Text Analysis II: Measuring Instructional Practices

ISEA Session 10

Jing Liu
University of Maryland
April 4, 2025

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. Introducing the MPowering Teachers system
2. Why measuring instruction?
3. Workflow of using NLP to measure instruction
4. Case study: measuring the uptake of student ideas
5. Ongoing efforts

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

The Current System of Human Observation and Feedback

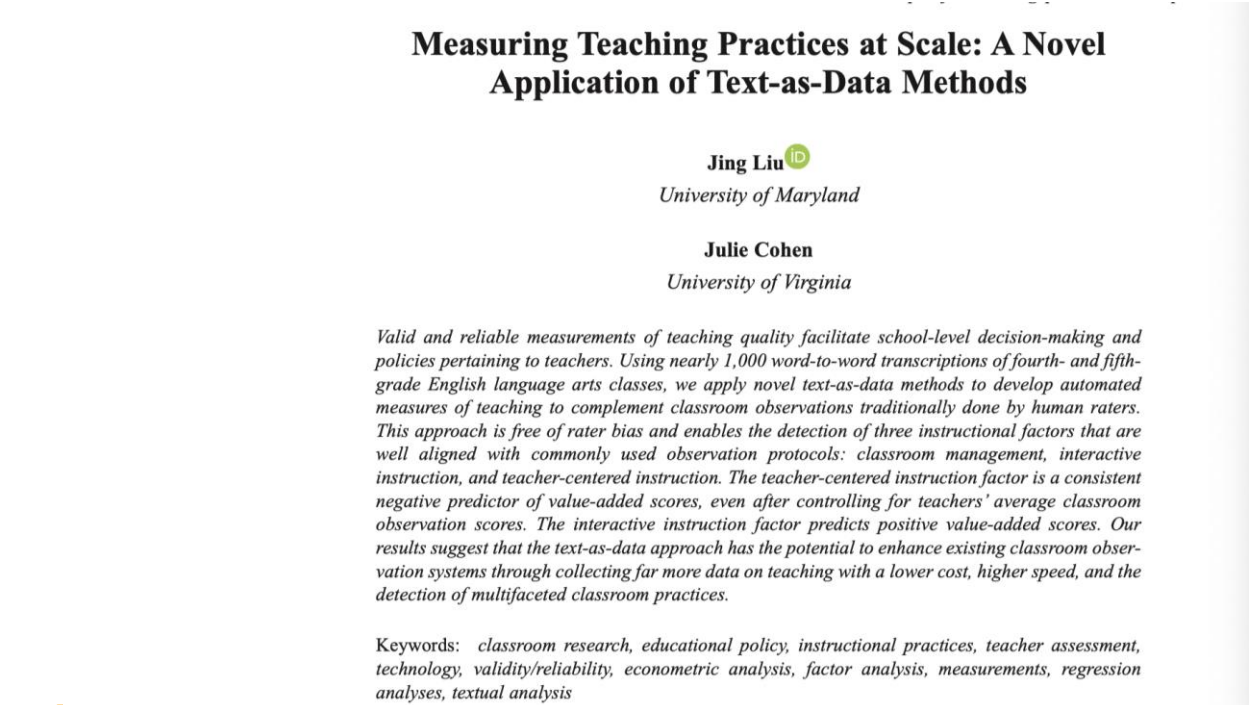
- Widely used in the US and the world to evaluate teaching practices across early childhood, K-12, and higher education (Kane & Staiger, 2012; Pianta & Hamre, 2009; Cohen & Goldhaber, 2016; Hill & Grossman, 2013)
- Resource intensive: an average public school teacher only receives formative feedback 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

Natural Language Processing (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)

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-4o; DeepSeek; Claude; etc
Pros: ● Downloadable → more transparency & control ● Needs little compute ● Can achieve similar performance to LLMs when sufficient labeled data is available ● Local deployment → much more secure	Pros: ● Very good at few shot learning ● Can be tuned with instructions ● Might be better at recognize implicit teaching strategies
Cons: ● Require more training data ● Can't be tuned with instructions or via interacting with the model	Cons: ● Most cannot be downloaded, hence privacy concerns ● Significantly higher compute resources required

What Instructional Practices to Measure?

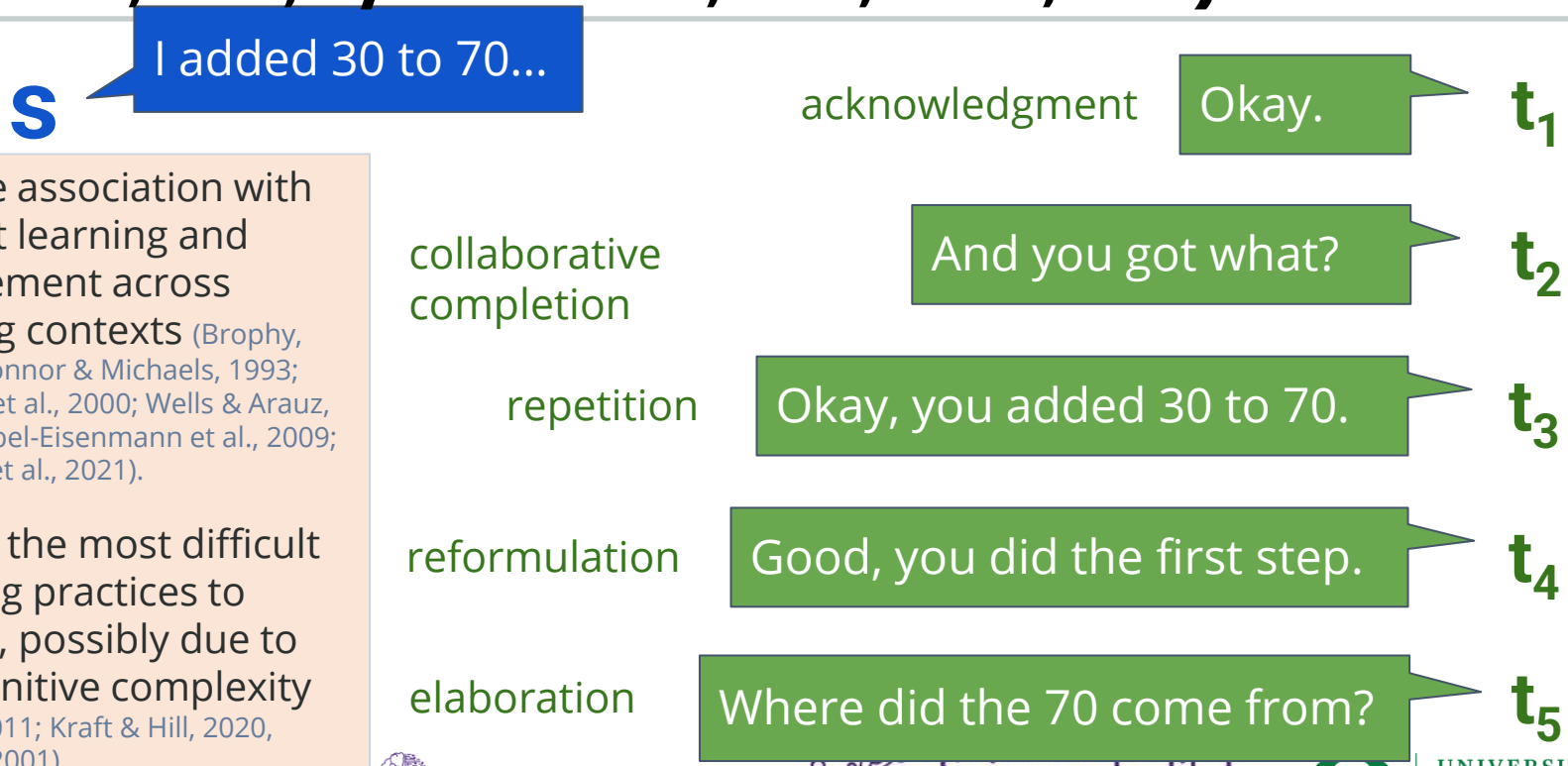
Starting with popular classroom observation tools!

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)**.

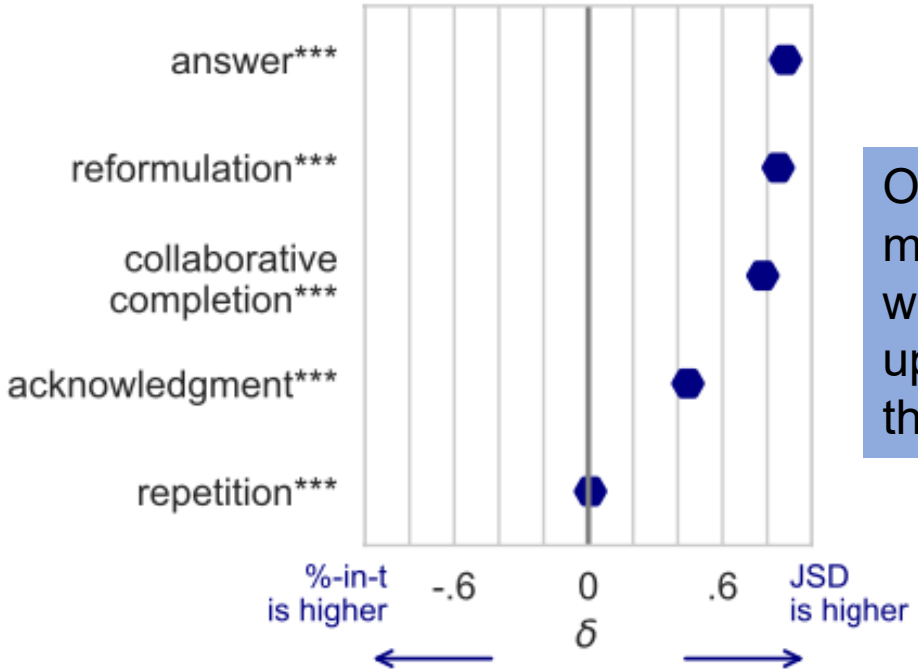
Validation Methods

- 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

Going Beyond Teachers' Uptake of Student Ideas

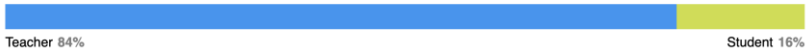
- Mathematical language (both teacher and student)
- Teacher focusing (open-ended) questions
- Student mathematical explanation and reasoning
- Classroom management and time on task
- Meta-cognitive modeling
- ……

Coming soon

- Attributing ideas to students
- Student small group productivity
- Student talk alignment with lesson objective(s)

Talk Distribution

Teacher Name has spoken 84% of the time in the class.

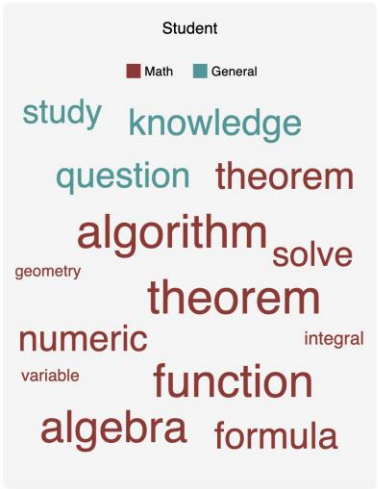


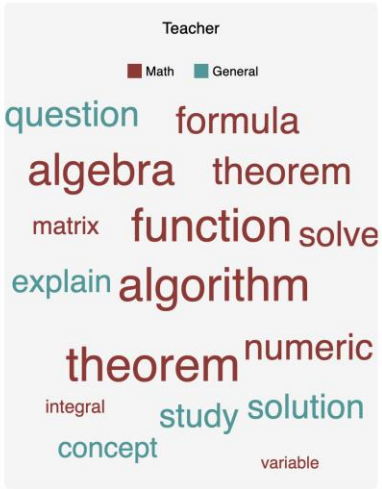
Talk Length

On an average, Teacher Name spoke 34 words continuously whereas students spoke 5 words continuously.

Top Words

Following are the commonly used words in the session.





Talk Moments Summary

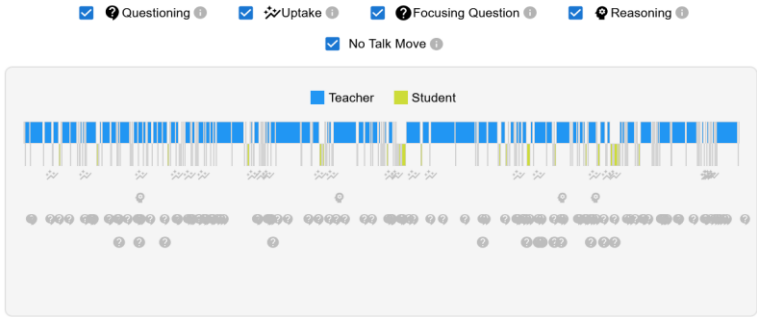
Summary of the different Talk moments observed in the class.

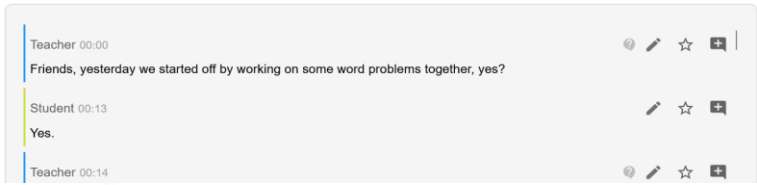
Uptake	Reasoning	Questioning	Focusing Question
24	4	111	13

Custom Moment 1	Custom Moment 2
0	0

Talk Moments

The chart below can help you explore when and how different talk moves were used in the class session.





NLP can also facilitate in-depth analysis of a variety of classroom dynamics

IES Institute of Education Sciences
AmplifyLearn.AI
UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS
UNIVERSITY OF OREGON

Educator Attention: How computational tools can systematically identify the distribution of a key resource for students

Qingyang

A Quantitative Study of Mathematical Language in Upper Elementary Classrooms

Zachary Himmelsbach, Heather C. Hill, Jing Liu, Dorottya Demszky

Sit Down Now: How Teachers' Language Reveals the Dynamics of Classroom Management Practices

Mei Tan, Dorottya Demszky

IES Institute of Education Sciences

AmplifyLearn.AI

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

UNIVERSITY OF OREGON

Bottleneck for Developing More Automated Measures

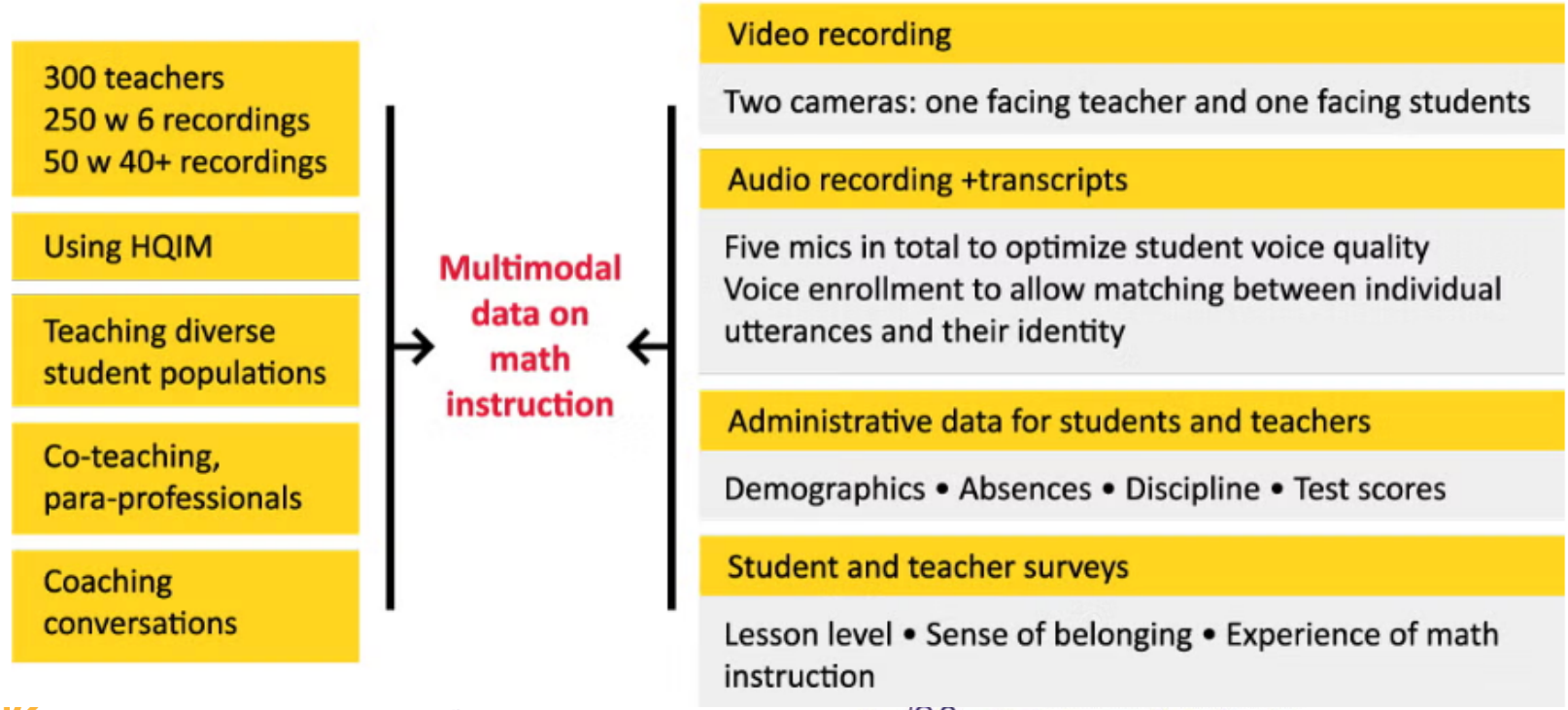
- Overall lack of data on classroom discourse
  - NCTE
  - MET
- Existing data do not have sufficient teaching practices that are high quality
  - The sparsity issue (uneven distribution of ratings)
- Quality of student speech data is quite low in existing datasets
  ⟶ key to developing student-centered measures
- Solutions
  - Better method
  - Better data

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."

Ongoing Multimodal Data Collection



Code Demo

- Edu-Convikit
  https://github.com/stanfordnlp/edu-convokit
- Funneling-focusing questions
  https://github.com/sterlingalic/funneling-focusing
- Uptake
  https://github.com/ddemszky/conversational-uptake

Assignment

Option 1: Use Edu-Convokit to analyze ncte_single_utterances.csv by using the pre-installed annotator. Conduct a descriptive analysis to answer

1) On average, what is the distribution of talk time between teachers and students?
2) Does teacher uptake of student ideas increase or decrease over the course of a lesson?
3) what lexical features separate instances of student reasoning vs. the rest of their speech?

Option 2: Use the annotation for student reasoning to train a classifier. You might want to compare the machine learning approach and an LLM-based approach to see which one works better.

I will see you at ASU GSV if you happen to go as well!

> Appendix

External Validation #1:

NCTE dataset [Kane et al., 2015]

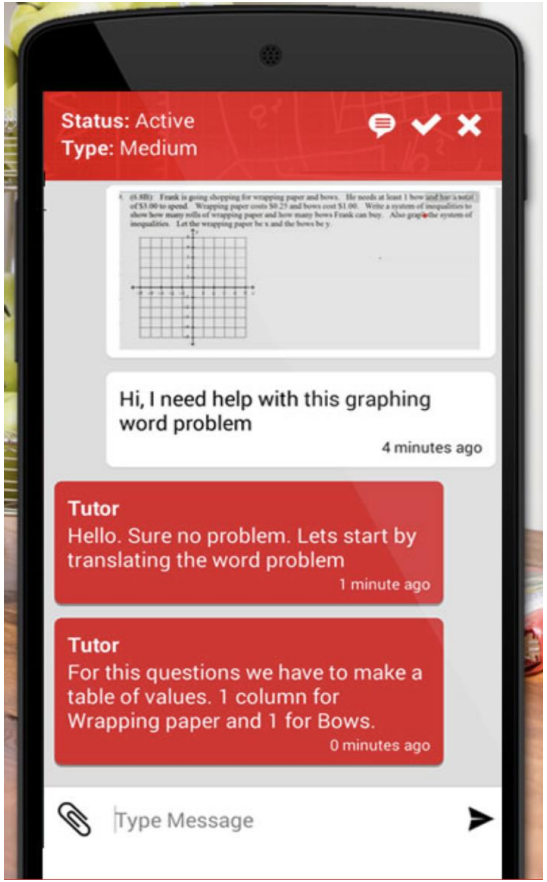
- N=55k (S, T) pairs
- elementary math classrooms
- spoken (in-person)
- whole class (20-30 students)
- external measures:
  - use of student contributions
    - **β=0.101*****
  - math instruction quality
    - **β=.091*****

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***