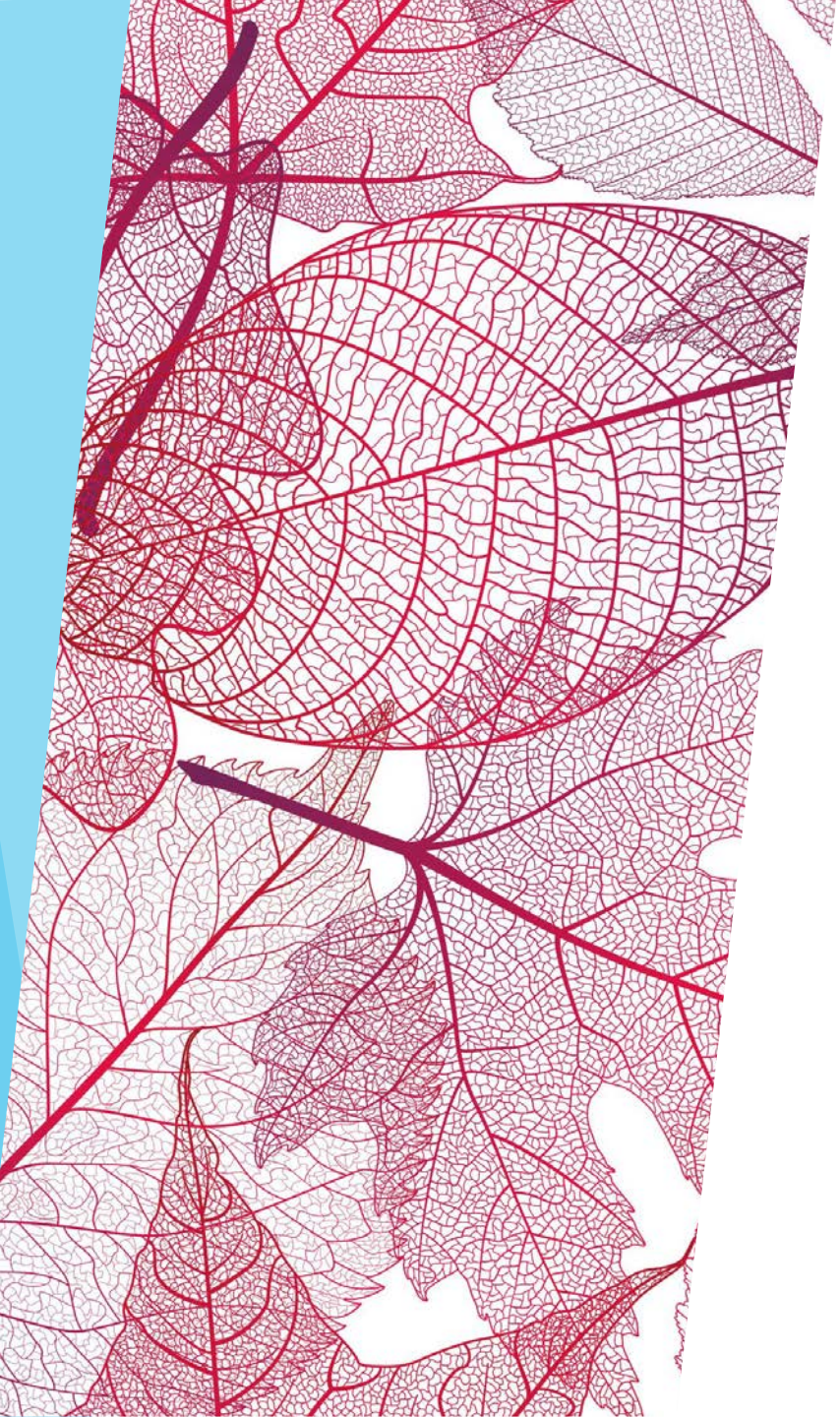
Using Video to Study Mathematics Instruction at Scale

With bonus section on the mysterious afterlife of some of that video

Heather Hill
Harvard GSE &
Annenberg Institute at Brown

This material is based upon work supported by the National Science Foundation under grant numbers 0918383 and 1348144 as well as IES grant number R305C090023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Institute for Educational Sciences.



A bit of background

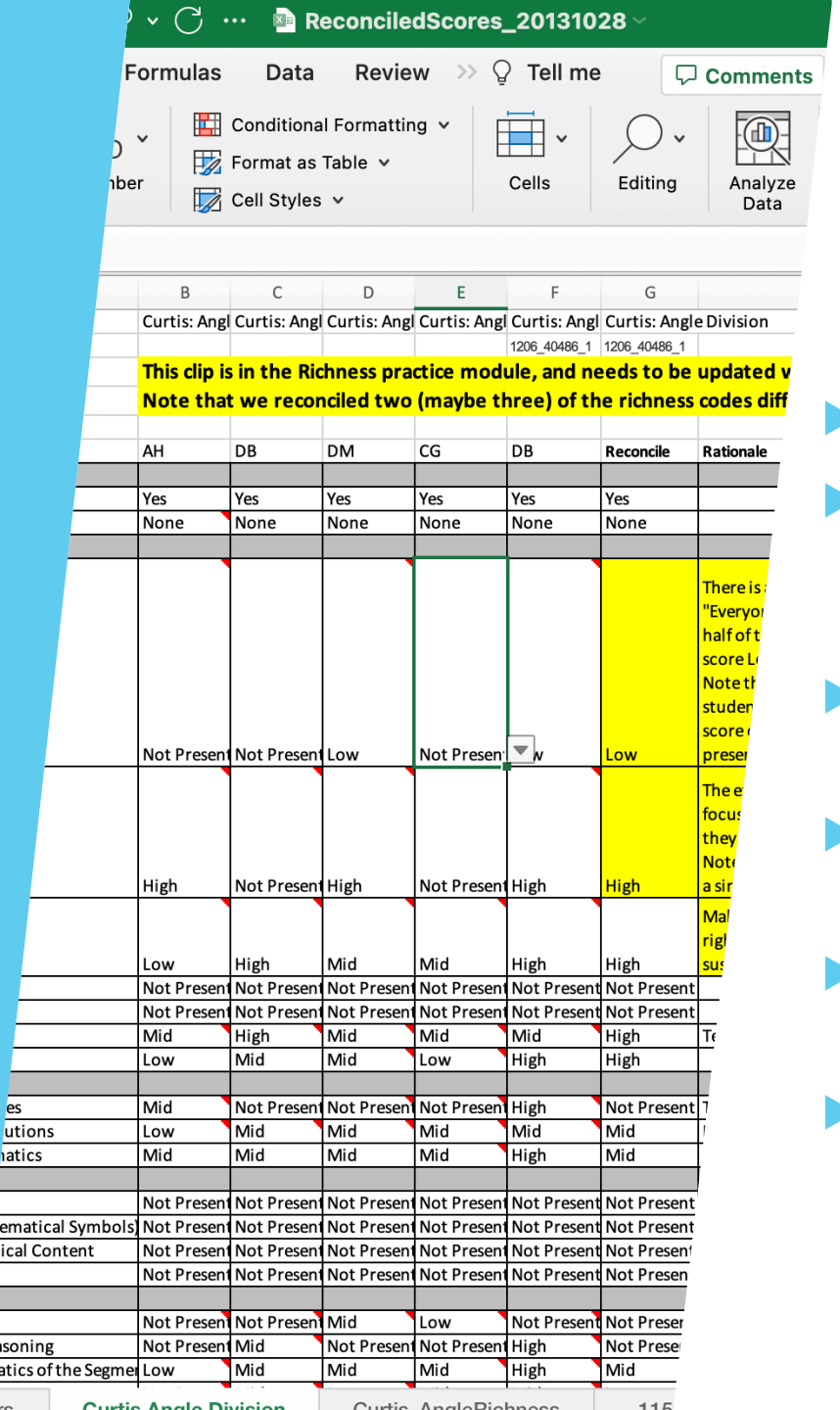
Three large-scale quantitative projects measuring teacher knowledge, math instruction, student math outcomes

Project 1: National Center for Teacher Effectiveness

- Goal: To correlate aspects of math knowledge, mathematics classroom instruction to student outcomes
- Video study design: Three videos per year x three years x ~300 teachers (~2700 videos)
  - Larger number of videos/teacher yields more accurate teacher scores
  - Focus: teachers and teacher-student interactions = board capture, high-quality audio
- Scored using the Mathematical Quality of Instruction instrument
  - Transcribed videos before scoring to improve accuracy

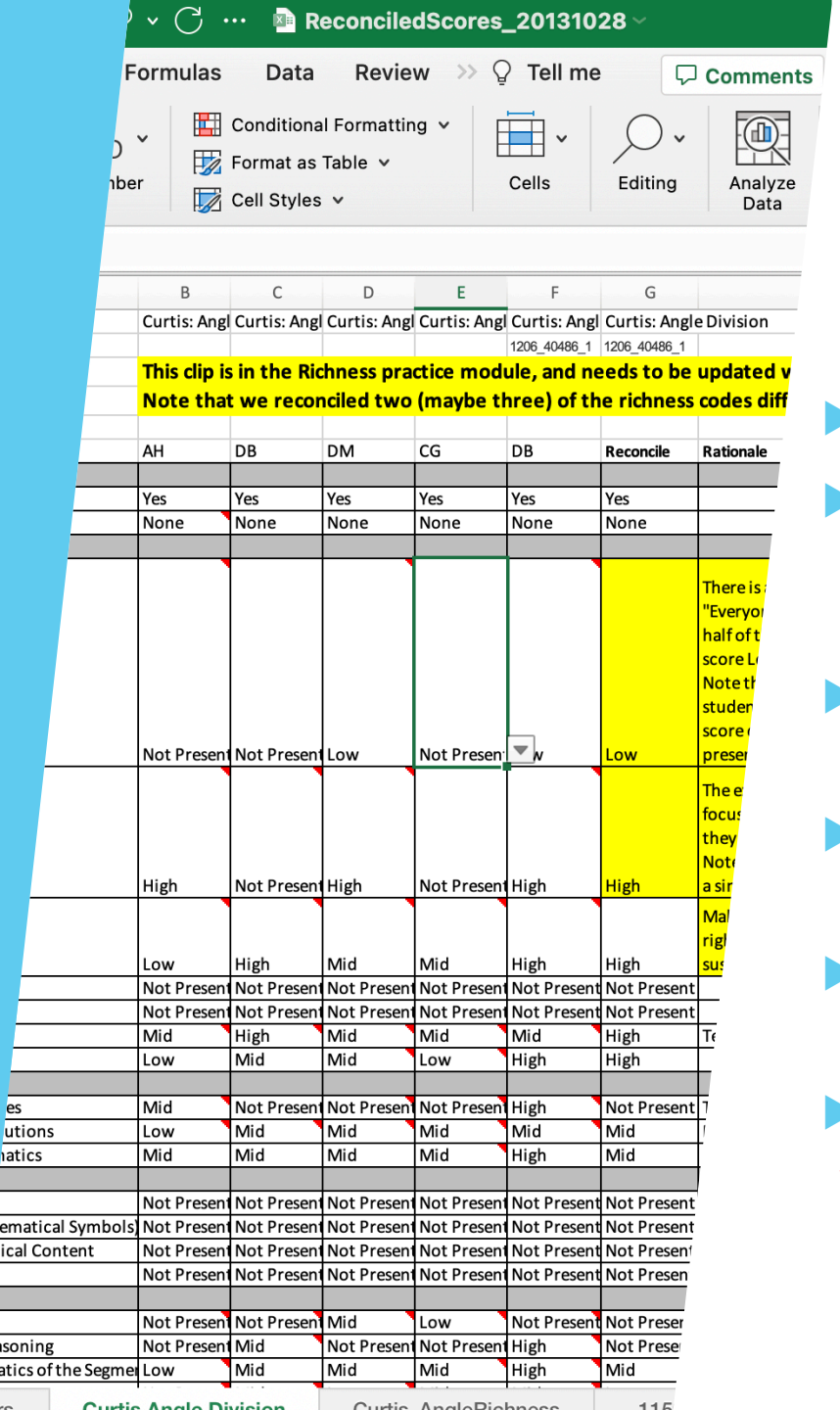
Projects 2 & 3: Impact Evals of Math PD

- Randomized experiments of PD meant to improve teacher knowledge and increase the use of student-centered instruction
- ~720 and ~625 lessons, respectively (~5-6 lessons/teacher)
- Scored with MQI
- Again focused on teachers and teacher-student interactions



Critical across all studies: Coders

- Hired large number of coders
- Pre-screened coders based on mathematical knowledge and teaching experience
- Developed online training for MQI (~20 hours)
- Developed certification assessment for MQI
- Supplied ongoing webinars to monitor coding accuracy
- For MET and NCTE, supplied dozens of master-coded videos



Critical across all studies: Coders

- Hired large number of coders
- Pre-screened coders based on mathematical knowledge and teaching experience
- Developed online training for MQI (~20 hours)
- Developed certification assessment for MQI
- Supplied ongoing webinars to monitor coding accuracy
- For NCTE, supplied dozens of master-coded videos

LOTS of master scoring

Wisdom of practice (?)

Tension	• Mechanical scoring vs. holistic scoring
Racial bias	• Campbell & Ronfeldt – small, but observable in MET data
Infrastructure	• Stuff breaks, funding sources are few, online requires care and feeding
Noise	• Observations scores are noisy due to both occasion, rater effects and error
Human scoring is expensive	• Prohibitive for many RCTs

The mysterious afterlife of the NCTE transcripts….

The NCTE Transcripts:
A Dataset of Elementary Math Classroom Transcripts

Dorottya Demszky
Stanford University
ddemszky@stanford.edu

Heather Hill
Harvard University
heather_hill@gse.harvard.edu

Abstract

Classroom discourse is a core medium of instruction – analyzing it can provide a window into teaching and learning as well as driving the development of new tools for improving instruction. We introduce the largest dataset of mathematics classroom transcripts available to researchers, and demonstrate how this data can help improve instruction. The dataset consists of 1,660 45-60 minute long 4th and 5th grade elementary mathematics observations collected by the National Center for Teacher Effectiveness (NCTE) between 2010-2013. The

major foundations (e.g., Gates, Spencer[2]). A key step in this direction is to measure and facilitate the use of effective mathematics teaching practices, an effort that draws on a long history of research (e.g., Brophy, 1984; Sedova et al., 2019). Instructional measurement has traditionally relied on resource-intensive classroom observation. Recent natural language processing models, trained on manually scored classroom transcripts, enable measuring effective instructional practices in scalable and adaptable ways (Kelly et al., 2018; Suresh et al., 2019; Demszky et al., 2021b; Alic et al., 2022; Hunk-

Natural Language Processing Approaches to Measuring Math Instruction

- Goal: Develop automated scoring algorithms
- Use classroom transcripts, parsed into teacher-student utterance pairs
- Use established NLP packages where commonalities with key classroom features
  - Teacher uptake of student ideas, student reasoning
- Use human labeling of subset of data to “train” AI to detect patterns in conversations
  - Ultimately what we have used to build measures

What are we measuring?

NLP measures	Correlation to value-added*	Correlation to same-item human-rated MQI scores
Students on task	0.04~	**0.02***
Uptake of student contribution	**0.12***	**0.11***
Focusing questions	**0.23***	**0.23***
Student reasoning	**0.19***	**0.31***

$p < 0.05$

In regressions controlling for teacher gender, race, classroom composition (FRPL, etc)

NLP TBD

- High-quality audio capture
- Dialect bias
- “Big brother” guardrails
- Better for teacher feedback?

Any questions?

heather_hill@harvard.edu