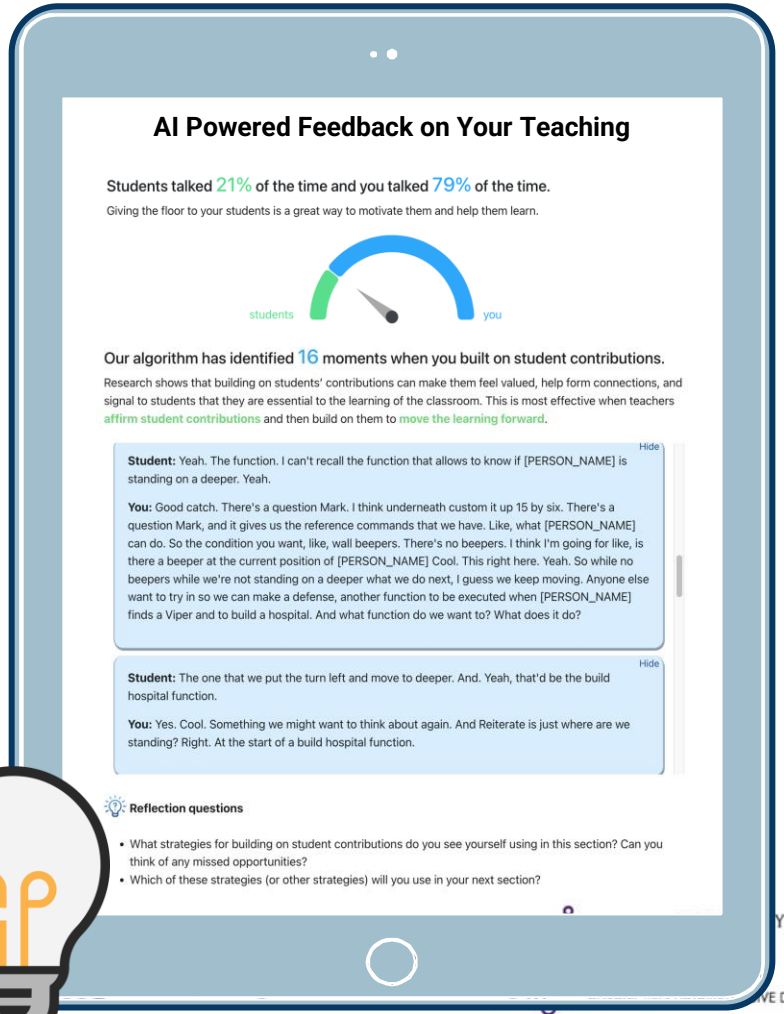
- Randomized encouragement study
- all instructors have access to feedback
- A random 50% of instructors receive ema reminders
- Feedback after each section (5x total)





UNIVERSITY of WASHINGTON eScience Institute ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON


AI Powered Feedback on Your Teaching
Students talked 21% of the time and you talked 79% of the time.
Giving the floor to your students is a great way to motivate them and help them learn.
students
you
Our algorithm has identified 16 moments when you built on student contributions.
Research shows that building on students' contributions can make them feel valued, help form connections, and signal to students that they are essential to the learning of the classroom. This is most effective when teachers affirm student contributions and then build on them to move the learning forward.
Hide
Student: Yeah. The function. I can't recall the function that allows to know if [PERSON_NAME] is standing on a deeper. Yeah.
You: Good catch. There's a question Mark. I think underneath custom it up 15 by six. There's a question Mark, and it gives us the reference commands that we have. Like, what [PERSON_NAME] can do. So the condition you want, like, wall beepers. There's no beepers. I think I'm going for like, is there a beeper at the current position of [PERSON_NAME] Cool. This right here. Yeah. So while no beepers while we're not standing on a deeper what we do next, I guess we keep moving. Anyone else want to try in so we can make a defense, another function to be executed when [PERSON_NAME] finds a Viper and to build a hospital. And what function do we want to? What does it do?
Hide
Student: The one that we put the turn left and move to deeper. And. Yeah, that'd be the build hospital function.
You: Yes. Cool. Something we might want to think about again. And Reiterate is just where are we standing? Right. At the start of a build hospital function.
Reflection questions
What strategies for building on student contributions do you see yourself using in this section? Can you think of any missed opportunities?
Which of these strategies (or other strategies) will you use in your next section?

Research questions

1. Does the feedback improve instructors' practice?
    - Uptake, questions, repetition, and instructors' talk time

2. Does the feedback impact student engagement and satisfaction?
    - Assignment completion
    - Class attendance
    - End-line survey about their perceptions

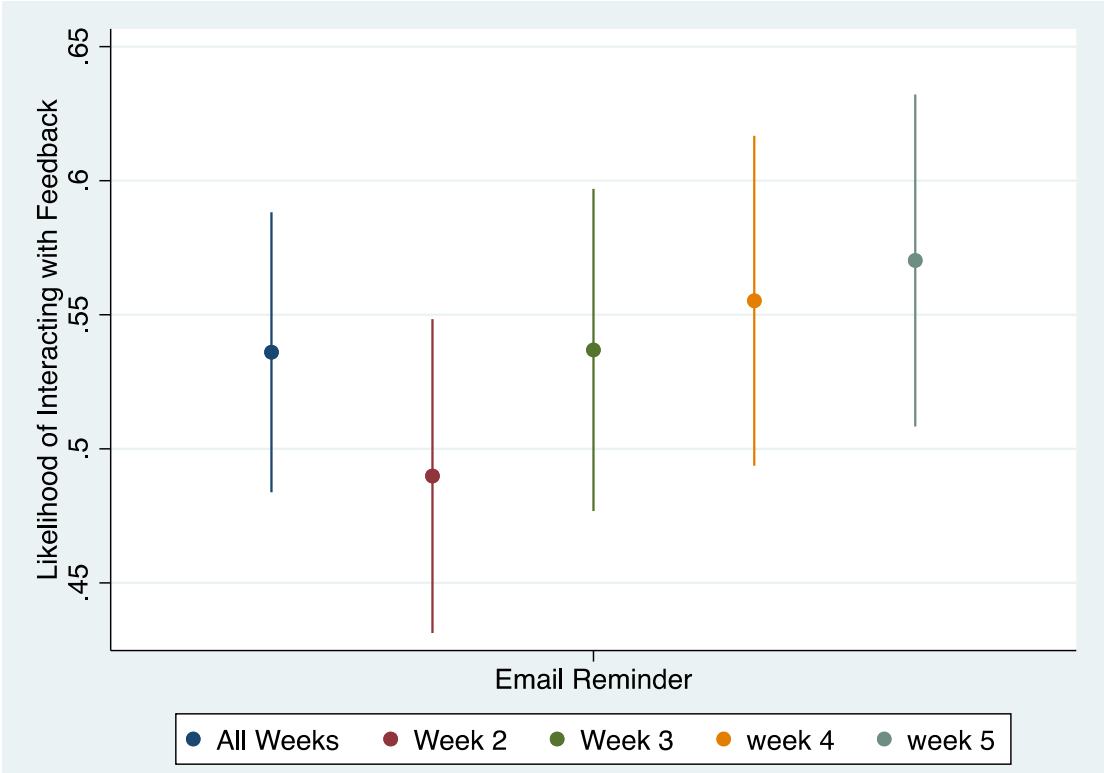
Identification Strategy: 2SLS Estimator

$$Y_{it} = \beta_0 + \beta_1 Feedback_{it} + \beta_2 \boldsymbol{X}_i + \varepsilon_{it}$$

- $i,\ t$ index instructors and a specific instructional week, respectively
- Whether an instructor changed their behavior in week t may be affected by random assignment through
  - whether they checked the feedback in week t
  - whether they checked the feedback in prior weeks
- $Feedback_{it}$ is defined as whether instructor i checked the NLP-based feedback *at least once* prior to the instructor's section in week t
- The email reminder (randomization) serves as an instrument for $Feedback_{it}$
- β_1 measures the impact of *ever interacting* with the automated feedback
- $\boldsymbol{X}_i$ includes student and instructor characteristics and pre-intervention teaching practices (week 1)

First Stages: By Week

Outcome: Whether an instructor ever checked the feedback



First-stage F statistics = 34.151

1. Across all instruction weeks, the email reminder increases treated instructors' likelihood of checking the feedback at least once to **71.2%**, **four times** the rate in the control group (17.6%).

2. The take-up appears to be the **strongest in week 2**, which is after the first email reminder.

Effects of Automated Feedback on Teaching Practices

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

1. Instructors' interaction with the feedback induced by the randomized email reminder improved their use of uptake **by 1.13 times per hour (13.2%)**

2. The improvement in uptake is driven primarily by **more sophisticated strategies** such as increased questioning rather than repetition or talk time.

TOT Effects on Student Outcomes

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

Heterogeneity

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

RCT with Polygence

- Tutoring is a quickly expanding form of instruction, especially by serving as a learning recovery tool post-pandemic
- Polygence: A research mentorship platform for high schoolers
- 1:1 online tutoring mainly offered by Ph.D. students
- N=414 mentors
- Randomly assigned half to receive automated feedback on uptake; the other half has no access to such feedback

Results

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.

1. Treated tutors improved their use of uptake by 0.57 times/hour.
2. The improvement in uptake is driven by both increased questioning and repetitions.
3. We also observe a reduced teacher-student talk ratio.
4. Results are broadly aligned with the Code in Place study.

ITT on Project Outcomes

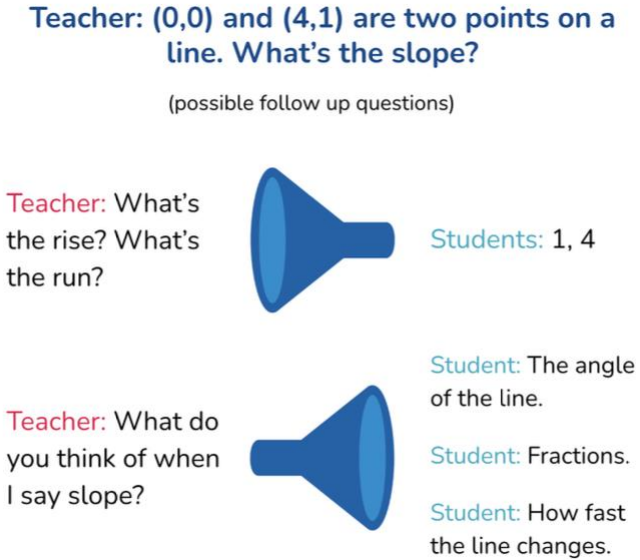
	(1) Mentor NPS	(2) Student NPS	(3) Student Mentor Review Score	(4) Student Optimism About Acad. Future	(5) Published Work
Treatment	0.230+	0.310*	0.020	0.391*	0.013
	(0.124)	(0.129)	(0.028)	(0.152)	(0.025)
Control Mean	9.144	8.093	4.871	8.155	0.107
R2	0.075	0.066	0.088	0.087	0.039
Observations	558	503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

Note: NPS=Net promoter score (On a scale from 0 to 10, how likely are you to recommend this product/company to a friend or colleague?)

RCT with TeachFX in Utah

- First large-scale RCT that tests the efficacy of automated feedback in in-person, K-12 classrooms.
- In partnership with TeachFX, a company that delivers automated feedback to teachers based on classroom recordings via a phone application.
- N=523 math or science teachers teaching in Utah public schools.
- All teachers have access to TeachFX's feedback, but half of them are randomly assigned to additional weekly feedback on **focusing questions** through email.
- Collected both quantitative data and rich interview data to understand teachers' perceptions of the tool.

Identifying Focusing Questions



- Binary classification machine learning model
- Fine-tuning Bert based on labeled data from the NCTE dataset, augmented by 694 annotated examples from TeachFX
- 84% accuracy on a held-out set from TeachFX (Alic et al., 2022)

Treatment

- Teachers enrolled in the study on a rolling basis and then got randomized assigned to the treatment or control group
- An email early every Tuesday morning which contained both the number of focusing questions they asked in all class recordings in the previous week as well as a display of, at most, the top 3 chosen focusing questions
- Top questions are identified by two math instructional coaches
- The email also contains a link to the focusing question insights on the TeachFX app
- Terminated treatment after 5 weeks of recordings for a given teacher

Hi Anthony,

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

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

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

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

Click here to hear how your students responded

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

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

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

Reflect

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

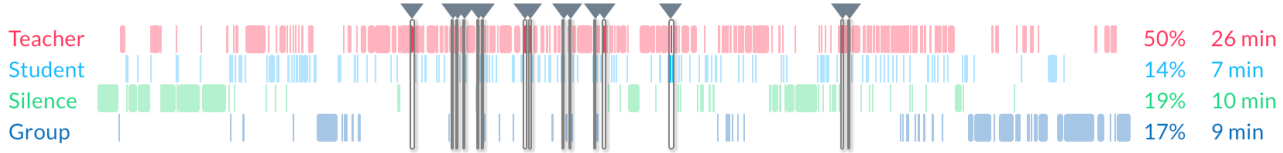
A1 Correlation

Thu | March 2 | 7:41 AM | 51 min

share delete

Listen back to these questions.

Here are 15 focusing questions you asked in this lesson.



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

Reflect:

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

Research Questions

- To what extent do K-12 teachers engage with the automated feedback on focusing questions?
- Does the automated feedback on focusing questions impact instruction, including teachers' use of focusing questions, student talk time, and student reasoning?
- How do teachers perceive the automated feedback on both focusing questions and other teaching practices? What are the barriers for them to engage with the feedback?

Descriptive Statistics

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

RQ1. Engagement with the feedback email and the TeachFX platform is limited

- We tracked email opens and views of the focusing question insight on the TeachFX platform
- Treated teachers opened the email Between 55-61% of teachers opened their emails across weeks, but only 17-22% of them viewed the focusing insight page.
- On average, teachers opened 1.8 emails (SD=1.9) out of 5 throughout the RCT.
- The intervention increases views of the TeachFX platform for the treated (21% of the time) vs. control teachers (15% of the time) ($p<0.05$)

RQ2. The treatment improved teachers' use of focusing questions, but not other related teaching practices or student engagement.

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

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

RQ3. Many barriers prevent teachers from fully engaging with the feedback

1. Low recording and transcription quality

"It has some really obvious flaws in the recording. And so a lot of us are like, 'Oh, I did not say that.' . . . I know that that's a hang-up for a lot of teachers."

2. Time constraint

"I think for me the hard [thing] is like didn't have time to sit and read it when it would come in, and then I would forget about it."

3. Concerns about data privacy issues involved in automated feedback

"Nobody likes listening to themselves and being observed and things like that, so like finding ways to be able to share things that we're happy about without feeling like... I don't know, like you're going to be criticized."

Conclusions

- Automated feedback to teachers using NLP-based measures shows promising effects in improving a key teaching practice in in-person teaching contexts, but failed to generate tangible effects on related teaching practices
- Many barriers prevent teachers from fully engaging with the feedback
- Data availability and logistics constraints prevent us from conducting more in-depth analysis on how the intervention works for different groups of teachers and in what contexts
- We are integrating coaching routines with automated feedback to enhance the take-up and effectiveness of our approach
- New RCTs are in the pipeline

Discussion

1. What might be some other applications of AI-based instructional measurement you can think of other than giving teachers feedback?

2. How do you plan to evaluate whether such application works or not?

Assignment

Based on what we introduced today, make a plan for a data science-powered RCT you would like to run in education. You need to think through and describe the following elements in 1-2 page memo:

- **The underlying theory of change**
- **Target population and sample**
- **Intervention design**
- **Outcomes you would like to measure**
- **Implementation**
- **Analytic approach**