CS 293/EDUC 473

Deploying & Evaluating the Effectiveness of NLP-Powered Tools

Stanford
GRADUATE SCHOOL OF
EDUCATION

Reminders & announcements

- HW3 due tonight at midnight
- Practice pitch round 2 on Wednesday
  - **last chance** to receive feedback from the class
  - will be split into **2 parallel sessions – announced soon!**
  - **peer reviews will be reassigned according to sessions**
  - please stick to **4 mins**

Teacher Review Panel

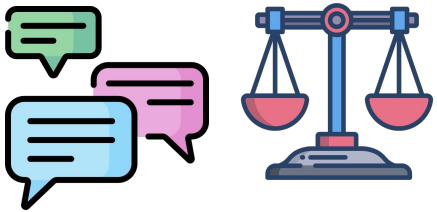
Jesus Rojas	I teach 6th grade science in Menlo Park and I am actively looking for ways to integrate more technology into my curriculum design and instruction. I am looking forward to collaborating and seeing the broad ideas that will be shared in this project.
Kavitha Satya-Mohandoss	I teach Algebra1 and 2 at USC East College Prep, LA. I am always looking for ways to make math learning engaging and enjoyable. I am also Chair of Outreach for the LACounty Science and Engineering Fair. My mission is to inculcate the value of two most important attributes exclusive to our planet- time and the human connection, in my students, and others around me. I am excited to learn from your students.
Hanna Crowe	I am a high school math teacher in Los Angeles. When I was in college I worked at the ITS Help Desk troubleshooting technology for other teachers. This showed me the impact of using technology in class, because I was able to see what happened when tech went wrong. I also felt like I was on the cutting edge of classroom technology, because we were constantly rolling out new initiatives for educators to try.
Nicole Elenz-Martin	I have served in the San Mateo Union High School District for the past 18 years as a Spanish Teacher, AVID Teacher, Instructional Technology Coordinator, Instructional Coach, and most recently as a site Administrator (Assistant Principal overseeing Curriculum and Instruction, as well as Technology). I am currently coaching and mentoring Elementary Administrators in the Bay Area as well, so I am seeing and accessing elementary school classrooms and curriculum too. Lastly, my own two children are in middle school, so I often see their access to and am very involved in their curricular areas. I am passionate about AI making learning more robust and exciting, as opposed to "making us less smart and more dependent", of course, and would love to see what you have to share from the teacher, mentor, administrator, and parent perspective!

Teacher Review Panel

Taylor Pacheco	I am an Algebra 1 teacher at Pueblo High School in the Tucson Unified School District (Arizona). I began STEP in person in 2019 and graduated virtually from STEP in 2020. I began my teaching career online and transitioned in person in 2021. I use AI to brainstorm lesson plans, worksheets, find the right words for an email, etc. and encourage students to use AI to help themselves get "unstuck".
Rahim Strong	Science has always been a love of mine since grade school. It's not surprising that after many years in the marine conservation field, I became a science teacher. Currently I teach at Downtown Charter Academy in Oakland CA. I have always been one for technology in the classroom, employing new tools as they become available. My teaching experiences have taken me from some of the poorest areas in California, to the Ultra-Wealthy students of MiSK schools in Saudi Arabia.
Sergio Estrada	I am an instructional coach. I taught secondary science for 8 years in El Paso, Texas. I love integrating technology in my teaching to be more efficient. I am always looking for technology that will make my teaching easier without sacrificing rigor or emotinoal support that I provide. For example, I do not like Edpuzzle as I am not there to truly check for understanding at a deep level. I have used Swivl to record my classes and reflect, I have also used TeachFX to help others coach. I am weary of using technology to teach as I am not sure we are at a place where it can provide the emotional support students needs.

1

Measure an educationally important discourse phenomenon



2

Validate the measure using existing data

3

Deploy the measure to give teachers feedback

Recap: Developed & validated an unsupervised measure for uptake using secondary data.

Case study for today's class:

1

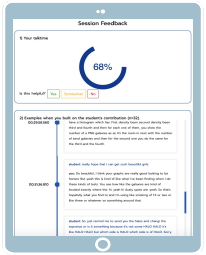
Measure an educationally important discourse phenomenon

2

Validate the measure using existing data

3

Deploy the measure to give teachers feedback



Steps to running an experiment

1. Set up the backend (i.e. NLP pipeline)
2. Develop the frontend with users
3. Test the end-to-end tool with users
4. Figure out the experiment setup
    a. Who are the participants? What are the conditions? What **quantitative** data will you collect as outcomes, covariates, etc.? Can you also collect **qualitative** data?
5. Run the experiment
    a. Constantly monitor, because there will be bugs
6. Analyze collected data
    a. Pre-registration highly encouraged!
7. Report & disseminate results

Steps to running an experiment

1. **Set up the backend (i.e. NLP pipeline)**

Behind the scenes

1. Record sections

2. Transcribe & anonymize recording

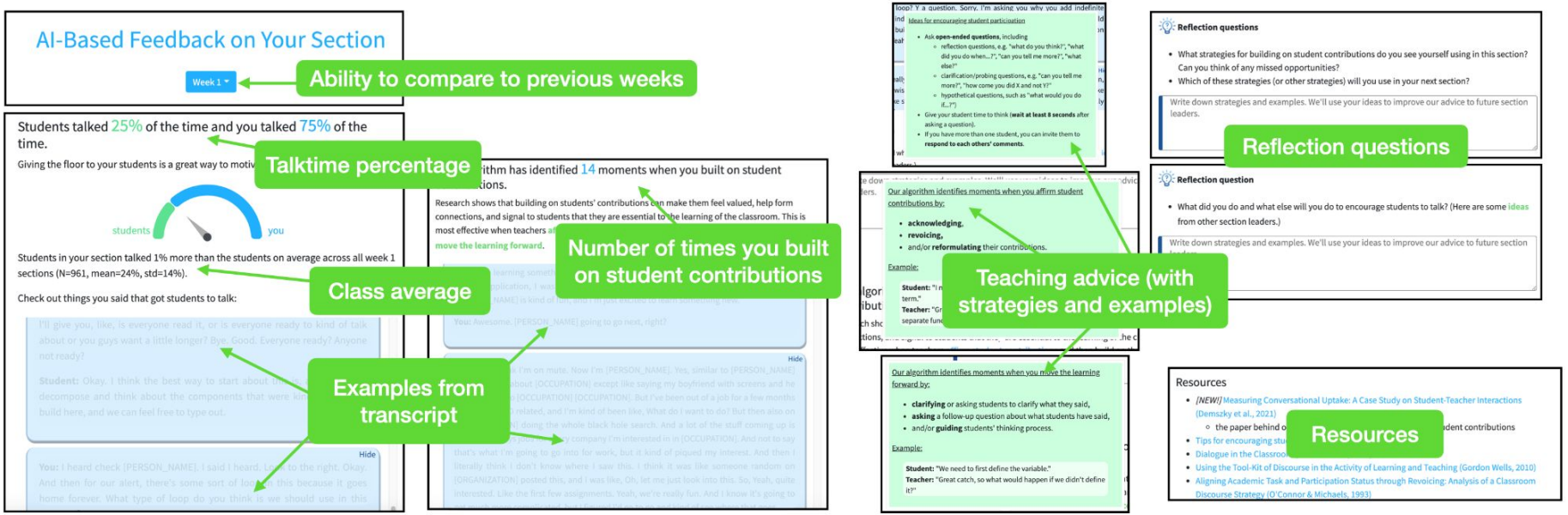
3. Analyze transcript

4. Generate feedback

Steps to running an experiment

1. Set up the backend (i.e. NLP pipeline)
2. **Develop the frontend with users**
3. **Test the end-to-end tool with users**

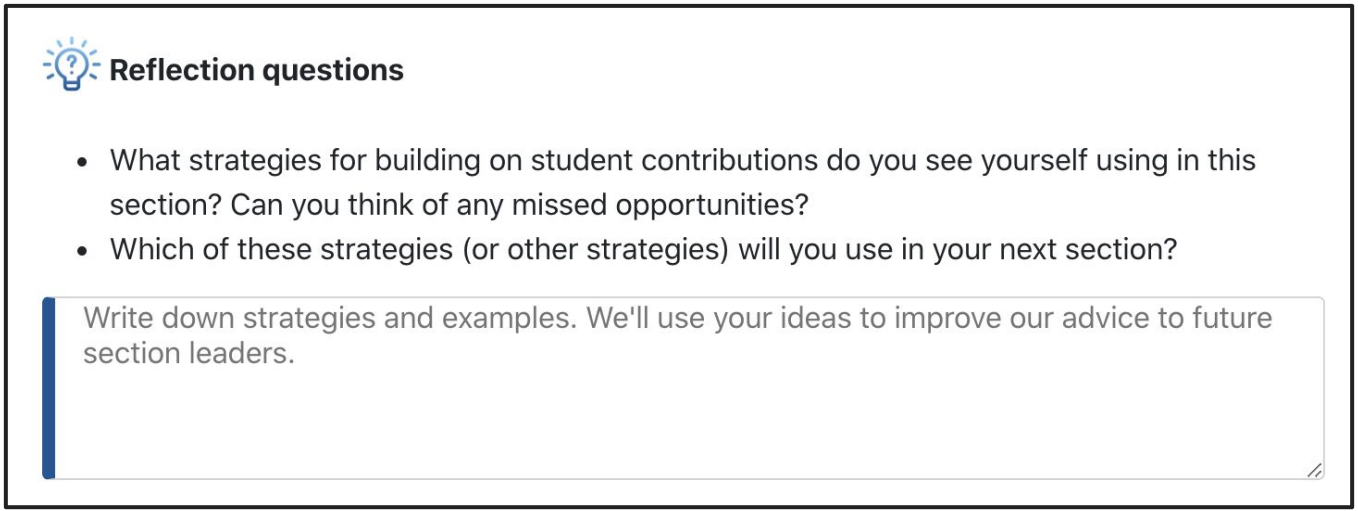
Code in Place NLP Feedback App



Design principles for reflective feedback

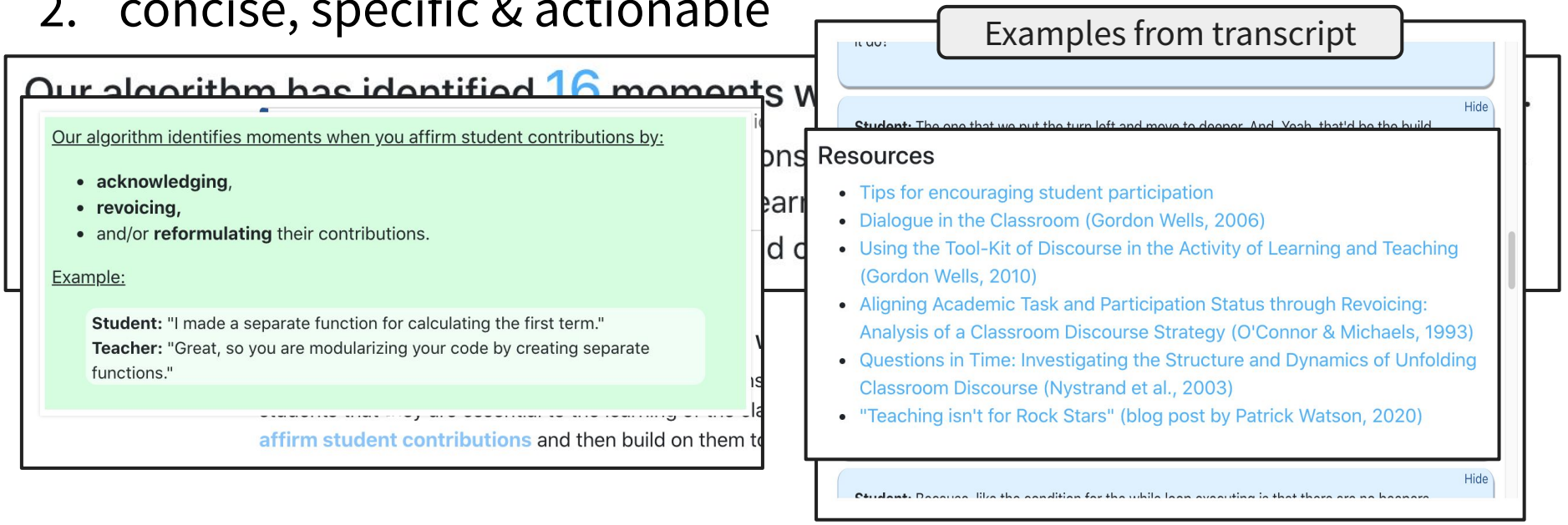
1. non-judgmental & private

"This feedback is meant to give you an opportunity to reflect and to support your professional development. It is not meant as an evaluation."



Design principles for reflective feedback

1. non-judgmental & private
2. concise, specific & actionable



Design principles for reflective feedback

1. non-judgmental & private
2. concise, specific & actionable
3. timely & regular

Steps to running an experiment

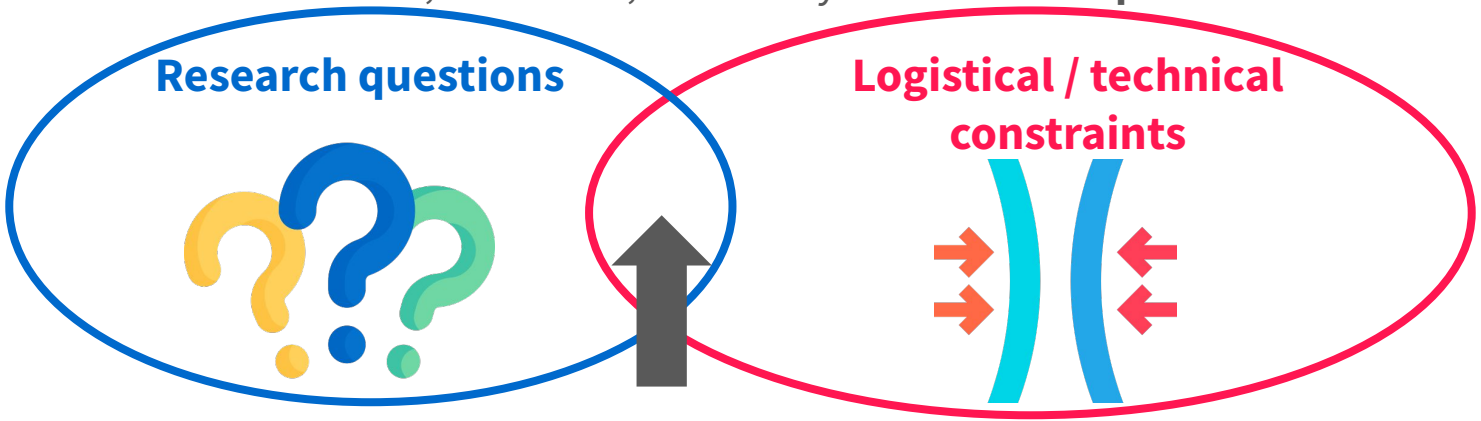
1. Set up the backend (i.e. NLP pipeline)
2. Develop the frontend with users
3. Test the end-to-end tool with users
4. **Figure out the experiment setup**
    a. Who are the participants? What are the conditions? What **quantitative** data will you collect as outcomes, covariates, etc.? Can you also collect **qualitative** data?

Steps to running an experiment

1. Set up the backend (i.e. NLP pipeline)
2. Develop the frontend with users
3. Test the end-to-end tool with users
4. **Figure out the experiment setup**
    a. Who are the participants? What are the conditions? What **quantitative** data will you collect as outcomes, covariates, etc.? Can you also collect **qualitative** data?

Steps to running an experiment

1. Set up the backend (i.e. NLP pipeline)
2. Develop the frontend with users
3. Test the end-to-end tool with users
4. **Figure out the experiment setup**
    a. Who are the participants? What are the conditions? What **quantitative** data will you collect as outcomes, covariates, etc.? Can you also collect **qualitative** data?



3 Platforms

Code in Place

Small group sections

- large sample size
- virtual → ease of integration + better transcription quality
- shared curriculum
- low attendance & lack of robust student outcomes
- limited information on teachers and students

Polygence

1:1 Research Mentorship

- moderate sample size
- virtual → ease of integration + better transcription quality
- more information on mentors & students
- student demographics are not very diverse
- lack of robust student outcomes

TeachFX

K-12 classrooms

- formal teaching context
- given a district partnership, teacher & student demographic / outcome information could be obtained
- existing infrastructure for automated feedback
- experiment confounded by other TeachFX feedback
- low transcription quality (esp. for students)

Code in Place

- democratize access to teaching and learning how to code
- 5-week free online course led by Stanford
- volunteer section leaders
- **12k students + 1.2k section leaders (spring 2021)**

Research questions

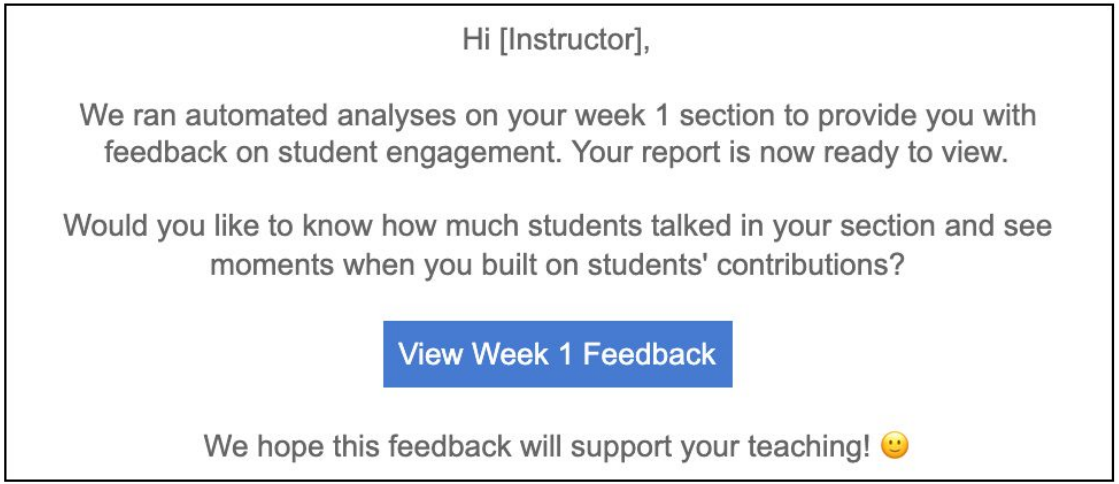
- Does the feedback improve instructors' practice?
- Does the feedback impact student engagement and satisfaction?

Other questions (if time):

- Does uptake correlate with other positive aspects of teaching?
- Do instructors find this feedback helpful?

Setup

- Randomized encouragement study
  - all instructors have access to feedback
  - 50% of instructors receive email reminders

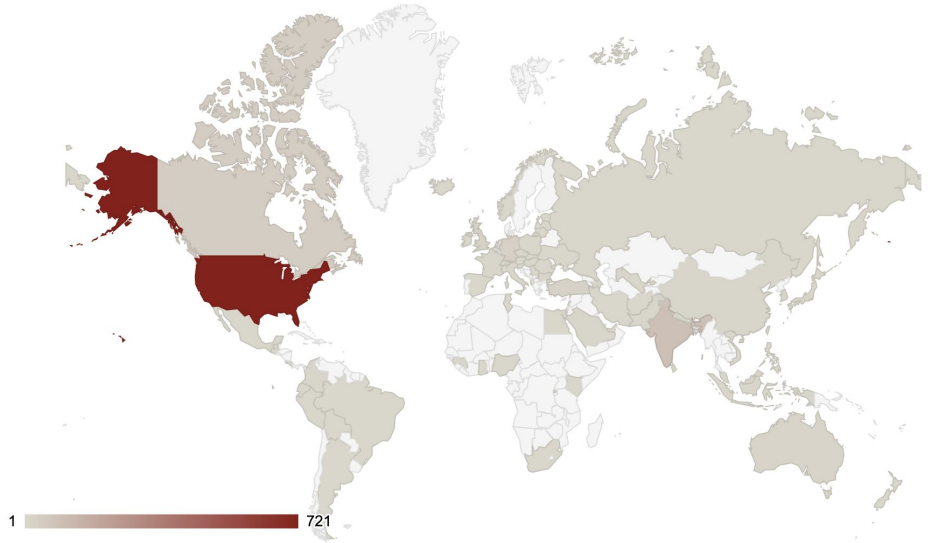


Setup

- Randomized encouragement study
  - all instructors have access to feedback
  - 50% of instructors receive email reminders
- Feedback after each section (5x total)
- Collected data:
  - transcripts
  - whether instructors checked the feedback
  - final survey from instructors and from students
  - student attendance

Data & Participants

- ~3k transcripts
- 880 instructors
  - 89 countries (64% USA, 8% India, 3% Canada, 2% Germany, 2% Turkey, 2% UK, 1% each in other countries)
  - 64% male
  - avg. age is 29



The study has run. Now what?

Analytical steps

1. Explore your data **WITHOUT** looking at the treatment variable
    a. understand which variables are useable (e.g. missingness, distribution)
2. Plan out each of your analyses
3. Pre-register your research questions, hypotheses and analyses (e.g. on Aspredicted or SocialScienceRegistry)
    a. this is not required, but highly encouraged because it facilitates **scientific integrity,** and forces you to think through everything very carefully before you actually run things
    b. example from our work: https://www.socialscienceregistry.org/trials/11258
4. Conduct your analyses

Randomization check

TABLE 2

Randomization Check

Variable	Control *M*	Treatment *M*	*p* value	*n*
Female	0.33	0.31	.52	918
Age	28.88	30.41	.04	917
First-time Code in Place instructor	0.8	0.78	.41	918
In Africa	0.02	0.02	.87	918
In Asia	0.16	0.18	.37	918
In Australia	0.01	0.02	.36	918
In Europe	0.12	0.11	.44	918
In North America	0.68	0.66	.54	918
In South America	0.01	0.01	.82	918
Offered Week 1 section	0.96	0.96	.63	918
Number of uptakes per hour (Week 1)	11.28	10.94	.41	880
Number of questions per hour (Week 1)	32.73	32.28	.66	880
Number of repetitions per hour (Week 1)	34.54	34.23	.77	880
Teacher talk time proportion (Week 1)	0.76	0.76	.96	880

Note. Joint *F* statistic is 0.81. First-time instructor indicates instructors who taught the first time in Code in Place. As this course is voluntary, 38 instructors did not show up in the first section (post randomization), and we thus exclude them from our analysis. We also do not have their Week 1 discourse features.

Research questions

- Does the feedback improve instructors' practice?
- Does the feedback impact student engagement and satisfaction?

Intent to treat
(preferred)

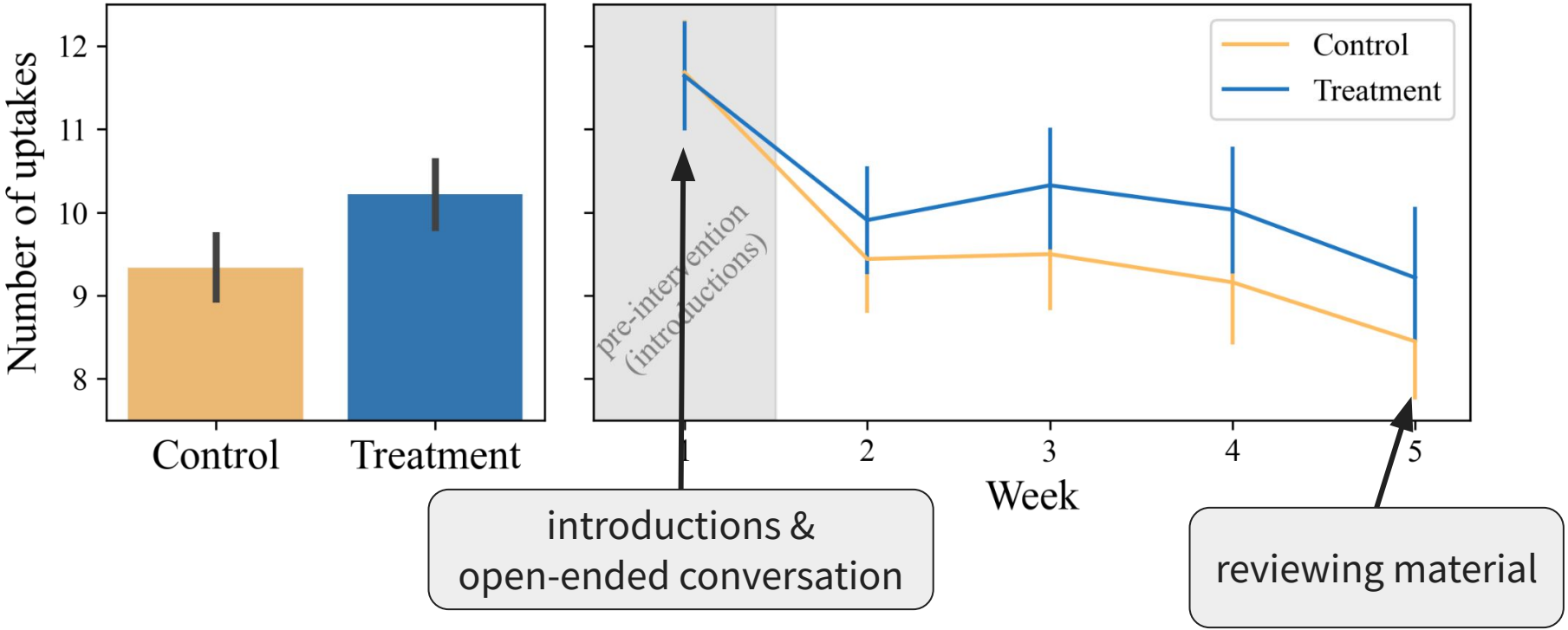
How does treatment status (i.e. receiving the email), regardless of whether a teacher used the feedback affect their practice?

Treatment on the treated
(helps explain effect)

How does checking the feedback affect teachers' practice?

Significant increase in uptake in the treatment group!

= Intent to treat



Does **the feedback** improve uptake?

= Treatment on the treated (ToT)

What method to use that accounts for selection bias? (e.g. people who are motivated to improve their instruction might be more likely to check the feedback)

Does **the feedback** improve uptake?

2 Stage Least Squares Estimator (2SLS)

See video by Ben Lambert to learn more.

1

Checked feedback? **Condition**

$$Feedback_i = \pi_0 + \pi_1 T_i + \pi_2 X_i + \varepsilon_i$$

Covariates

Instructors who got emails were **3.6x more likely** to check the feedback!

Does **the feedback** improve uptake?

2 Stage Least Squares Estimator (2SLS)

1

Checked feedback? **Condition**

$$Feedback_i = \pi_0 + \pi_1 T_i + \pi_2 X_i + \varepsilon_i$$

Covariates

2

Teacher practice
(e.g., number of uptakes)

$$Y_i = \beta_0 + \beta_1 \hat{Feedback}_i + \beta_2 X_i + \mu_i$$

Estimate for checking feedback

Covariates

Covariates

- Instructor demographics
  - In USA, age, is female
- Student demographics
  - In USA, age (bucketed), is female
- First week (pre-intervention) discourse measures
  - Uptake, repetition, questions, talk time
- Week number

Instructors take up student contributions **~2.2 additional times** per section (**~24% increase**) as a result of the feedback

Dependent variable	2SLS Estimate
Number of uptakes	2.209* [±1.070]

*$p < 0.05$, controlling for section duration and teacher-level covariates

Instructors ask **~6.2 additional questions** per section (**~22% increase**) as a result of the feedback

Dependent variable	2SLS Estimate
Number of uptakes	2.209* [±1.070]
Number of teacher questions	6.210* [±2.882]

*$p < 0.05$, controlling for section duration and teacher-level covariates

Instructors **do *not* do more repetition** of student utterances as a result of the feedback

Despite the fact that repetition correlates with uptake in the data ($r=0.80$, $p<0.01$) → suggests that teachers improve on uptake using more sophisticated techniques

Depender	
Number	
Number of teacher questions	6.210* [±2.882]
Number of repetitions	4.355 [±3.478]

*$p < 0.05$, controlling for section duration and teacher-level covariates

But wait, there’s a catch!

How do you measure “checking feedback”?

- ➔ Our original measure = “did the instructor check their **prior week’s** feedback?”
- ➔ Reviewer 2: Change in practice can is not only affected by whether they opened their feedback the **week right before** but also if they opened it in **any prior week**. This violates assumptions for the two stage least squares regression. You should instead use “**ever opened the feedback until week X”** as the instrument.

Updated estimates after reviewer 2's feedback!

Dependent variable	2SLS Estimate
Number of uptakes	1.125* [±0.491]
Number of teacher questions	3.169* [±1.344]
Number of repetitions	1.947 [±1.606]

*$p < 0.05$, controlling for section duration and teacher-level covariates

Results didn't change significantly, but the takeaway is: think through all your **assumptions** & get feedback on your analyses from many perspectives!!!

Research questions

- Does the feedback improve teacher practice?
- Does the feedback impact student engagement and satisfaction?
    - students' end-of-course survey responses (16% response rate)
    - student attendance

Feedback improves students' **response rates to survey**

Dependent variable	2SLS Estimate
% of students responding to survey	0.069* [±0.029]

*$p < 0.05$, controlling for teacher-level covariates

Feedback improves students' **overall course ratings**

Dependent variable	2SLS Estimate
% of students responding to survey	0.069* [±0.029]
% of students recommending the course (7+ rating)	0.078* [±0.029]

*$p < 0.05$, controlling for teacher-level covariates

Feedback improves students' ratings of **section helpfulness**

Dependent variable	2SLS Estimate
% of students responding to survey	0.069* [±0.029]
% of students recommending the course (7+ rating)	0.078* [±0.029]
% of students rating the section as helpful	0.046* [±0.022]

*$p < 0.05$, controlling for teacher-level covariates

Feedback did not have a significant effect on student attendance

Dependent variable	2SLS Estimate
% of students responding to survey	0.069* [±0.029]
% of students recommending the course (7+ rating)	0.078* [±0.029]
% of students rating the section as helpful	0.046* [±0.022]
Student attendance	0.364 [±0.364]

*$p < 0.05$, controlling for teacher-level covariates

Takeaways

- Explore validity of data & define research questions **before** running ITT / ToT analyses
- ITT is generally preferred, but ToT can help explain magnitude and treatment mechanisms
- Run your assumptions by other people (ask them to be your reviewer 2), especially those with a **stats background**

Extra slides

Experimental validation

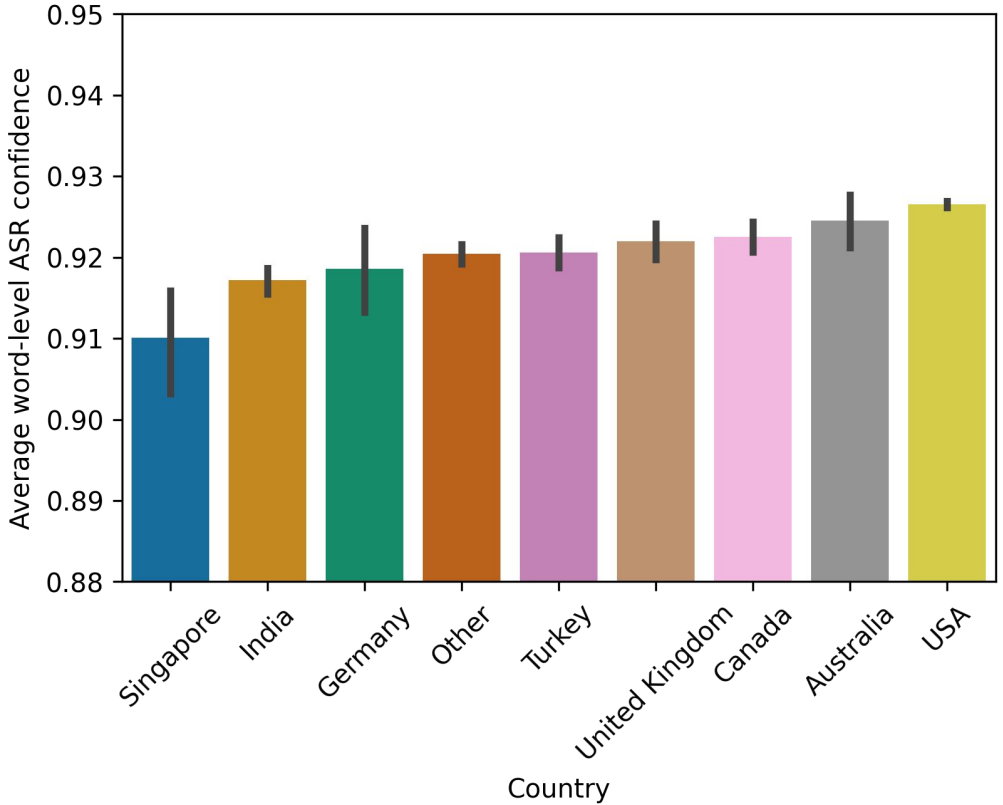
Randomization was performed pre-intervention.

Variable	**Treatment**	**Control**	***t*-statistic**	***p*-value**
Number of instructors	568	568	N/A	N/A
% female	33%	32%	0.38	0.71
% in USA	63%	63%	0.12	0.90
% returning instructors	21%	19%	0.80	0.42
Avg age	29.0	28.5	1.64	0.10

Intervention data statistics

Transcripts	pre-intervention N (week 1)	945
	N (weeks 2-5)	3,002
Instructors	N (weeks 2-5)	880
	country	89 unique countries; 64% USA, 8% India, 3% Canada, 2% Germany, 2% Turkey, 2% UK, 1% each in other countries
	gender	65% male, 33% female, 1% non-binary, 1% missing
	age	M=29, STD=11

ASR confidence by country



Developing an equitable ASR model

- Create a representative dataset
- Evaluation framework [Demszky et al., 2020]
- Develop custom models

- Join forces with related efforts to make ASR more equitable:
  - Koenecke et al., 2020
  - Koh et al., 2020
  - Aloufi et al., 2020

Uptake correlates with **number of teacher questions**

Independent variable	Coef. (Mixed effects model)
Number of teacher questions	0.336*** [±0.003]

***p < 0.001, controlling for section duration

Uptake correlates with amount of **revoicing / repetition**

Independent variable	Coef. (Mixed effects model)
Number of teacher questions	0.336*** [±0.003]
Number of repetitions (%-IN-T > 0)	0.254*** [±0.004]

***p < 0.001, controlling for section duration

Uptake correlates with the **number of students speaking** in class

Independent variable	Coef. (Mixed effects model)
Number of teacher questions	0.336*** [±0.003]
Number of repetitions (%-IN-T > 0)	0.254*** [±0.004]
Number of students speaking	0.680*** [±0.049]

***p < 0.001, controlling for section duration

Uptake correlates with the **number of students attending** class

Independent variable	Coef. (Mixed effects model)
Number of teacher questions	0.336*** [±0.003]
Number of repetitions (%-IN-T > 0)	0.254*** [±0.004]
Number of students speaking	0.680*** [±0.049]
Student attendance	0.323*** [±0.048]

***p < 0.001, controlling for section duration

Uptake correlates **negatively** with average **teacher utterance length**

Independent variable	Coef. (Mixed effects model)
Number of teacher questions	0.336*** [±0.003]
Number of repetitions (%-IN-T > 0)	0.254*** [±0.004]
Number of students speaking	0.680*** [±0.049]
Student attendance	0.323*** [±0.048]
Avg teacher utterance length	-0.002*** [±0.000]

***p < 0.001, controlling for section duration

Uptake correlates **negatively** with **teacher talktime proportion**

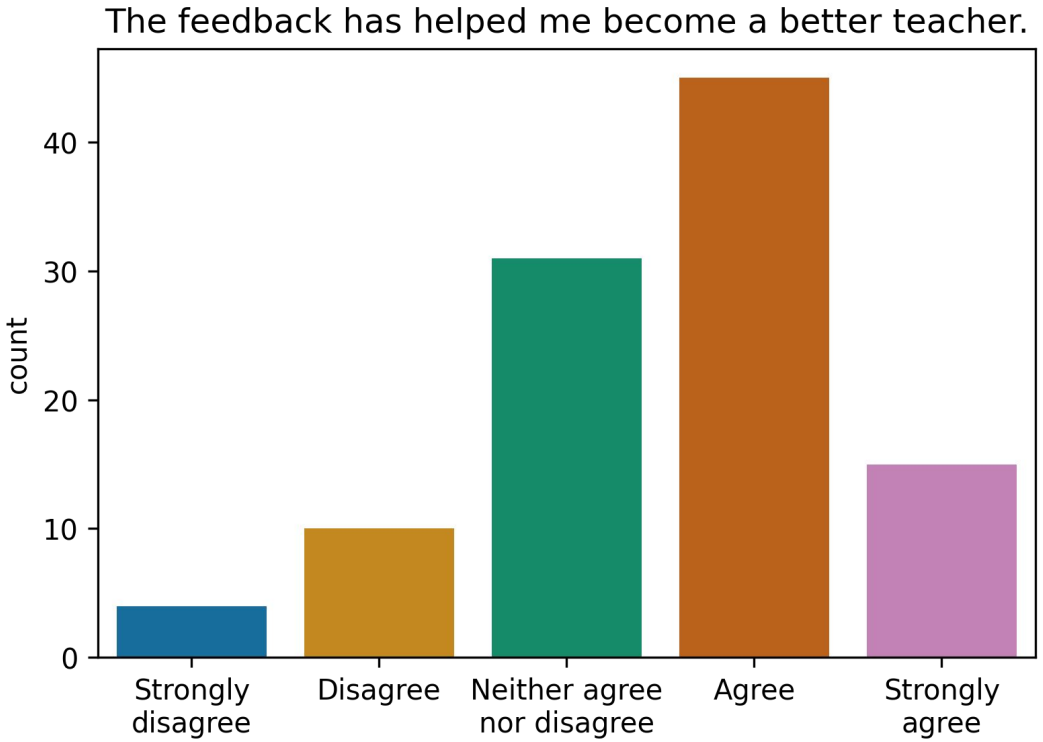
Independent variable	Coef. (Mixed effects model)
Number of teacher questions	0.336*** [±0.003]
Number of repetitions (%-IN-T > 0)	0.254*** [±0.004]
Number of students speaking	0.680*** [±0.049]
Student attendance	0.323*** [±0.048]
Avg teacher utterance length	-0.002*** [±0.000]
Teacher talktime proportion	-17.207*** [±0.704]

***p < 0.001, controlling for section duration

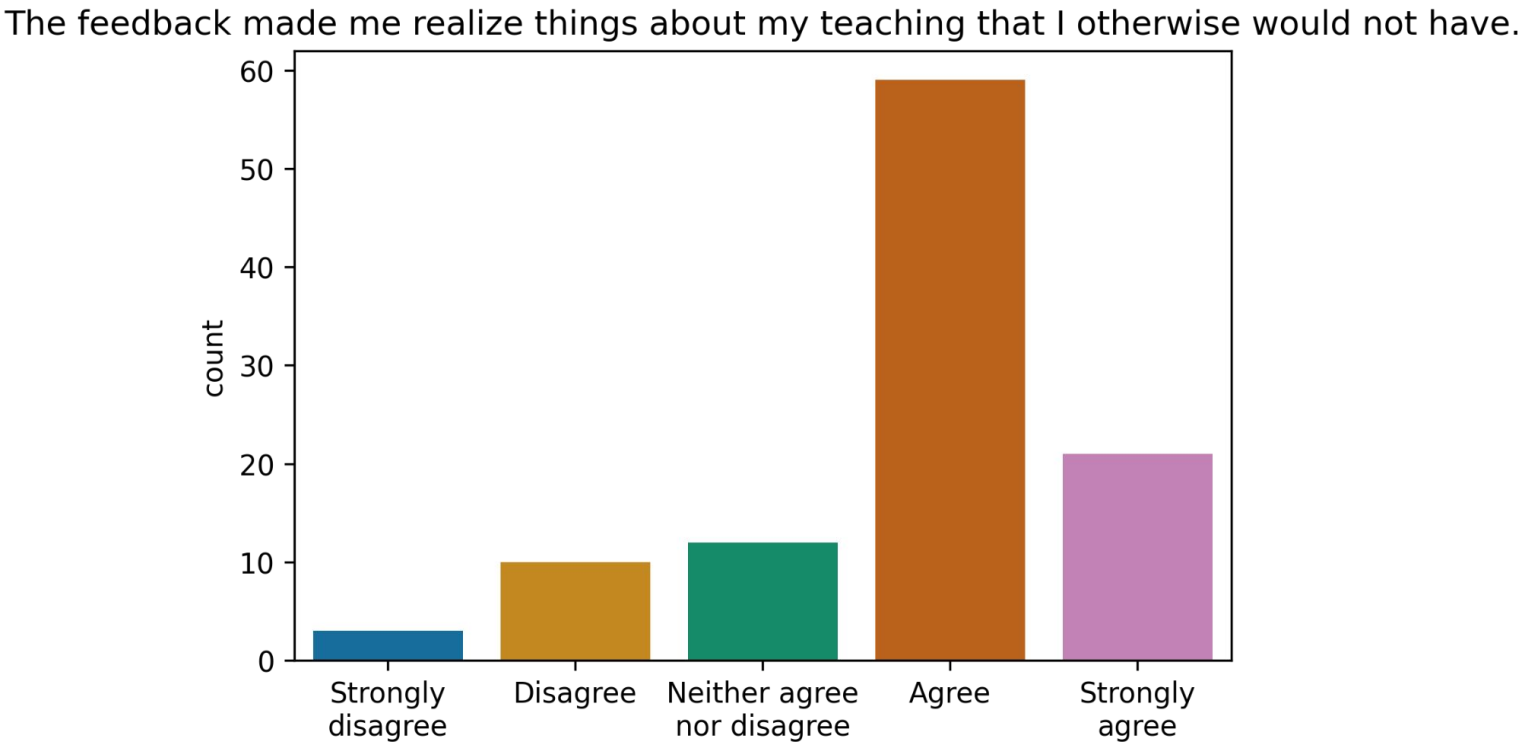
Final survey for teachers

- **Surveyed a random sample of 200 teachers anonymously**
  - Incentive: lottery for 10 x $40 gift cards
  - Teachers could be from either condition
- **71% response rate (N=142)**
- **73% reported to have looked at the feedback at least once**
  - Main reason for not looking: did not know about it (80%)

The majority of teachers (57%) said the tool helped them become a better teacher.

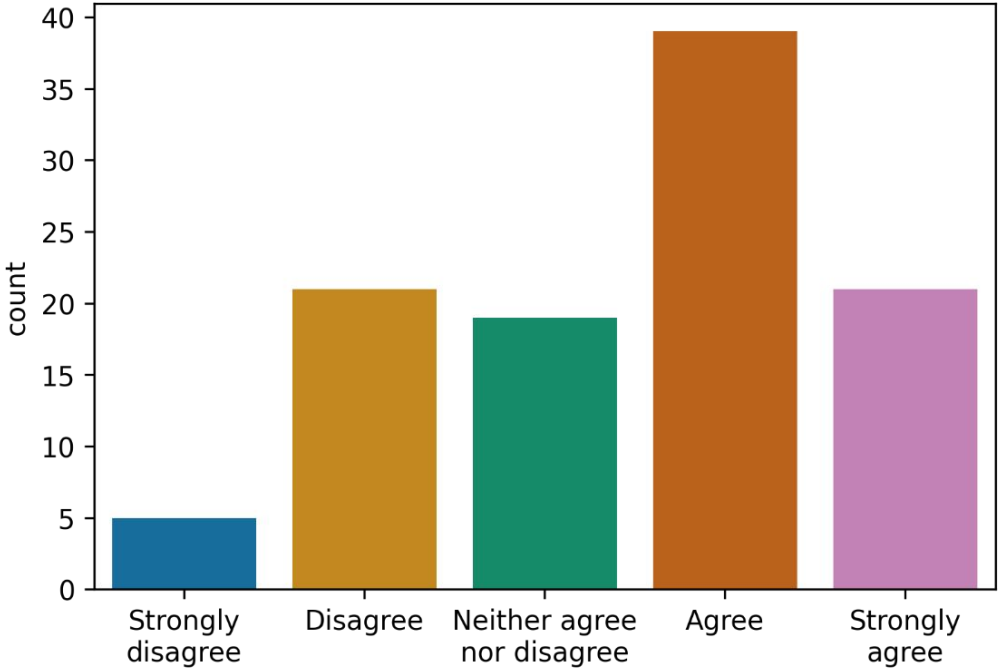


The vast majority of teachers (76%) said the tool made them realize things about their teaching they otherwise wouldn't have.

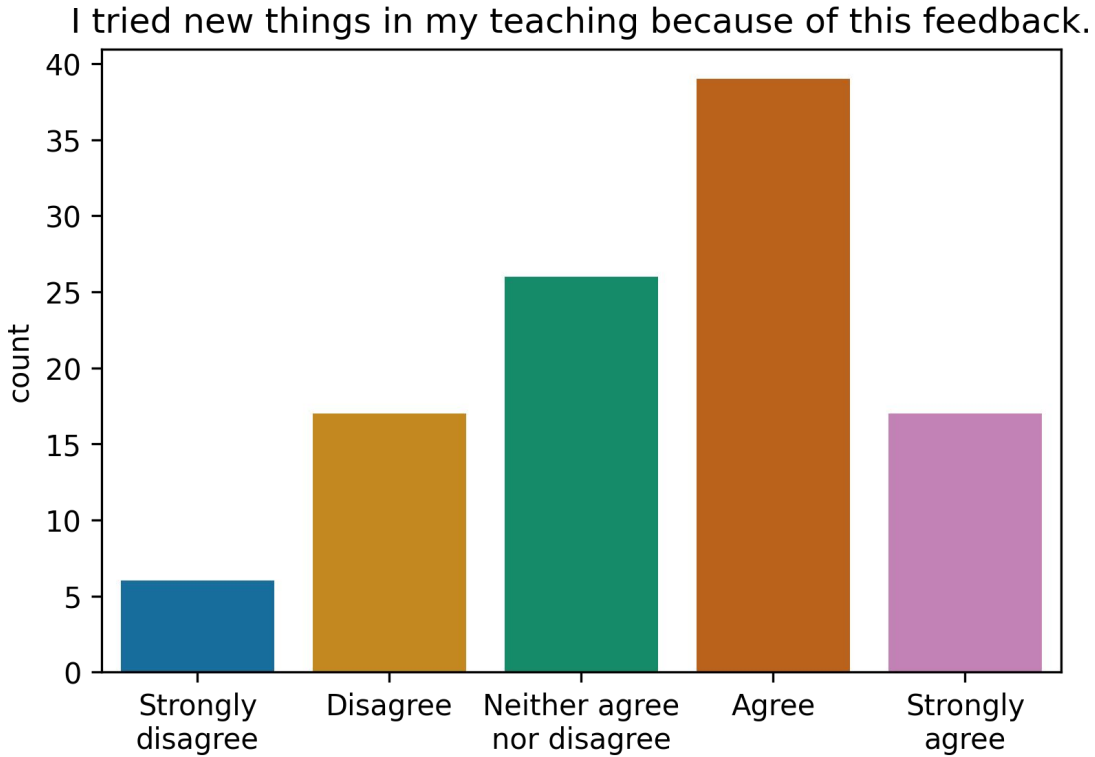


The majority of teachers (57%) said the tool made them pay more attention to who was getting a voice in their class.

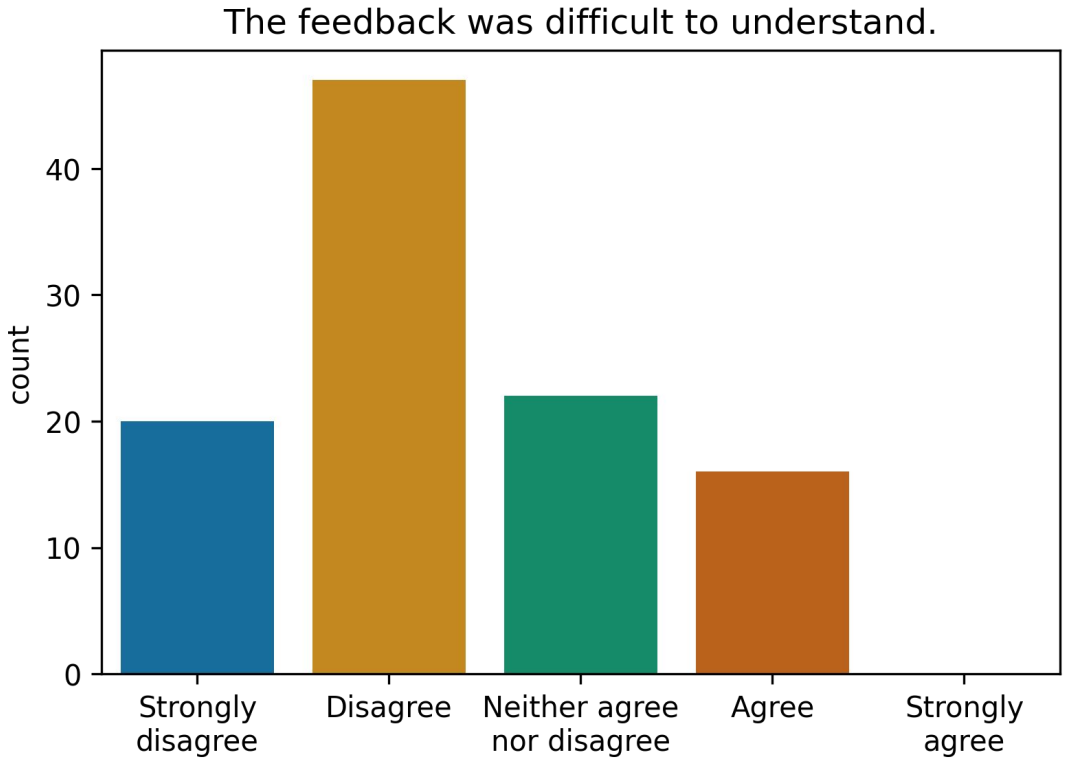
The feedback made me pay more attention to who was getting a voice in my class than I otherwise would have.



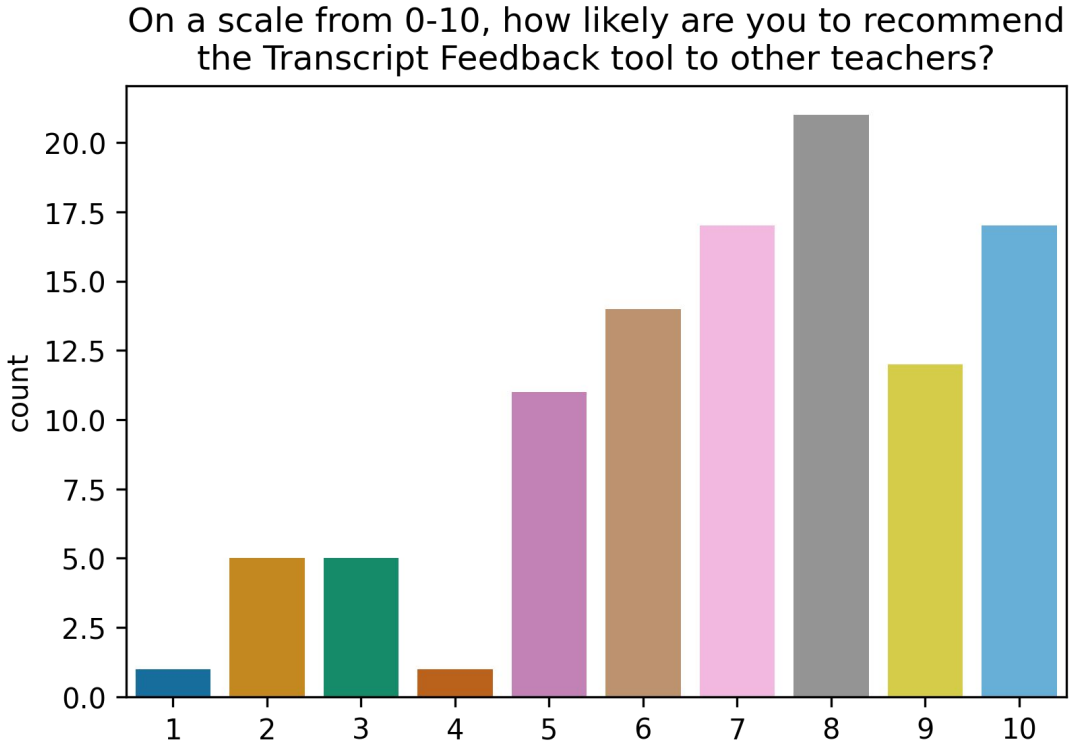
The majority of teachers (53%) said they tried new things in their teaching as a result of the feedback.



The majority of teachers (64%) said the feedback wasn't difficult to understand.

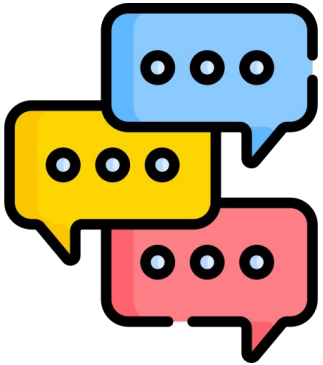


Teachers gave an average score of **7 out of 10** in terms of how likely they are to recommend the tool



Suggestions for improvement (open-ended responses)

- improve **ASR quality** (20 out of 62 mentions)
- incorporate **chat** (8 out of 62 mentions)c



Open-ended comments

The transcript feedback tool was really helpful and gave me insights and data that I couldn't have possibly had otherwise! Keep up with the great work, hope this becomes a standard tool for teachers all over the globe :)

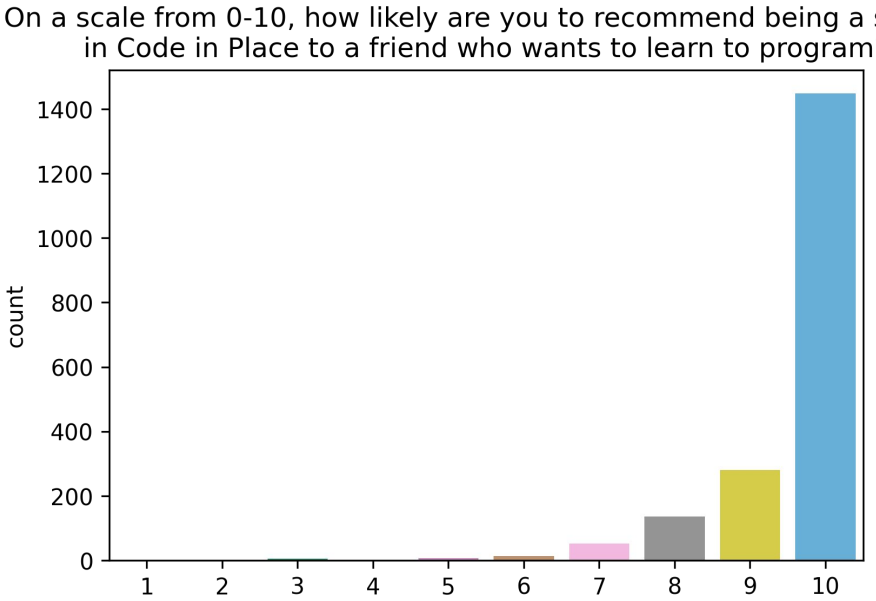
I think, overall, it was a helpful tool. The data provided is very wholesome and focuses on the growth of the teacher in terms of understanding his/her teaching style and is also a constant reminder of to incorporate a pedagogy that involves dialogue.

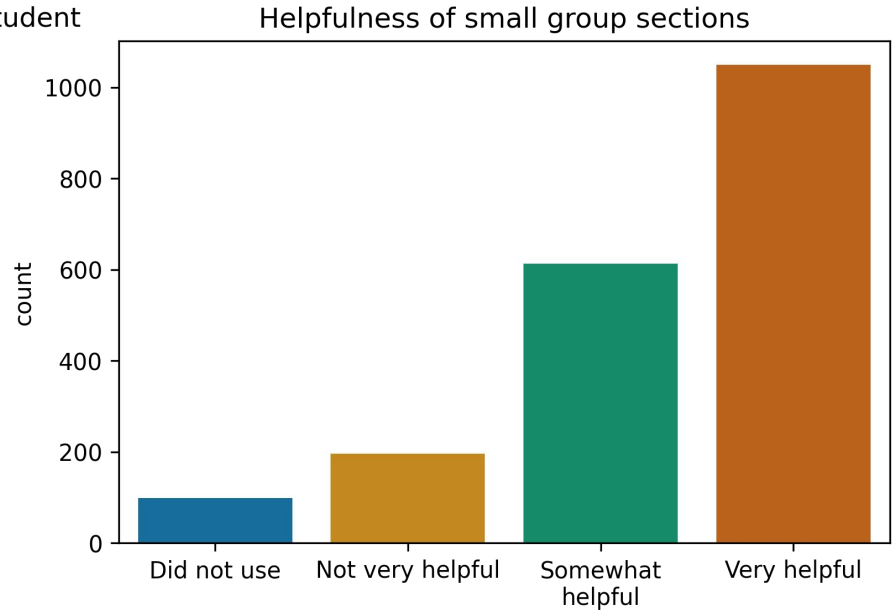
I think it was useful, however it would be nice to have recommendations regarding the demographics of the students. For example, my group was primary from India and I'm from Colombia and because of our cultures we have been thought very different ways of interacting and engaging in class. So while I was trying to do group al activities were everyone interacts, my students wanted to listen to me talk during all the section and wouldn't answer unless I called them to answer.

Such an amazing tool!! I have always been looking forward to this every week.

Student final survey statistics

Response rate: 16% (N=1958 out of 12179)





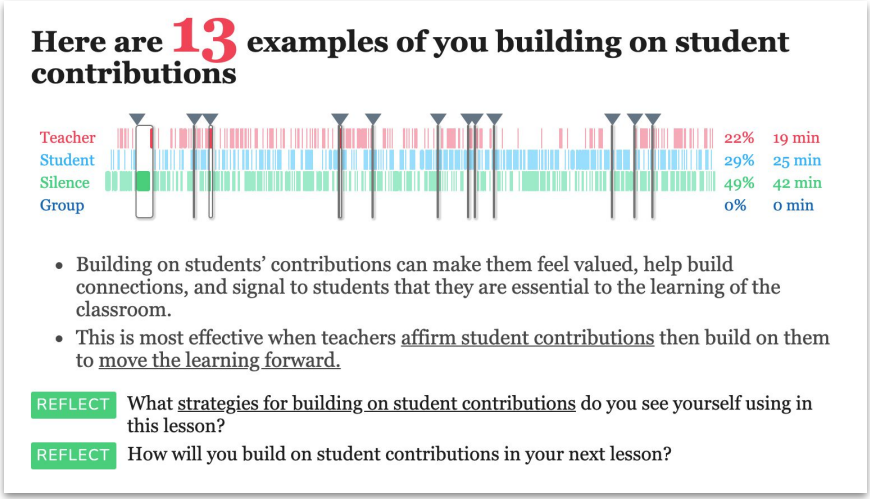
TeachFX Study

Blog posts for teachers about uptake

Jan 18 • Written By Guest User

6 Practices for Building on Student Contributions

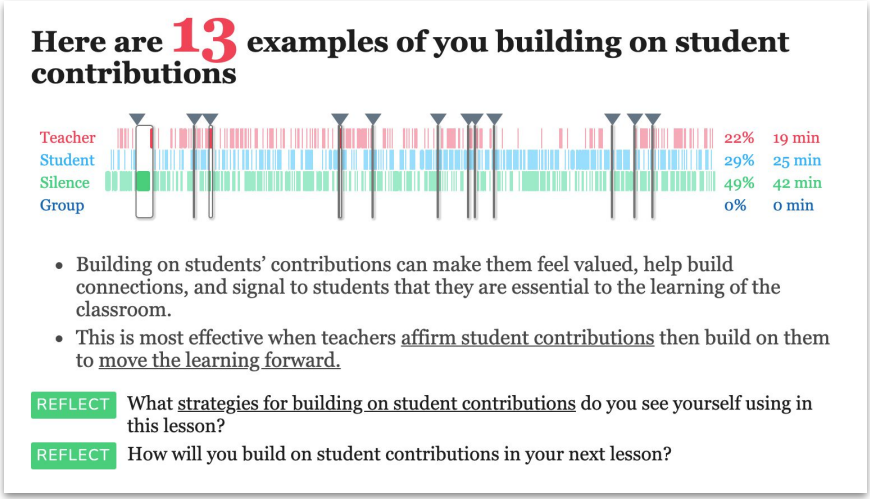
Ms. Detroit, 5th grade science teacher



A screenshot of one of Ms. Detroit's class reports

- **"No benchmarks" problem**. "The biggest question for me is, is it enough?" Me: "What would enough look like?" Her: "If there are 13 examples over the course a 1.5hr class here...I think I would like to have at least double that so I'd have every like 3-5 minutes: questioning, building on students contributions, questioning, building on students contributions, without getting too much into the back and forth...but I don't want my building on student contributions to make this more teacher led" (meaning, she was worried that if she focused too much on building on student contributions, she'd end up with too much teacher talk)
    - Note: "Double" is a VERY ambitious goal. We should be prompting teachers on how to set realistic goals to incrementally improve their practice. She has no way of knowing that 13 is a VERY high number of uptake examples to get, relative to how many examples are surfaced in a typical class report.
- **Particularly valued the 6 strategies for their "how-to" value, for helping her reflect on how to teach better**. Didn't seem to register that the strategies were listed on the slide to inform how the algorithm was working. She had taken a screenshot of the 6 strategies and stored it on a folder on her computer for future reference.
- **Very trusting of the accuracy of the data**. The slide read, "Here are 13 examples of you building on student contributions" -- she took it as a given that these 13 examples were all of the examples of her building on student contributions. Did not seem to question that.
- **Understood "building on student contributions" to be synonymous with "follow up questions"**. A big part of her interest in this insight was because she is working on asking better follow up questions.

Ms. Detroit, 5th grade science teacher

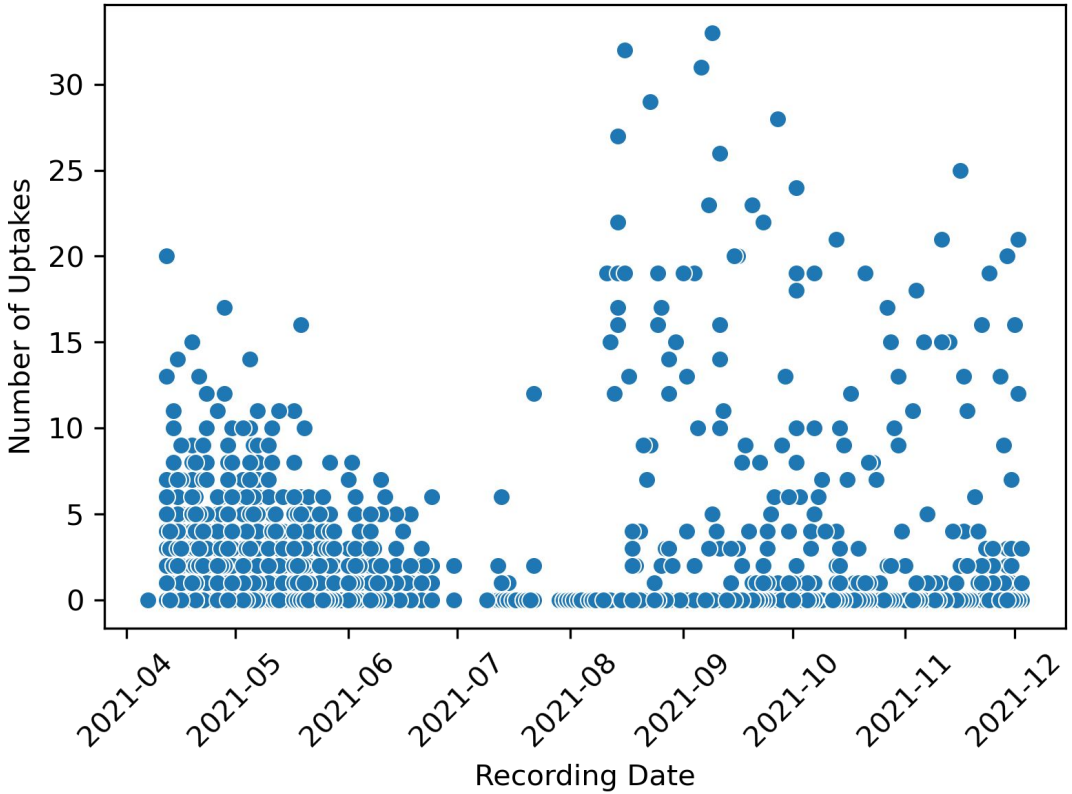


A screenshot of one of Ms. Detroit's class reports

- "No benchmarks" problem.
- Particularly valued the 6 strategies for their "how-to" value, for helping her reflect on how to teach better
- Very trusting of the accuracy of the data. Understood "building on student contributions" to be synonymous with "follow up questions".

Promising preliminary results!

TeachFX



Increase in uptake over time by user

Mixed Linear Model Regression Results

Model:	MixedLM	Dependent Variable:	num_uptake_zscore
No. Observations:	4951	Method:	REML
No. Groups:	195	Scale:	0.7943
Min. group size:	1	Log-Likelihood:	-6564.3453
Max. group size:	352	Converged:	Yes
Mean group size:	25.4		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	0.026	0.036	0.717	0.473	-0.045	0.096
timestamp_zscore	0.149	0.019	8.058	0.000	0.113	0.185
duration_zscore	0.180	0.015	12.396	0.000	0.152	0.209
Group Var	0.129	0.023				

Teachers who looked at the feedback increase their uptake significantly more

Mixed Linear Model Regression Results

Model:	MixedLM	Dependent Variable:	num_uptake_zscore
No. Observations:	4951	Method:	REML
No. Groups:	195	Scale:	0.7998
Min. group size:	1	Log-Likelihood:	-6575.2499
Max. group size:	352	Converged:	Yes
Mean group size:	25.4		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	0.007	0.035	0.198	0.843	-0.062	0.076
teacher_interacted_prev[T.True]	1.103	0.177	6.249	0.000	0.757	1.449
duration_zscore	0.173	0.015	11.884	0.000	0.144	0.202
Group Var	0.119	0.021				

Polygence study

M-Powering Teachers

POWERING TEACHERS

Data-driven, non-judgmental feedback on instructor's discourse, encouraging **dialogic teaching practices.**

Stanford | GRADUATE SCHOOL OF EDUCATION

M-Powering Teachers

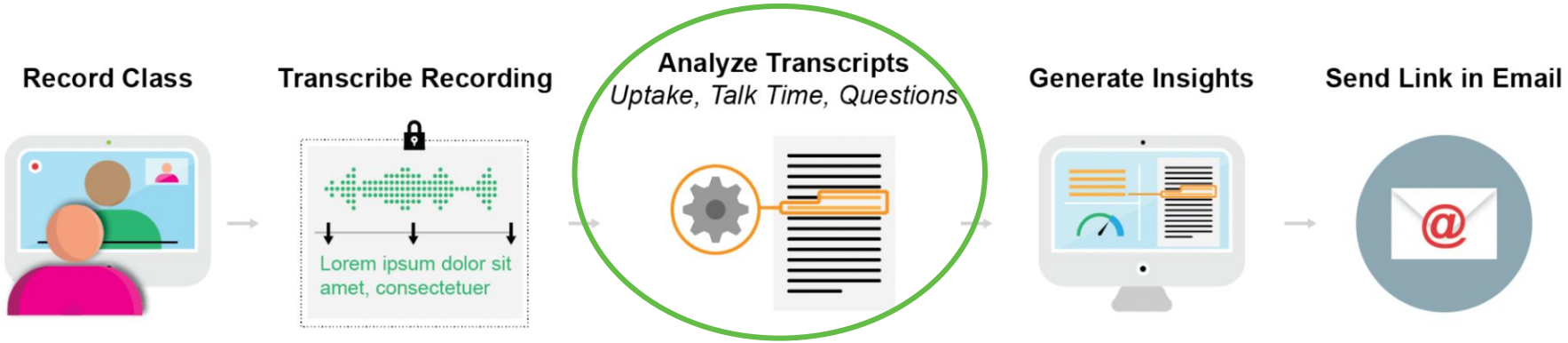
Data-driven, non-judgmental feedback on instructor's discourse, encouraging **dialogic teaching practices.**

Stanford | GRADUATE SCHOOL OF EDUCATION

M-Powering Teachers

POWERING TEACHERS

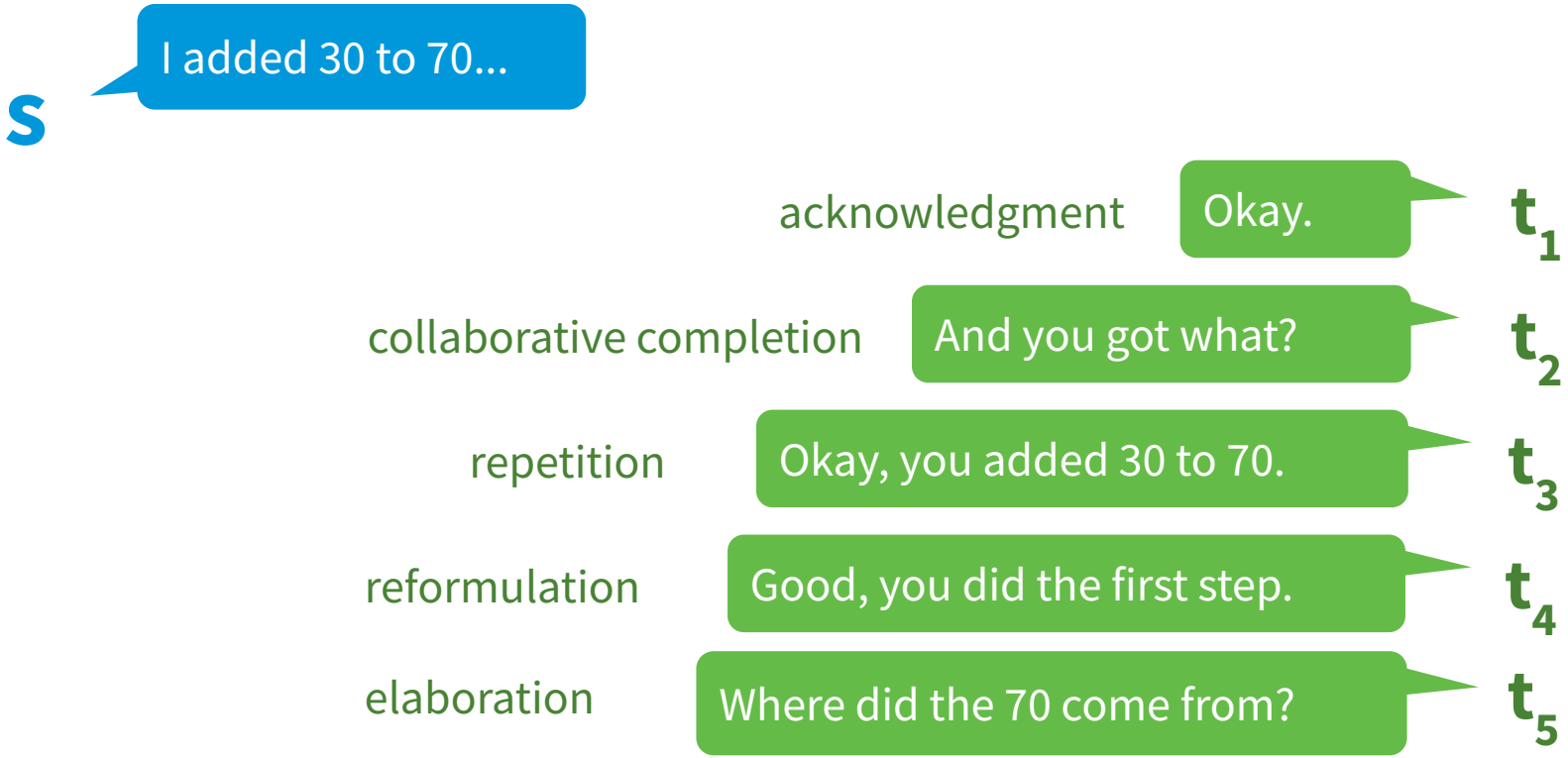
Data-driven, non-judgmental feedback on instructor's discourse, encouraging **dialogic teaching practices.**



Focus:
Teachers' **uptake**
of student ideas

Stanford | GRADUATE SCHOOL OF EDUCATION

Uptake is to build on the interlocutor's contribution



When teachers take up student ideas, ...

- They **amplify student voices** and promote **dialogic instruction**
  [Wells, 1999; Nystrand et al., 1997]
- **Students learn and do better**
  [Brophy, 1984; O'Connor & Michaels, 1993; Nystrand et al., 2003]

Our measure of uptake

- Unsupervised NLP measure, powered by an LLM (Bert)
- Correlates positively with expert observation scores and value-added scores

Measuring Conversational Uptake:
A Case Study on Student-Teacher Interactions

Dorottya Demszky[1] Jing Liu[2] Zid Mancenido[3] Julie Cohen[4]
Heather Hill[3] Dan Jurafsky[1] Tatsunori Hashimoto[1]
[1]Stanford University [2]University of Maryland [3]Harvard University [4]University of Virginia
{ddemszky, thashim}@stanford.edu

Abstract

In conversation, *uptake* happens when a speaker builds on the contribution of their interlocutor by, for example, acknowledging, repeating or reformulating what they have said. In education, teachers' uptake of student contributions has been linked to higher student achievement. Yet measuring and improving teachers' uptake at scale is challenging, as existing methods require expensive annotation by experts. We propose a framework for computationally measuring uptake, by (1) releasing a dataset of student-teacher exchanges extracted from US math classroom transcripts annotated for uptake by experts; (2) formalizing uptake as pointwise Jensen-Shannon Divergence (PJSD), estimated via next utterance classification; (3) conducting a linguistically-motivated comparison of different unsupervised measures and (4) correlating these measures with educational outcomes. We find that although repetition captures a significant part of uptake, PJSD outperforms repetition-based baselines, as it is capable of identifying a wider range of uptake phenomena like question answering and reformulation. We apply

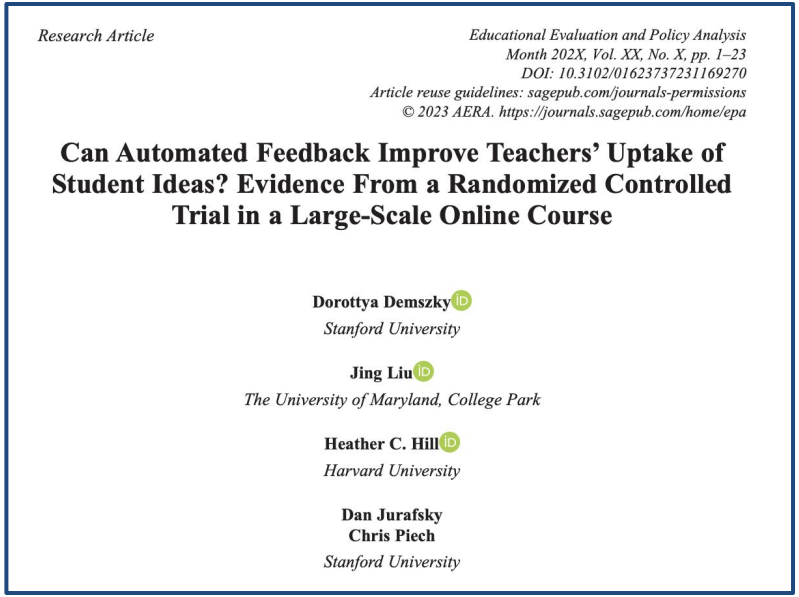
Figure 1: Example student utterance s and possible teacher replies t, illustrating different uptake strategies.

which is especially important in contexts like education. Teachers' uptake of student ideas promotes dialogic instruction by amplifying student voices and giving them agency in the learning process, unlike monologic instruction where teachers lecture at students (Bakhtin, 1981; Wells, 1999; Nystrand et al., 1997). Despite extensive research showing

Demszky et al., ACL '21

Prior success of M-Powering Teachers

- Online small group instruction for programming
- Automated feedback **improves instructor's uptake of student ideas by 13%** and increases students' satisfaction with the course and assignment completion



Research Article

Educational Evaluation and Policy Analysis
Month 202X, Vol. XX, No. X, pp. 1–23
DOI: 10.3102/01623737231169270
Article reuse guidelines: sagepub.com/journals-permissions

Can Automated Feedback Improve Teachers' Uptake of Student Ideas? Evidence From a Randomized Controlled Trial in a Large-Scale Online Course

Dorottya Demszky
Stanford University

Jing Liu
The University of Maryland, College Park

Heather C. Hill
Harvard University

Dan Jurafsky
Chris Piech
Stanford University

Demszky et al., EEPA '23

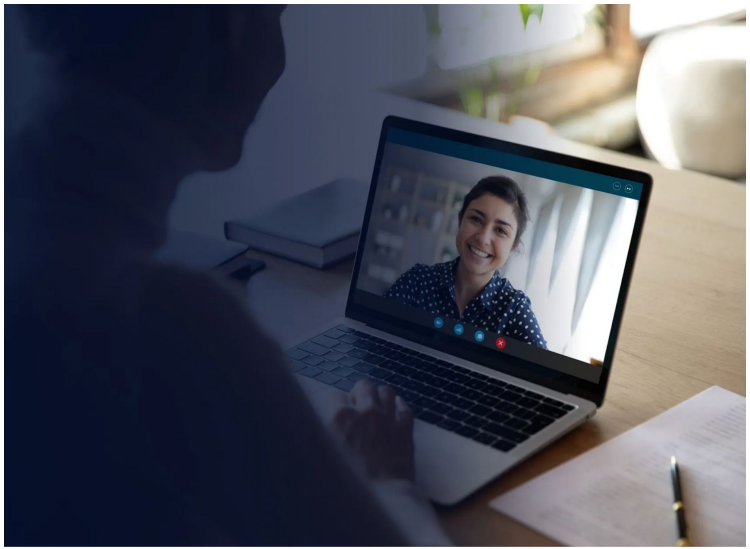
POWERING TEACHERS

Stanford | GRADUATE SCHOOL OF EDUCATION

Does the positive impact of M-Powering Teachers generalize to a 1:1 teaching context?

1:1 Research Mentorship Program (Polygence)

- Students are mostly in high school
- Mentors are usually graduate students
- Most mentors and students are in the US
- Students and mentors meet for 10 sessions finished over ~4 months
- The program takes place via Zoom



Research Questions

RQ1 What percentage of mentors engage with the automated feedback?

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

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

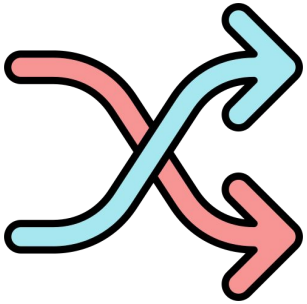
RQ4 What is the impact of automated feedback on project outcomes?

Participants

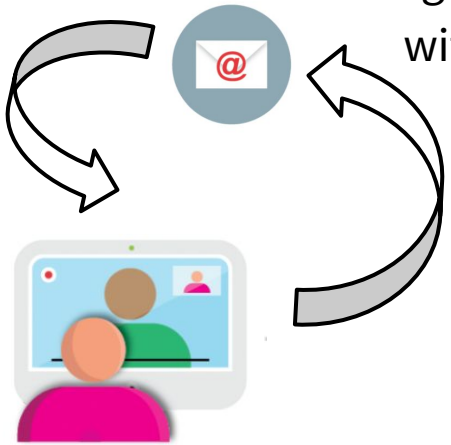
- 414 mentors
  - signed up after April, 2022
- 624 students

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

Experimental Design



Random assignment upon signup

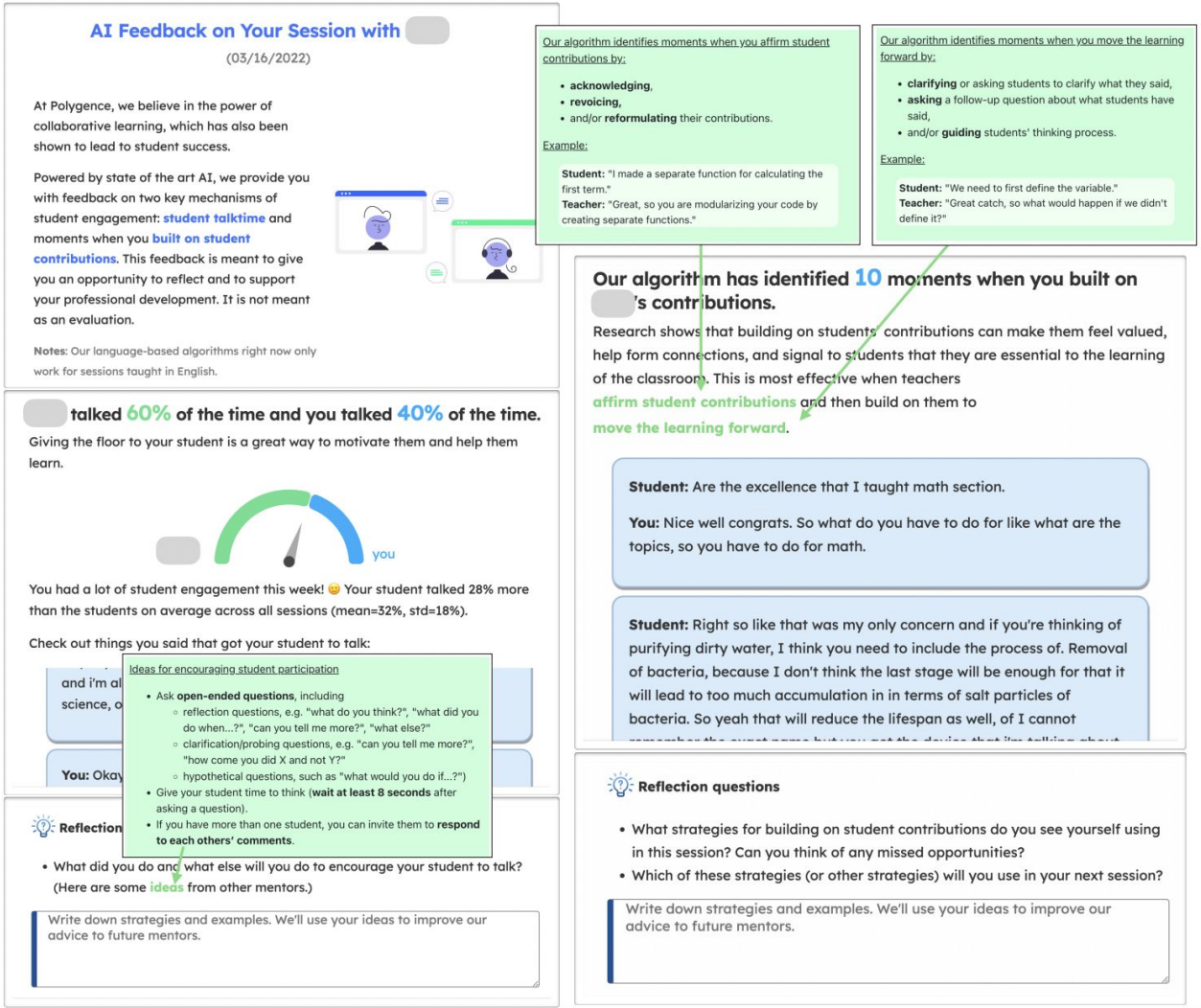


Treatment group gets feedback within 1 day via email link

Teaching session x 10

Project completion

Interface of M-Powering Teachers



Results

RQ1 What percentage of mentors engage with the automated feedback?

Results

RQ1 What percentage of mentors engage with the automated feedback?

84% of mentors checked the feedback at least once, mostly in the first session (74%) then less frequently

Results

RQ1 What percentage of mentors engage with the automated feedback?

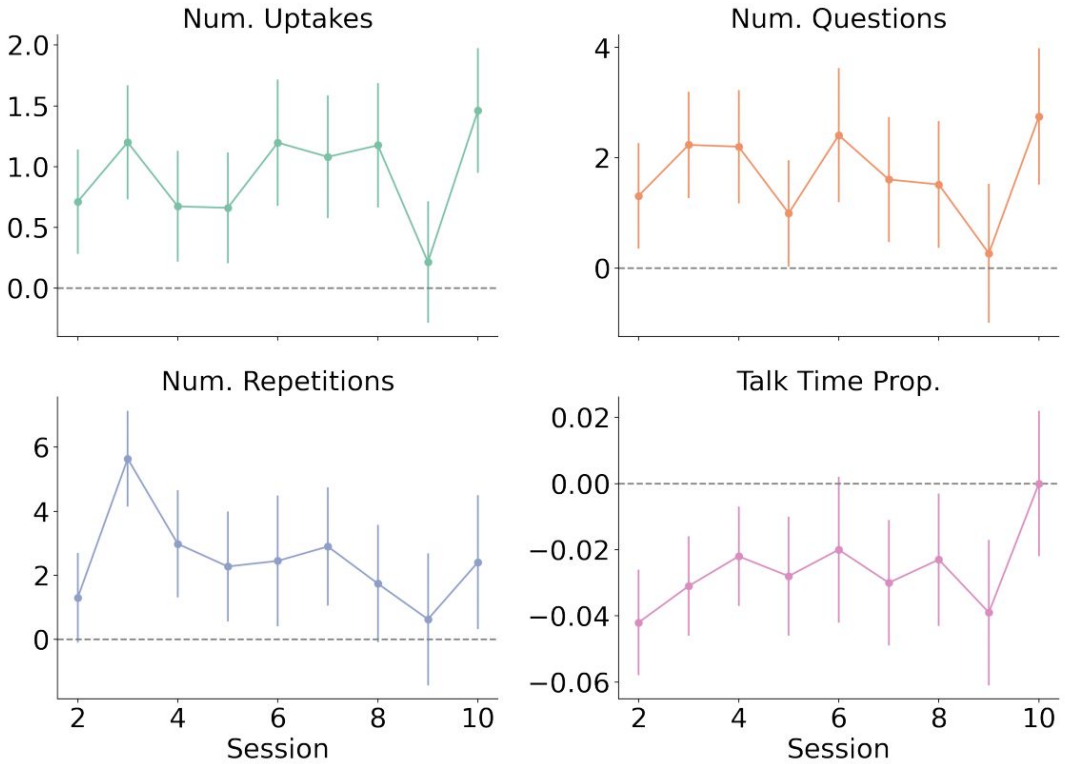
RQ2 What is the impact of automated feedback on mentors' instruction?

Mentors who receive feedback…

- **take up student ideas 9 % more** ($p < 0.05$)
- ask 6% more questions ($p < 0.1$)
- repeat student contributions 6 % more ($p < 0.05$)
- talk 5% less ($p < 0.01$)

Controlling for mentor and student demographic features.

The trends persist over time



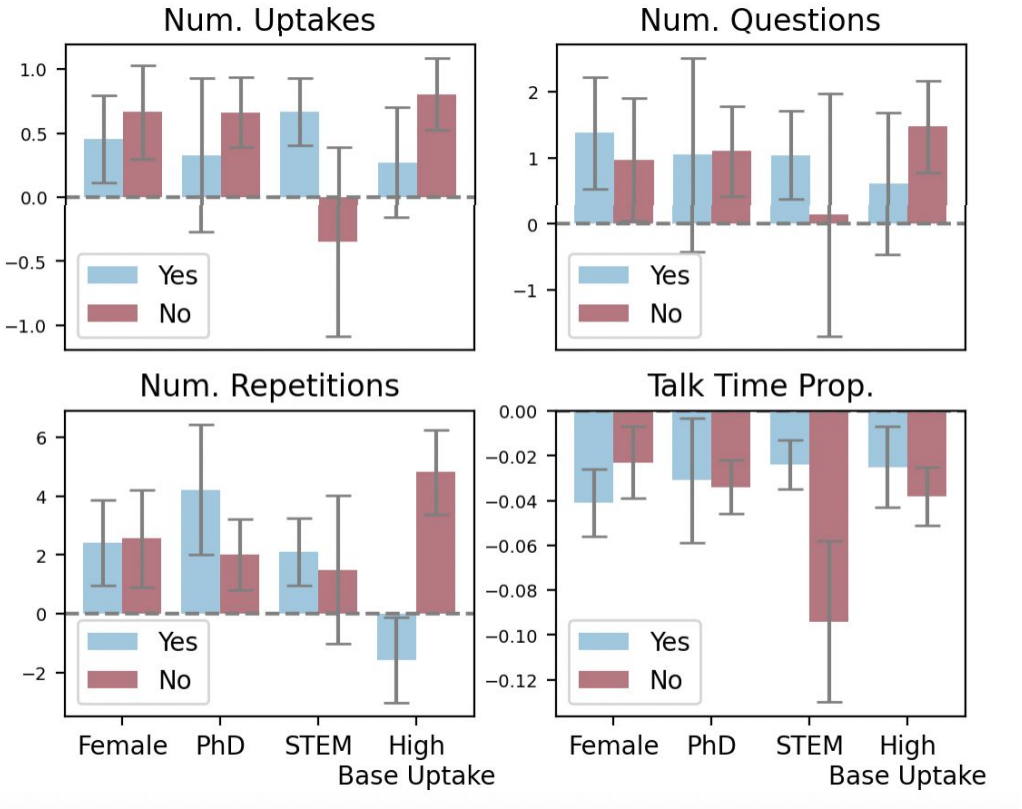
Results

RQ1 What percentage of mentors engage with the automated feedback?

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

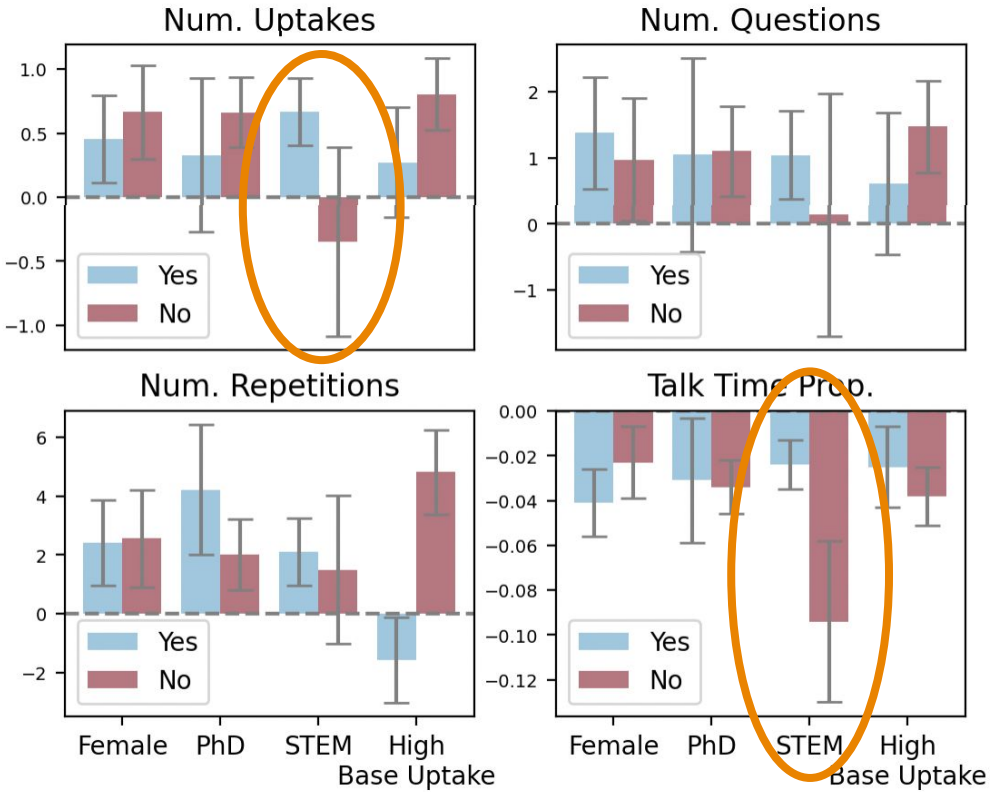
RQ3 Does the automated feedback have a differential impact on different groups of mentors?

Trends are largely consistent across mentor subgroups



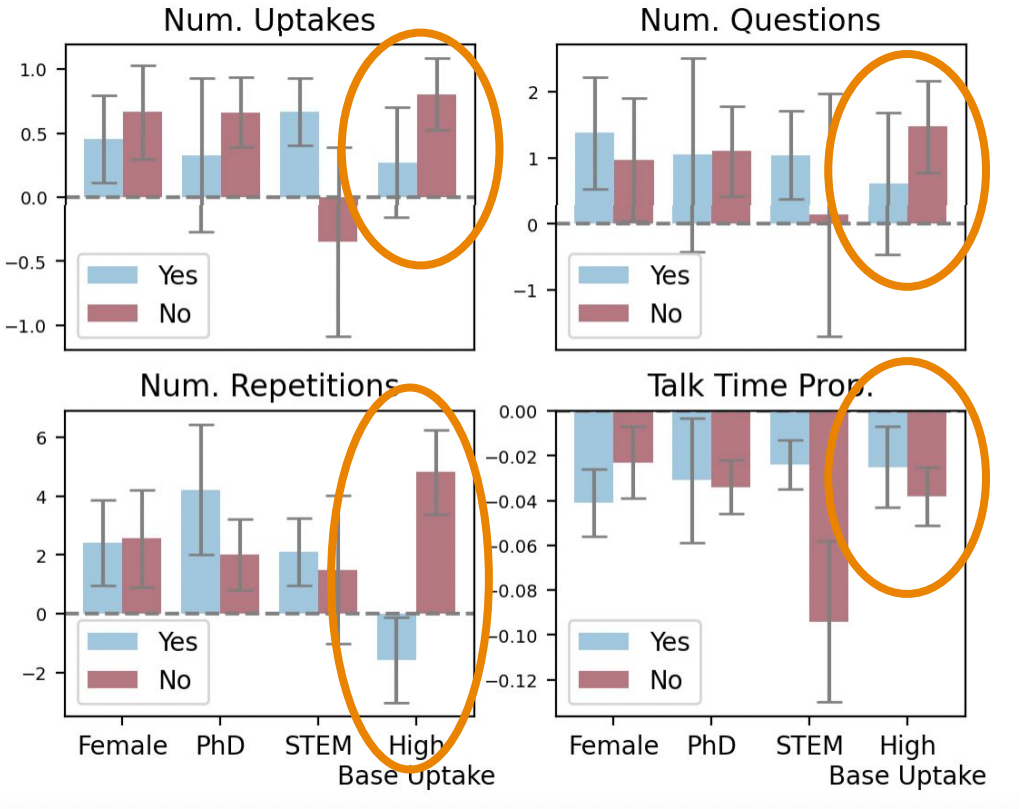
STEM vs non STEM mentors

STEM mentors increase their uptake somewhat more while non STEM mentors decrease their talk time more



Low vs high baseline uptake

Those with low baseline uptake respond better to feedback



Results

RQ1 What percentage of mentors engage with the automated feedback?

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

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

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

As a result of the feedback, …

- mentors gave 3% higher NPS scores ($p<0.1$)
- students gave 4% higher NPS scores ($p < 0.05$)
- students were 5% more relative optimism about their academic future ($p< 0.05$)
- there was no impact on mentor review scores or publication status (missing data issue).

Open questions

- How do we facilitate teachers' engagement with the feedback?
- How do we navigate the trade-off between diversity & flexibility of feedback with user-friendliness?
- How do we incorporate generative AI safely and robustly?
- How do we adapt the feedback to in person contexts?

Current & Future Work

Improve Feedback by Working Closely with Educators

Integrate Feedback into Professional Learning Frameworks

Facilitate Safe & Equitable Access

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

	(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