AI for Instruction Measurement and Feedback

AERA Virtual Research Learning Series

Jing Liu
University of Maryland
June 11, 2024

IES Institute of Education Sciences

AmplifyLearn.AI

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

UNIVERSITY OF OREGON

Overview of today

1. Why measuring instruction?
2. Workflow of using natural language processing (NLP) to measure instruction
3. Case study: measuring the uptake of student ideas
4. Field experiments

The Measurement of Effective Teaching Is Fundamental to Any Educational Improvement Efforts!

IES Institute of Education Sciences

AmplifyLearn.AI

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

UNIVERSITY OF OREGON

The Current System of Human Observation and Feedback

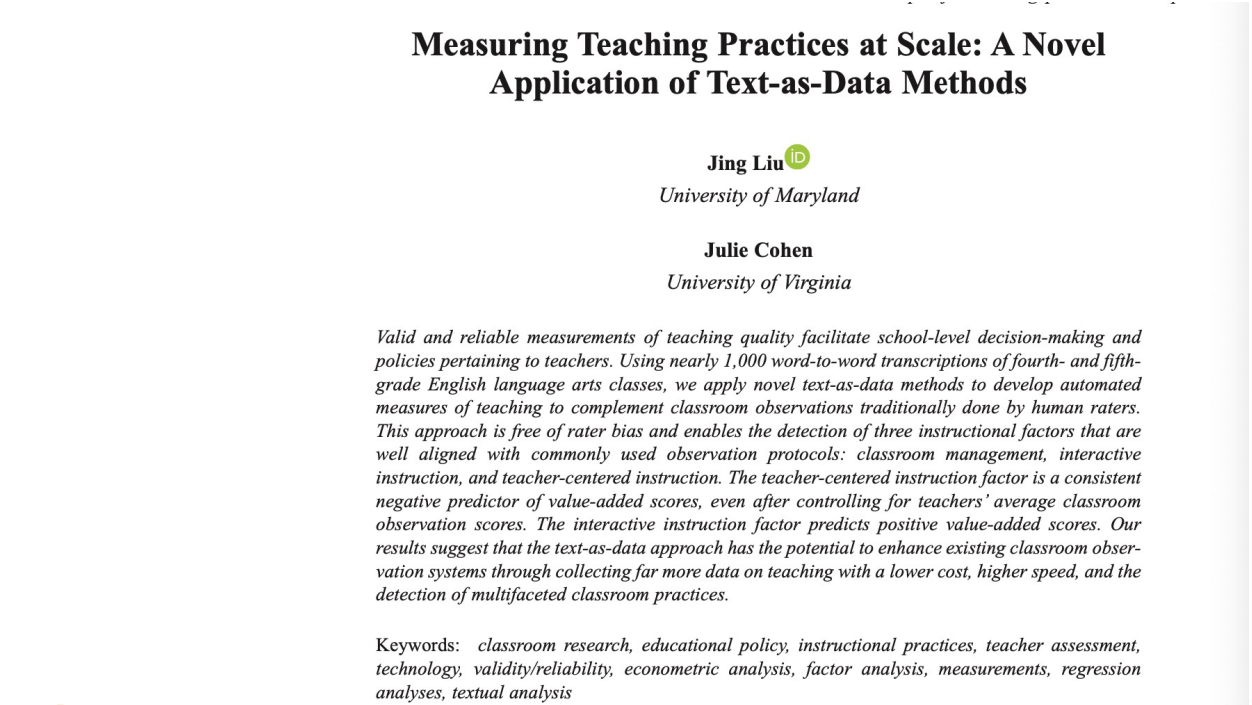
- Widely used in the US and the world to evaluate teaching practices across early childhood, K-12, and higher education (Kane & Staiger, 2012; Pianta & Hamre, 2009; Cohen & Goldhaber, 2016; Hill & Grossman, 2013)
- Resource intensive: an average public school teacher only receives formative feedback once or twice per year (Kraft & Gilmour, 2016)
- The quality of feedback varies: low rater consistency & prone to bias (Ho & Kane, 2013; Donaldson & Woulfin, 2018; Kraft & Gilmour, 2016)

IES Institute of Education Sciences

AmplifyLearn.AI

UNIV eScien ADVANCING DATA-IN

NLP Techniques Provides A Powerful Alternative to Human Observation



Measuring Teaching Practices at Scale: A Novel Application of Text-as-Data Methods

Jing Liu
University of Maryland

Julie Cohen
University of Virginia

Valid and reliable measurements of teaching quality facilitate school-level decision-making and policies pertaining to teachers. Using nearly 1,000 word-to-word transcriptions of fourth- and fifth-grade English language arts classes, we apply novel text-as-data methods to develop automated measures of teaching to complement classroom observations traditionally done by human raters. This approach is free of rater bias and enables the detection of three instructional factors that are well aligned with commonly used observation protocols: classroom management, interactive instruction, and teacher-centered instruction. The teacher-centered instruction factor is a consistent negative predictor of value-added scores, even after controlling for teachers' average classroom observation scores. The interactive instruction factor predicts positive value-added scores. Our results suggest that the text-as-data approach has the potential to enhance existing classroom observation systems through collecting far more data on teaching with a lower cost, higher speed, and the detection of multifaceted classroom practices.

Keywords: *classroom research, educational policy, instructional practices, teacher assessment, technology, validity/reliability, econometric analysis, factor analysis, measurements, regression analyses, textual analysis*

Liu & Cohen (2021)

IES Institute of Education Sciences

AmplifyLearn.AI

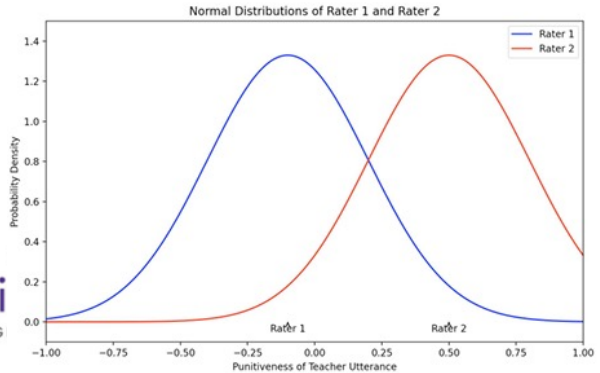
of WASHINGTON eScience Institute ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

NLP Measure Development Workflow

Annotation

- **Conduct high-quality annotation for model training and validation**
  - Actual sample size for annotation varies based on the nature of the measure and the "unit" of samples (i.e., sentences, paragraphs, chapters, etc)
  - Rule of thumb: 1K for discrete, low-inference measures; 2K for high-inference ones
  - Regardless of NLP model choice, you need a validation set
- **Achieving high interrater agreement is critical**
  - When possible, having multiple coders who have domain knowledge
  - Iteratively refine definition of a construct and coding scheme
  - Check the distribution of scoring for raters



Supervised vs. Unsupervised Modeling

Supervised models	Unsupervised models
Pros: ● Tends to perform better when sufficient labeled training data is available	Pros: ● Does not need labeled data for training ● Tends to transfer better across domains
Cons: ● Model performance tends to correlate directly with amount of labeled data, which in turn is expensive to collect ● Performance often generalizes less across domains	Cons: ● Not available / gets complicated for many high-inference constructs

Supervised modeling: LLMs or smaller models?

Smaller models (RoBERTa, BERT, etc.)	LLMs
Resources: https://simpletransformers.ai/; https://huggingface.co/docs/transformers/index	GPT-3.5; Llama 2; GPT-4 (instruct tuning)
Pros: ● Downloadable → more transparency & control ● Needs little compute ● Can achieve similar performance to LLMs when sufficient labeled data is available	Pros: ● Very good at few shot learning ● Can be tuned with instructions
Cons: ● Require more training data ● Can't be tuned with instructions or via interacting with the model	Cons: ● Most cannot be downloaded ● Many models can't be finetuned (e.g. GPT-4, Claude)

What Instructional Practices to Measure?

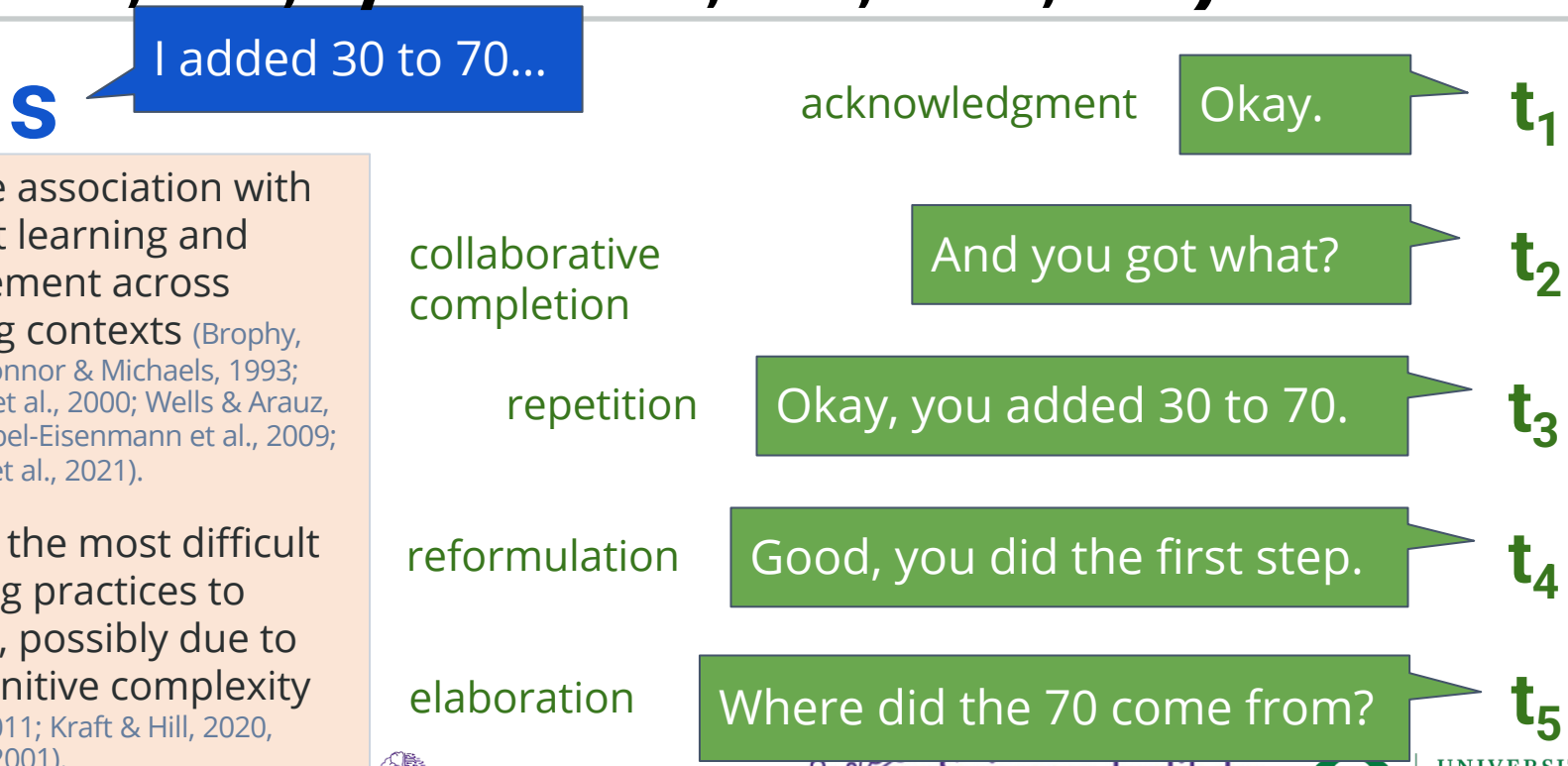
Starting with popular classroom observation tools!

Observation instrument	Developed by	Type of classes served
Classroom Assessment Scoring System	University of Virginia	English language arts and math
Framework for Teaching	Charlotte Danielson	English language arts and math
Protocol for Language Arts Teaching Observations	Stanford University	English language arts
Mathematical Quality of Instruction	University of Michigan	Math
UTeach Observational Protocol	University of Texas–Austin	Math

Kane & Staiger, 2012

What is Uptake?

(Collins, 1982; Nystrand et al., 1997; Wells, 1999).



- Positive association with student learning and achievement across learning contexts (Brophy, 1984; O'Connor & Michaels, 1993; Nystrand et al., 2000; Wells & Arauz, 2006; Herbel-Eisenmann et al., 2009; Demszky et al., 2021).
- Among the most difficult teaching practices to change, possibly due to the cognitive complexity (Cohen, 2011; Kraft & Hill, 2020, Lampert, 2001).

Data Source

- 4th and 5th grade elementary math classroom transcripts collected by the National Center for Teacher Effectiveness (NCTE) between 2010-2013 (Kane et al., 2015)
- 317 teachers
- 4 school districts in New England serving largely low-income, historically marginalized students
- Transcripts are anonymized

Annotation

- 3 raters / example with 13 raters who have prior experience with teaching/coaching
- Raters were given extensive training, and documentation w/ examples
- In the annotation interface, raters were presented with an (S, T) pair and asked
  - Does (S, T) relate to math?
    - (e.g. “Can I go to the bathroom?” is not related to math)
  - If both (S, T) relate to math, they were asked to rate T for “low”, “mid” or “high” uptake

Example	Label
S: ’Cause you took away 10 and 70 minus 10 is 60. T: Why did we take away 10?	high
S: There’s not enough seeds. T: There’s not enough seeds. How do you know right away that 128 or 132 or whatever it was you got doesn’t make sense?	high

Example	Label
S: ’Cause you took away 10 and 70 minus 10 is 60. T: Why did we take away 10?	high
S: There’s not enough seeds. T: There’s not enough seeds. How do you know right away that 128 or 132 or whatever it was you got doesn’t make sense?	high
S: Teacher L, can you change your dimensions like 3-D and stuff for your bars? T: You can do 2-D or 3-D, yes. I already said that.	mid
S: The higher the number, the smaller it is. T: You got it. That’s a good thought.	mid

Example	Label
S: ’Cause you took away 10 and 70 minus 10 is 60. T: Why did we take away 10?	high
S: There’s not enough seeds. T: There’s not enough seeds. How do you know right away that 128 or 132 or whatever it was you got doesn’t make sense?	high
S: Teacher L, can you change your dimensions like 3-D and stuff for your bars? T: You can do 2-D or 3-D, yes. I already said that.	mid
S: The higher the number, the smaller it is. T: You got it. That’s a good thought.	mid
S: An obtuse angle is more than 90 degrees. T: Why don’t we put our pencils down and just do some brainstorming, and then we’ll go back through it?	low
S: Because the base of it is a hexagon. T: Student K?	low

Use NLP to measure uptake

Next utterance classification
~ **Pointwise Jensen Shannon Divergence (PJSD)**

$$pJSD(t,s) := -\frac{1}{2}\Big(\log P(Z=1|M=t,s) + \mathbb{E}\log(1-P(Z=1|M=T',s))\Big) + \log(2)$$

where ***(S, T)*** is a teacher-student utterance pair, ***T'*** is a randomly sampled teacher utterance and $M := ZT + (1-Z)T'$ is a mixture of the two with a binary indicator variable ***Z*** **~ Bern(p=0.5)**.

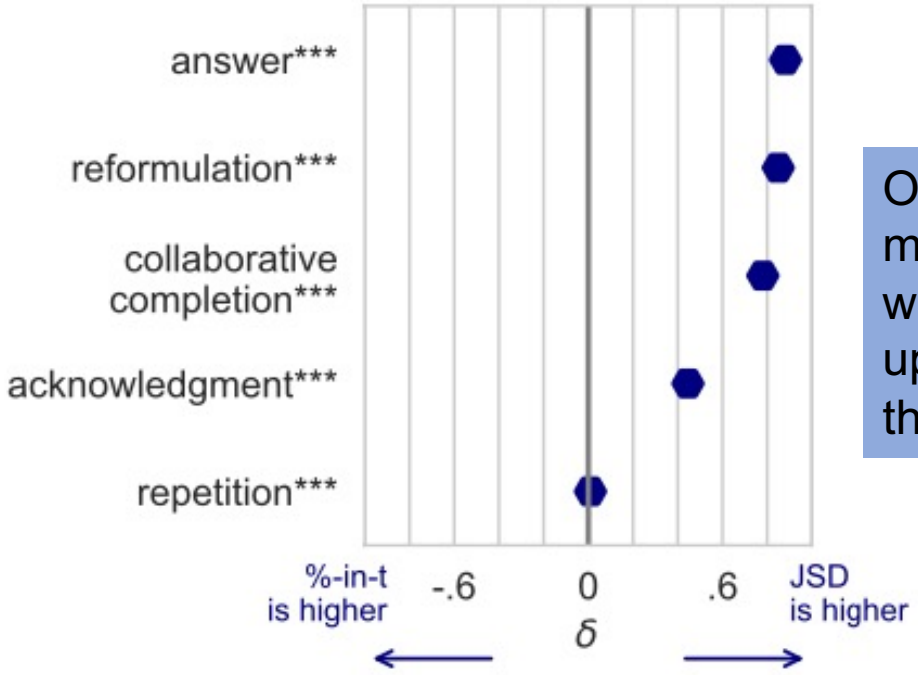
Validation Methods

- Comparison to expert annotation
- Linguistic analysis
- External validation

Validation #1: Comparison to expert labels

Model	Correlation with annotation
Sentence-Bert	0.390
Glove	0424
%-IN-S	0.449
Universal Sentence Encoder	0.448
Jaccard	0.450
BLEU	0.510
%-IN-T	0.523***
Our Uptake Measure	0.540***

Validation #2: Qualitative comparison via speech acts (Switchboard corpus)



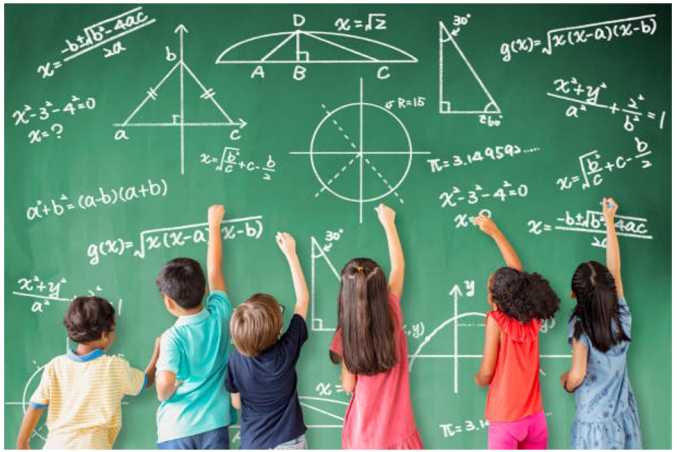
Our unsupervised method captures a wider range of uptake strategies than %-in-T.

Validation #3: Correlation with external measurements

- Obtain datasets with transcript-level external measurements
  - classroom observation scores
  - student satisfaction scores
- Generate aggregate uptake score for each transcript
- Correlate aggregate uptake score with external measurements

External Validation #1:

NCTE dataset [Kane et al., 2015]



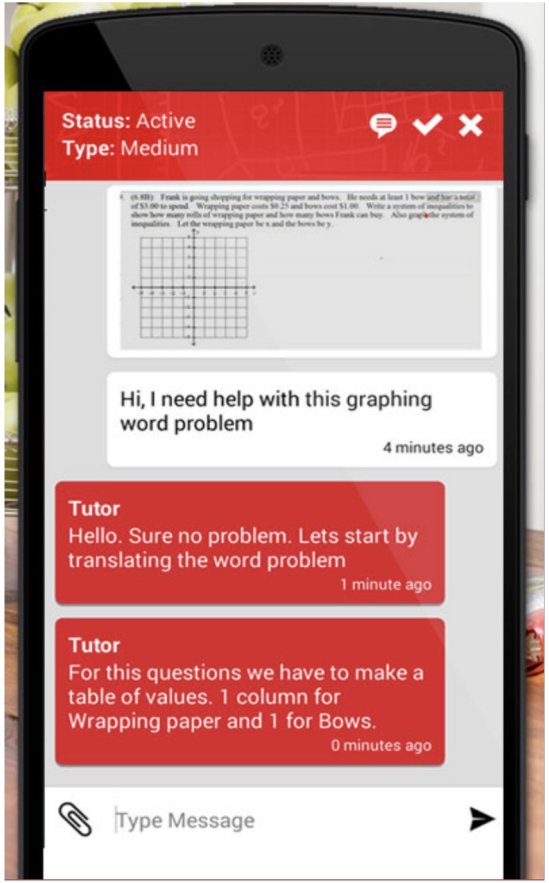
- N=55k (S, T) pairs
- elementary math classrooms
- spoken (in-person)
- whole class (20-30 students)
- external measures:
  - use of student contributions
    - **β=0.101*****
  - math instruction quality
    - **β=.091*****

OLS coefficients, *** p < 0.001

External Validation #2:

Tutoring dataset

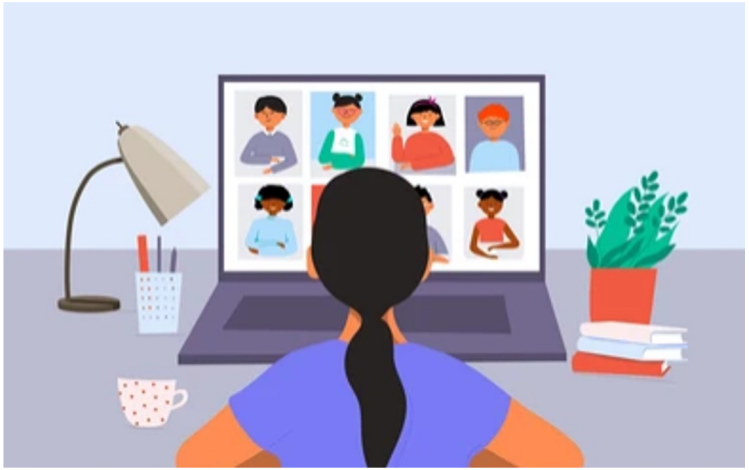
- N=85k (S, T) pairs
- math and science
- written (texts through app)
- 1:1
- outcomes:
  - external reviewer rating
    - **β=0.063*****
  - student satisfaction
    - **β=0.069*****



External Validation #3:

SimTeacher [Cohen et al., 2020]

- **not part of training data!**
- N=2.7k (S, T) pairs
- elementary literacy
- spoken (virtual)
- small group (5 students)
- outcomes:
  - quality of feedback
    - **β=.127***



Going Beyond Teachers' Uptake of Student Ideas

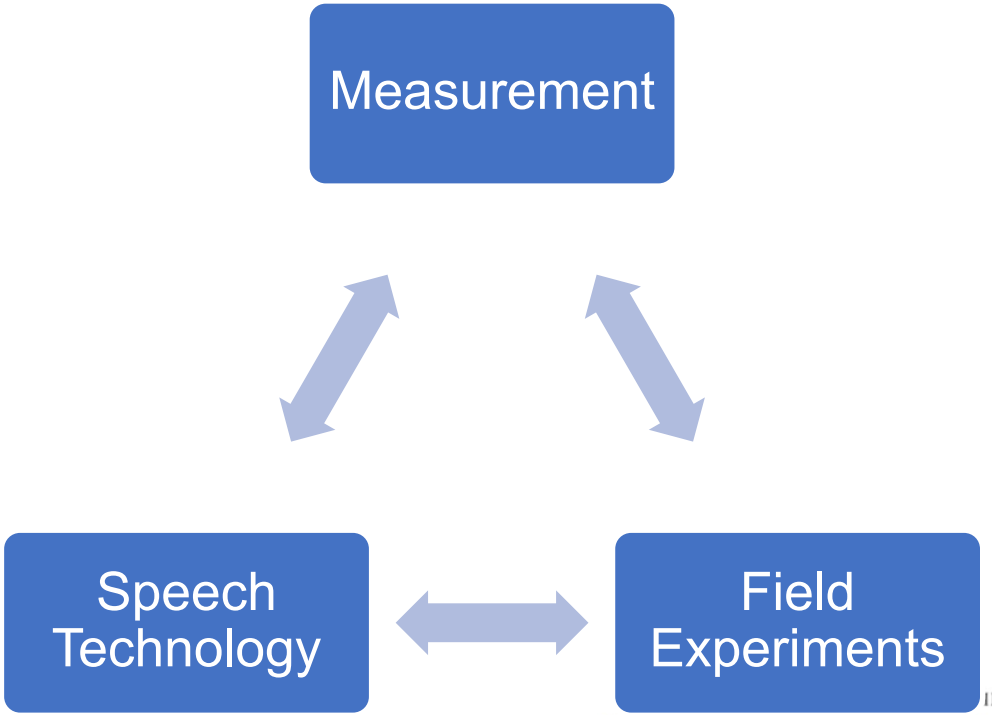
- Mathematical language (both teacher and student)
- Teacher focusing (open-ended) questions
- Student mathematical explanation and reasoning
- Classroom management and time on task
- Meta-cognitive modeling
- ……

The Promises and Pitfalls of Using Language Models to Measure Instruction Quality in Education (Xu, Liu et al., 2024)

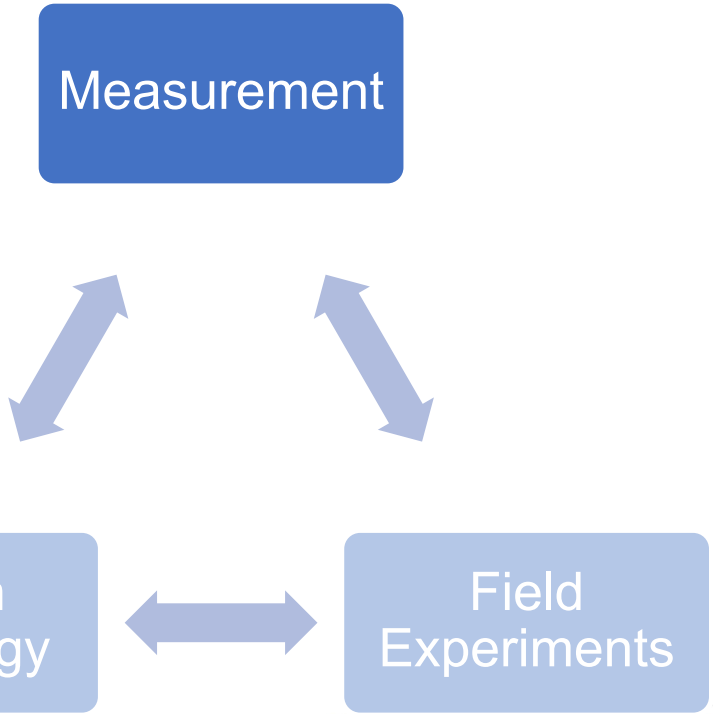
- Tackle two common challenges with using NLP to measure teaching
  - Very imbalanced distribution of scoring (lack of high-rating examples)
  - Long input, especially for high-inference teaching practices

- "Our results suggest that pretrained Language Models (PLMs) demonstrate performances comparable to the agreement level of human raters for variables that are more discrete and require lower inference, but their efficacy diminishes with more complex teaching practices that require further inferences."

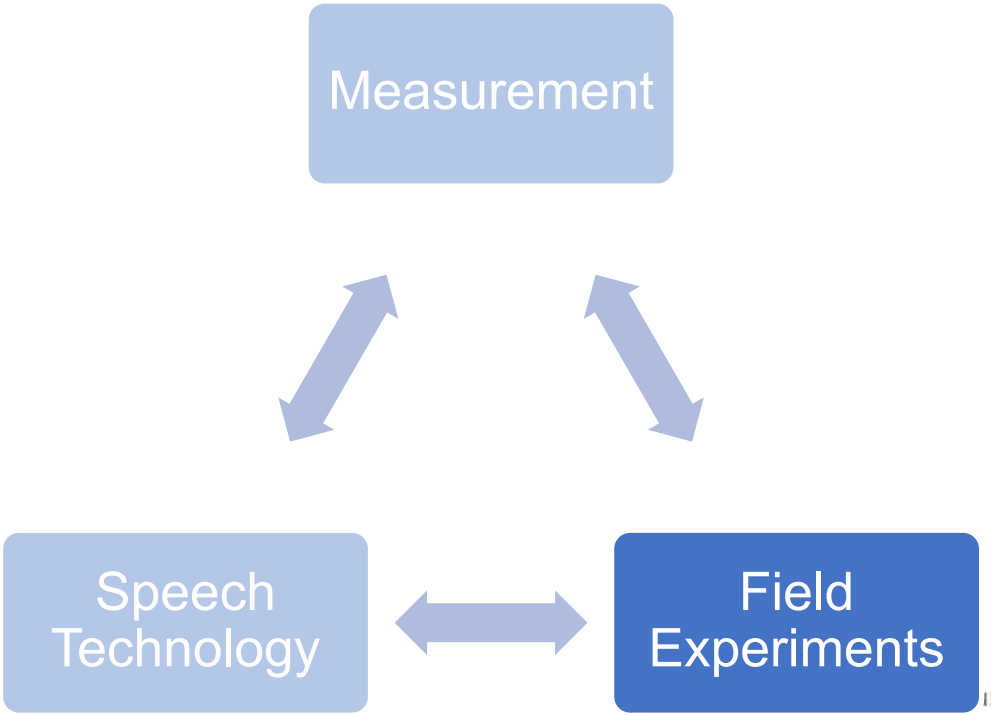
Using NLP to Measure and Improve Teaching: A Framework



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

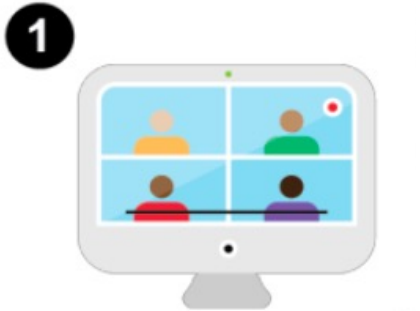
AmplifyLearn.AI

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

UNIVERSITY OF OREGON

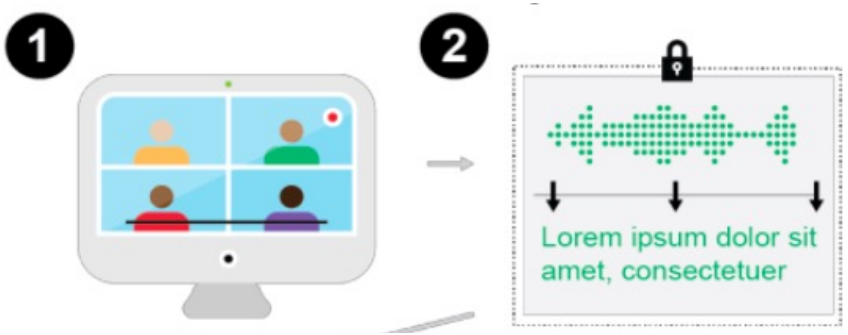
Pipeline of Giving Feedback

Recording



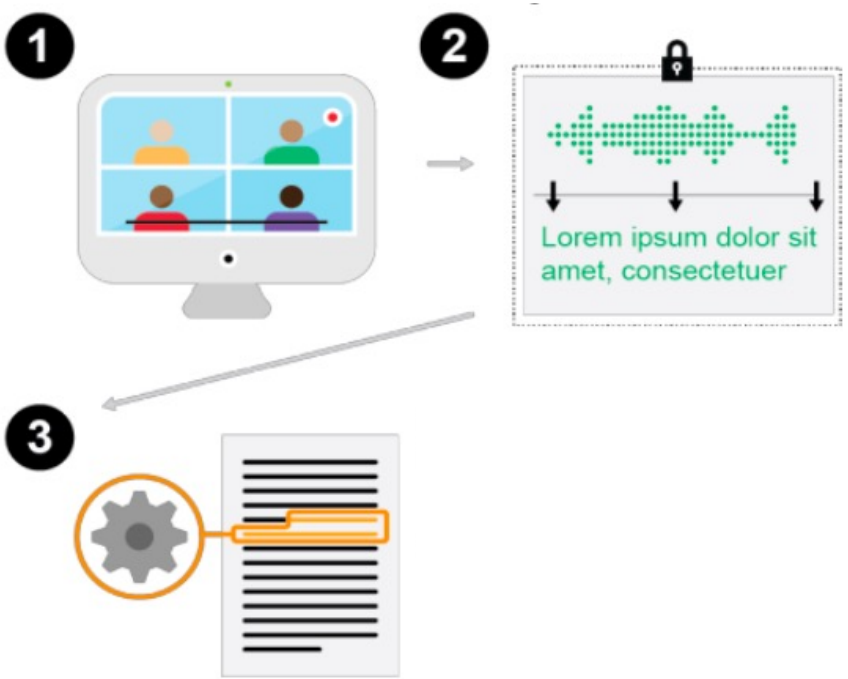
Pipeline of Giving Feedback

Recording

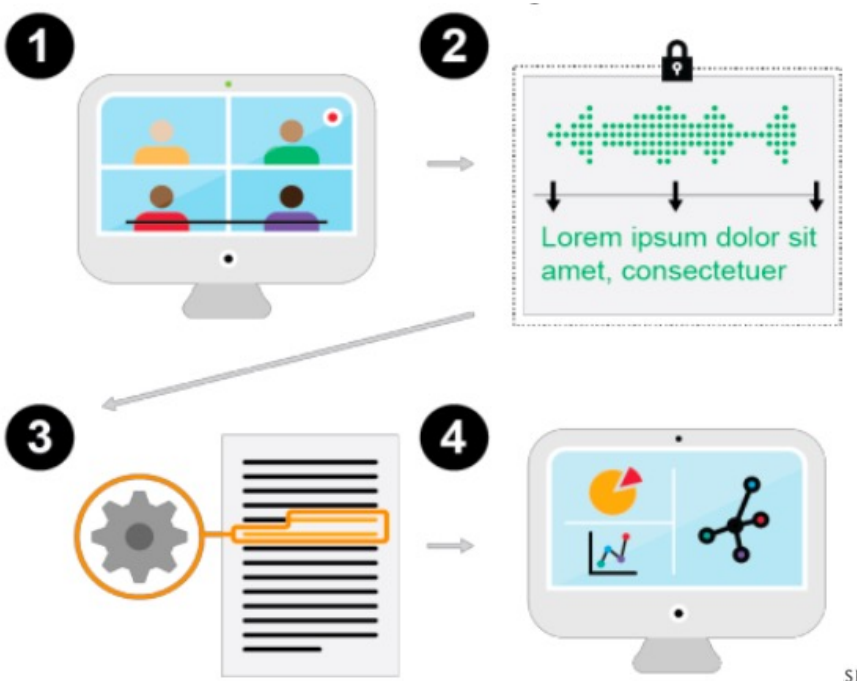


Transcription & Anonymization

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.

IES Institute of Education Sciences

AmplifyLearn.AI

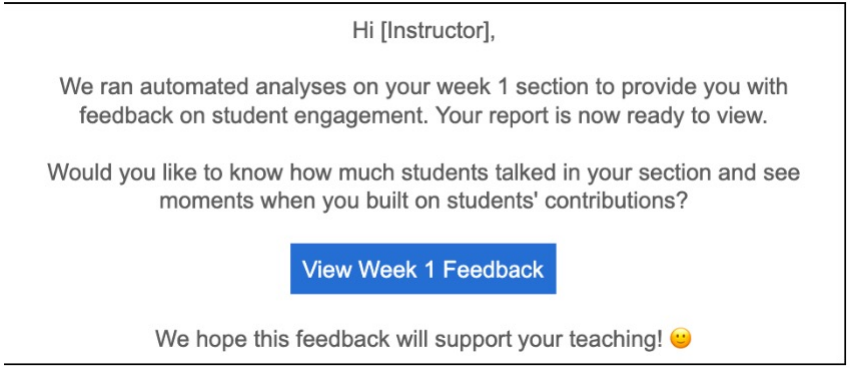
UNIVERSITY of WASHINGTON eScience Institute ADVANCING DATA-INTENSIVE DISCOVERY ACROSS FIELDS

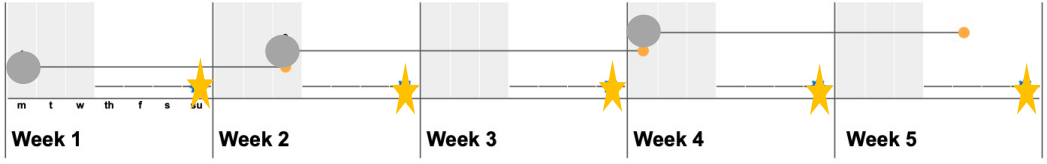
UNIVERSITY OF OREGON

RCT with Code in Place

RCT Design

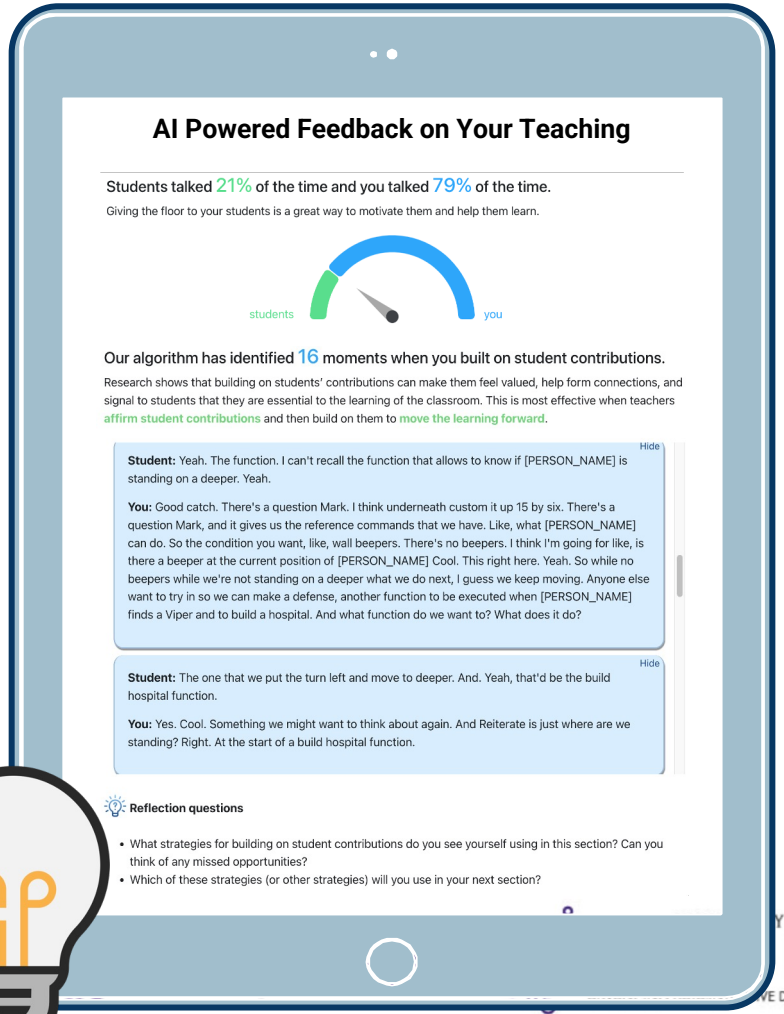
- Randomized encouragement study
- all instructors have access to feedback
- A random 50% of instructors receive ema reminders
- Feedback after each section (5x total)





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

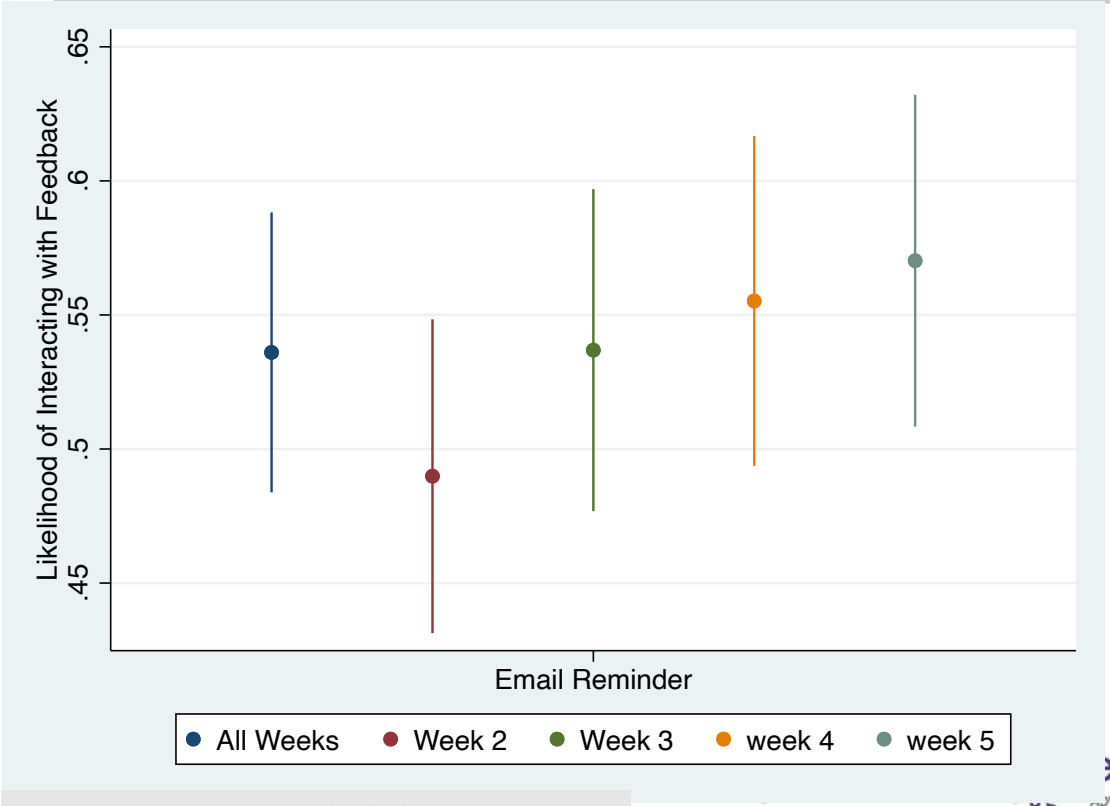
Identification Strategy: 2SLS Estimator

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

- $i,\ t$ 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