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 teacherstudent 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

	• G ·	••• 🖻 R	econcile	dScores ₋	_201310	28~	
F	ormulas	Data	Revie	w >> Ç	Tell me	P	Comm
		Conditiona	l Formattir	g v F		\bigcirc \checkmark	
C		Format as	Table 🗸			\mathcal{P}	
hbe	r 🛄				Cells	Editing	Anal
		Cell Styles	~				Da
	В	С	D	E	F	G	
	Curtis: Ang	Curtis: Ang	Curtis: Ang	Curtis: Angl	Curtis: Angl	Curtis: Angle	e Division
			_		1206_40486_1	1206_40486_1	
	This clip is	s in the Rid	hness pra	ctice mod	ule, and ne	eds to be	update
	Note that	<mark>t we reco</mark> r	nciled two	(maybe th	nree) of th	e richness	codes d
	AU	DP	DM	<u></u>	DB	Decencile	Patianala
	АП	06				Reconcile	Rationale
	Yes	Yes	Yes	Yes	Yes	Yes	
	None	None	None	None	None	None	
							Thoras in
							"Everyou
							halfoft
							scorel
							Noteth
							studen
							score
	Not Present	Not Present	Low	Not Presen	▼ _N	Low	preser
							Thee
							focus
							they
							Note
	High	Not Present	High	Not Present	High	High	a sir
							Mal
							rig
	Low	High	Mid	Mid	High	High	sus
	Not Present	Not Presen	Not Present	Not Present	Not Present	Not Present	
	Not Present	Not Present	Not Present	Not Present	Not Present	Not Present	
	Mid	High	Mid	Mid	Mid	High	Te
	Low	Mid	Mid	Low	High	High	-
	Mid	Not Droconi	Not Drocon	Not Drocon	High	Not Drocont	5
		Mid	Mid	Mid	Mid	Mid	;
	Mid	Mid	Mid	Mid	High	Mid	ŕ
	ivin d	ivii a	i i i i i i i i i i i i i i i i i i i	1711G	111811	IVIIG	
	Not Present	Not Present	Not Present	Not Present	Not Present	Not Present	
vmbols	Not Present	Not Present	Not Present	Not Present	Not Present	Not Present	
,	Not Present	Not Presen	Not Present	Not Present	Not Present	Not Present	
nt		Not Present	Not Present	Not Present	Not Present	Not Presen	
nt	Not Present	Hot I reserve					
nt	Not Present						
ent	Not Present	Not Presen	Mid	Low	Not Present	Not Preser	
ent	Not Present Not Present Not Present	Not Presen Mid	Mid Not Present	Low Not Present	Not Present High	Not Preser Not Prese	

utio

natics

ical

isonii atics

Critical across al 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

F	ormulas	Data	Revie	w »» Ç	> Tell me		Comm
	•	Conditiona	l Formattin	g v F		\bigcirc	
C	E E	Format as	Table 🗸			\sum	
hbei	r 📖				Cells	Editing	Analy
		Cell Styles	~				Dat
	P	6	D	F	E	G	
	Curtin Arrel	Currentian Arrian	Curtin An al	L Cumbias Ameri	Cuntin An el	Custia: An al	District
	Curtis: Angi	Curtis: Angi	Curtis: Angi	Curtis: Angi	1206 40496 1	1206 40496 1	Division
	This clip is Note that	in the Ric twe recor	hness pra nciled two	ctice mod (maybe th	ule, and ne nree) of th	eds to be e richness	updatec codes di
	AH	DB	DM	CG	DB	Reconcile	Rationale
	Voc	Voc	Vec	Voc	Vec	Voc	
	Yes	Yes	Yes	Yes	Yes	Yes	
	None	None	None	None	None	None	
	Not Present High	Not Present	Low High	Not Presen	▼ w High	Low	half of t score L Note th studer score i preser The e focus they Note a sir Mal
	Low	High	Mid	Mid	High	High	rigi su:
	Not Present	Not Present	Not Present	Not Present	Not Present	Not Present	
	Not Present	Not Present	Not Present	Not Present	Not Present	Not Present	-
		High	Mid	Mid	Wid	High	16
	LOW	wiid	WIId	LOW	High	пign	-
	Mid	Not Present	Not Present	Not Present	High	Not Present	7
	Low	Mid	Mid	Mid	Mid	Mid	ī
	Mid	Mid	Mid	Mid	High	Mid	t
							t
	Not Present	Not Present	Not Present	Not Present	Not Present	Not Present	
mbols)	Not Present	Not Present	Not Present	Not Present	Not Present	Not Present	
t	Not Present	Not Present	Not Present	Not Present	Not Present	Not Present	
	Not Present	Not Present	Not Present	Not Present	Not Present	Not Presen	
	Not Present	Not Presen	Mid	Low	Not Present	Not Preser	
	Not Present Not Present	Not Presen Mid	Mid Not Present	Low Not Present	Not Present High	Not Preser Not Preser	

es utio

natic

emat ical (

isonii atics

Critical across al studies: Coders

- Hired large number of coders
- Pre-screened coders based or mathematical knowledge and teaching experience
- Developed online train(0) fo MQI (~20 hours)
- Developed certileation assessment for MQI
- Supplied ongoing webinars to monit@coding accuracy
 - For NCTE, supplied dozens of waster-coded videos

Wisdom of practice (?)

Tension	 Mechanical scoring vs. holistic scoring
	Comphall 9 Donfaldt anall but
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, Spencer²). 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 resourceintensive 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?

NI P measures	Correlation to	Correlation to same-item human-rated MQI
NEI MEASUICS	value-added	300103
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