DO THE NUMBERS ADD UP? EXAMINING RELATIONSHIPS BETWEEN TEACHER VALUE-ADDED SCORES AND ALTERNATIVE INDICATORS OF TEACHER QUALITY

CORINNE HERLIHY
TUCSON, AZ
1-10-2014

Johanna Barmore
Mary Beisiegel
David Blazar
David Braslow
Lindsay Brown
Kate Buckley
Mark Chin
Julie Cohen
Douglas Corey
Claire C. Gogolen
Pam Grossman
Heather C. Hill
Andrea Humez
Erica Litke
Kathleen Lynch
Dan McGinn
Samantha Rabinowicz
Sara Roesler
Matthew Ronfeldt
Lucas Salzman
Douglas Staiger

National Center for Teacher Effectiveness

Center for Education Policy Research
HARVARD UNIVERSITY

ies INSTITUTE OF EDUCATION SCIENCES

CAN WE EXPLAIN VALUE-ADDED SCORES VIA TEACHER OR CLASSROOM CHARACTERISTICS?

- Traditional literature: Variable
  - Process product literature: 0.1-0.4
  - Principal use of performance rubrics: 0.30-.40 (Jacob & Lefgren, 2004; Kimball et al., 2004)
  - More modern studies with VAM scores as outcomes
    - Hill, Rowan & Ball (2005): Mathematical knowledge for teaching, small effect
    - Sadler et al. (2013): Teachers' knowledge of students' thinking
    - Bell et al. (2012): CLASS 0.3ish
    - Hill, Kapitula & Umland (2011): MQI 0.3-0.6ish
    - Schacter & Thum (2004): 0.6
    - Grossman et al. (2012), Pianta et al., (2008), Stronge (2011): Hard to tell, but not large effects
- *Considerable variation in correlations*
  - *Why??*
- What teaching characteristics explain value-added scores?
  - Explicit, organized instruction (process-product literature; Stronge, 2011; Grossman et al., 2012)
  - Classroom climate (Pianta et al., 2008)
  - Content-specific aspects of instruction (Hill, Kapitula, Umland, 2011; Grossman et al., 2012)
  - Inquiry?
  - Need exploratory research; can inform practice and improvement

WHAT WE DON'T KNOW

- The extent to which these correlations result from choices made during model specification process (of either VAMs or classroom indicators)
  - Using Validity Criteria to Enable Model Selection: An Exploratory Analysis (Chin, Hill, McGinn, Staiger, & Buckley)
- The extent to which these correlations vary by district or by test
  - How Well Do Teacher Observations Predict Value-Added? Exploring Variability Across Districts (Grossman, Cohen, Ronfeldt, Brown, Lynch, & Chin)
- Characteristics of instruction in high and low-VAM teacher classrooms
  - Examining High and Low Value-Added Mathematics Instruction: Can Expert Observers Tell the Difference? (Hill, Litke, Humez, Blazar, Corey, Barmore, Chin, Beisiegel, Salzman, Roesler, Braslow, & Rabinowicz)

DATASET - NCTE

- National Center for Teacher Effectiveness main study
  - Over 300 fourth and fifth grade teachers
  - Value-Added scores for teachers
    - Typical within-district HLM model (student prior achievement, demographics, peer & cohort effects)
    - State standardized test scores for ALL students from up to 4 years
    - Alternative test scores (fall & spring) for NCTE students for up to 2 years
  - Two years of videotaped lessons (up to 6 lessons per teacher)
    - Coded with the Mathematical Quality of Instruction instrument (MQI) and Classroom Assessment Scoring System (CLASS)
  - Other alternative indicators
    - Teacher knowledge, student perception surveys

USING VALIDITY CRITERIA TO ENABLE MODEL SELECTION: AN EXPLORATORY ANALYSIS

Mark Chin
Heather C. Hill
Dan McGinn
Doug Staiger
Katie Buckley

MOTIVATION

- Terminology: Test-based accountability metrics (TBAMs)
  - Value-added scores
  - Student growth percentile scores
- Little consensus across districts, states, and research organizations how best to specify TBAM models (Goldhaber & Theobald, 2012)
  - Problematic because ranking of teacher TBAM not preserved from model to model
  - Student demographic? Classroom composition? School fixed effects?
  - Single year? Multi year?
- Our proposal: Consider TBAM alignment with alternative, non-test-based measures of teacher and teaching effectiveness in deciding on what TBAM model to use

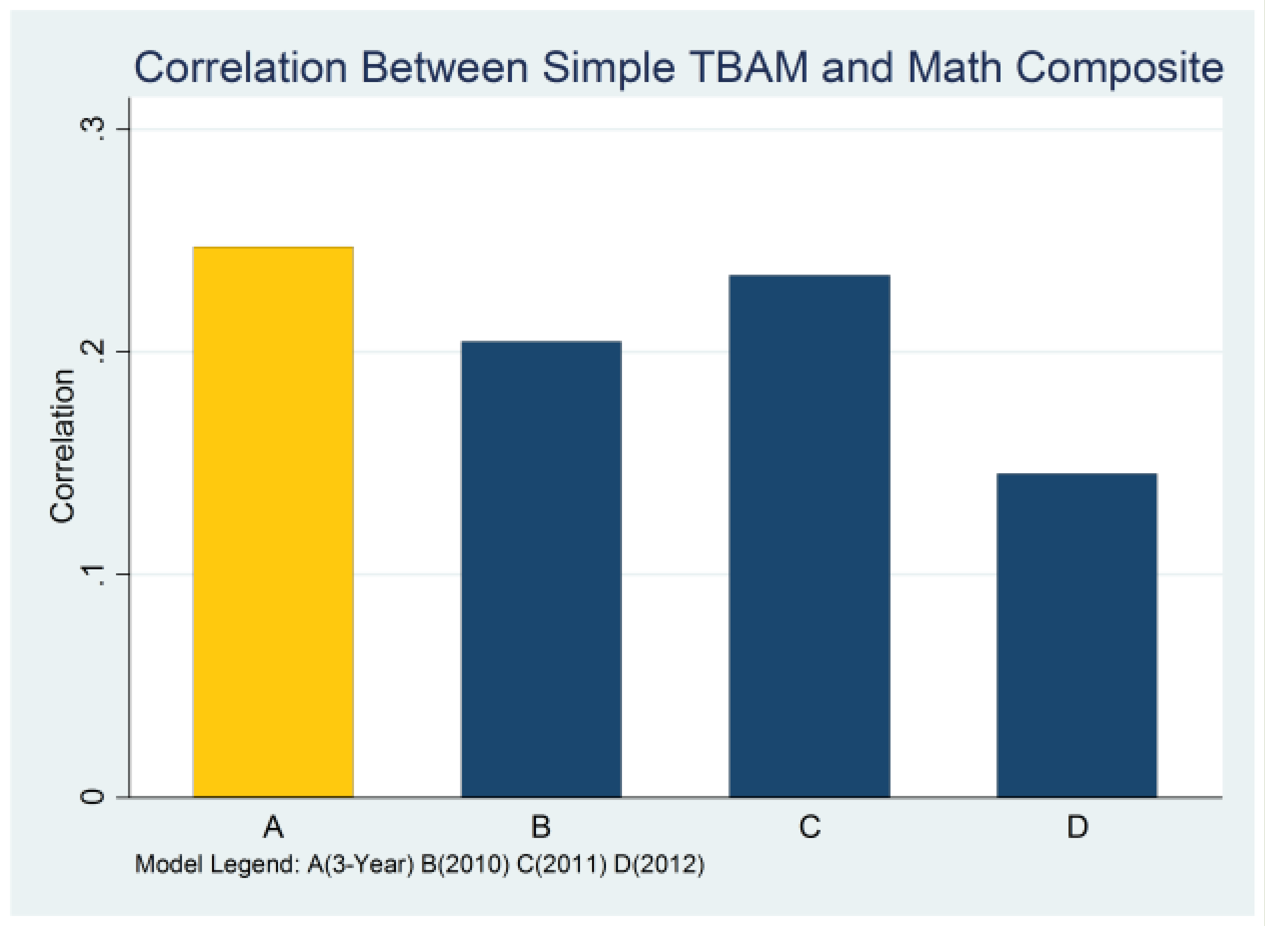
TBAM MODELS CONSIDERED

- Simple value-added model (VAM)
  - Student prior achievement
  - Student demographics
- Peer VAM
  - Simple VAM
  - Classroom aggregates
- School Fixed Effect VAM
  - Simple VAM
  - School Fixed Effects
- Student Growth Percentiles (SGPs)
  - Student prior achievement
  - Quantile regression

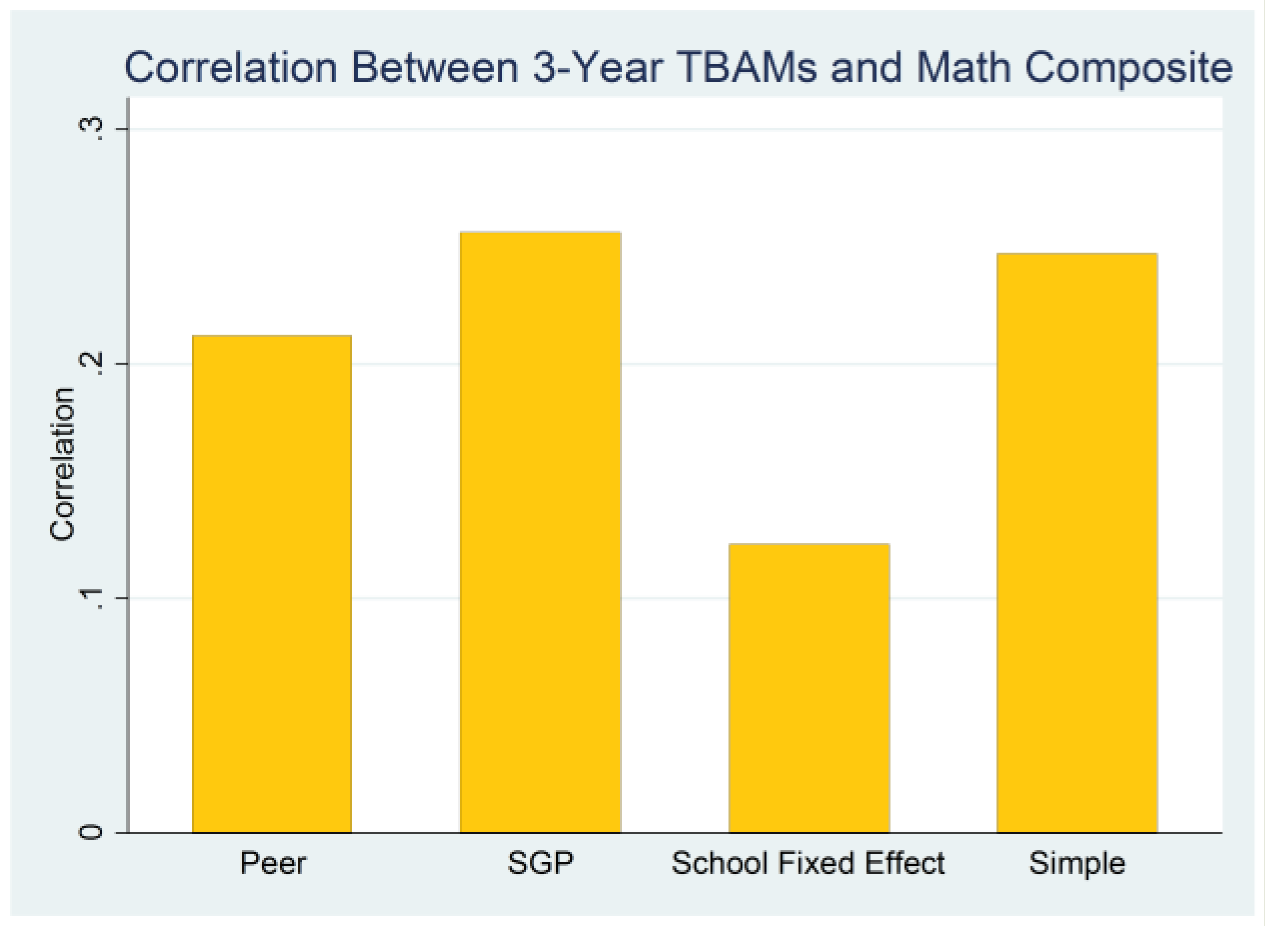
ALTERNATIVE MEASURES CONSIDERED

- Math Composite
  - Instruction
    - Mathematical richness
    - Mathematical errors and imprecisions
  - Math Knowledge
    - General
    - Specific to teaching
    - Knowledge of students’ performance
- Classroom Interaction Composite
  - Student perceptions
  - Instruction
    - Classroom Organization
    - Instructional and Emotional Support

CORRELATIONS BETWEEN TBAMS AND THE MATH COMPOSITE



CORRELATIONS BETWEEN 3-YEAR TBAMS AND THE MATH COMPOSITE



CONCLUSIONS

- Districts and researchers should strongly consider using 3-year TBAMs instead of 1-Year TBAMs
    - 1-year TBAM correlations with alternative indicator vary in their magnitude and significance depending on model and year
        - Researchers who use 1-year estimates may find differing results in their analyses depending on year of study
        - Districts will have different evaluations of teacher effectiveness for the same teacher depending on the year in consideration
    - 3-year TBAMs more strongly correlated to non-test-based measures
- Less controlled TBAM models tend to more strongly correlate to alternative indicators of teacher effectiveness
        - Variability in analyses in research or evaluations of teachers may be due to model choice – for a district or research organization who seeks simply the most alignment with alternative non-test-based indicators of effective teachers, choose the simple or SGP model

HOW WELL DO TEACHER OBSERVATIONS PREDICT VALUE-ADDED? EXPLORING VARIABILITY ACROSS DISTRICTS

Pam Grossman
Julie Cohen
Matthew Ronfeldt
Lindsay Brown
Kathleen Lynch
Mark Chin
David Blazar

RESEARCH QUESTIONS

- Do observational instruments predict student achievement equally well across different tests and district/state contexts?
- If correlations vary across districts or tests, can we identify factors that explain this variability?

CURRENT PRESENTATION: 2 STUDIES

	Study 1: Math	**Study 2: ELA**
# Districts	5 (in 4 states)	6 (in 6 states)
# Teachers	298	893
Grades	4th and 5th	4th-8th
Uniform test across districts?	NCTE	SAT-9
Tests that differed by district/state?	State assessments	State assessments
Uniform observation instrument across districts?	Mathematical Quality of Instruction (MQI)	Protocol for Language Arts Teaching Observation (PLATO)

HYPOTHESIS TESTING

- The NCTE tests we administered are CONSTANT from state to state
  - The relationship of MQI to student achievement on this alternative test should NOT vary between state to state, or district to district
- State tests differ from one another
  - The relationship of MQI to student achievement on state tests may vary from state to state
  - The relationship of MQI to student achievement on the test should NOT vary between districts within the same state (who take the same test)

STATISTICAL DIFFERENCE FROM DISTRICT TO DISTRICT?

Wald Test Results – Testing MQI Regression Coefficients on NCTE Student Achievement

		District B vs. District X				District D vs. District X			District G vs. District X		District N vs. District X
MQI Code	All = Beta	D	G	N	R	G	N	R	N	R	R
Richness											
Working with Students											X
Errors and Imprecision											
Common Core Student Practices											
Lesson-Level MQI											
Guess at Typical MQI											

Wald Test Results – Testing MQI Regression Coefficients on State Student Achievement

		District B vs. District X				District D vs. District X			District G vs. District X		District N vs. District X
MQI Code	All = Beta	D	G	N	R	G	N	R	N	R	R
Richness			X							X	
Working with Students	X	X	X					X		X	
Errors and Imprecision		X	X								
Common Core Student Practices											
Lesson-Level MQI	X		X							X	
Guess at Typical MQI	X	X	X							X	

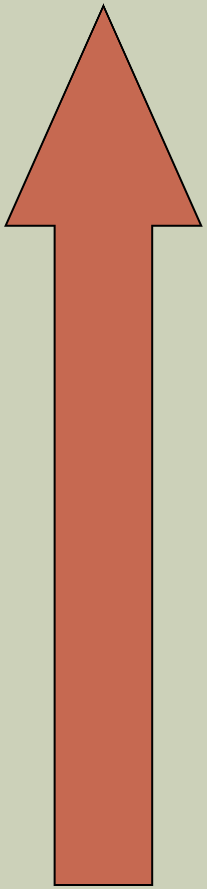
WHAT FACTORS CONTRIBUTE TO THESE DIFFERENCES?

- Exploring 2 possible factors:

  1. Tests’ Cognitive Demand

  2. Tests’ Item Formats

TESTS' COGNITIVE DEMAND
(SEC FRAMEWORK, PORTER, 2002)



5. Conjecture/generalize/prove

4. Solve non-routine problems

3. Communicate understanding

2. Perform procedures

1. Memorize

TESTS' COGNITIVE DEMAND
(SEC FRAMEWORK, PORTER, 2002)

Test	Mean	SD
Districts B & R	2.36	0.86
District D	2.13	0.69
District G	2.00	0.66
District N	2.04	0.81

TESTS’ ITEM FORMATS (AERA/NCME, 1999)

	Percent of Items		
Test	Multiple Choice	Short Answer	Open-Ended
Districts B & R	64	12	24
District D	86	12	2
District G	100	0	0
District N	100	0	0

CONCLUSIONS

- Relationships between teachers’ value-added and instructional quality vary by district
- Why this variability?
  - ‘Match’ between content of observational instrument and state assessment?
  - Cognitive demand

POLICY IMPLICATIONS

- Districts must think seriously about:
  - the alignment of the components of their evaluation system
  - the student outcomes they value and how those outcomes are measured
- Observation protocols may be better predictors for rigor of CCSS than some state VAM.

EXAMINING HIGH AND LOW VALUE-ADDED MATHEMATICS INSTRUCTION: CAN EXPERT OBSERVERS TELL THE DIFFERENCE?

Heather C. Hill
David Blazar
Andrea Humez
Erica Litke
Mary Beisiegel
Johanna Barmore
Mark Chin
Douglas Corey
Sara Roesler
Lucas Salzman
David Braslow
Samantha Rabinowicz

MOTIVATION AND RESEARCH QUESTIONS

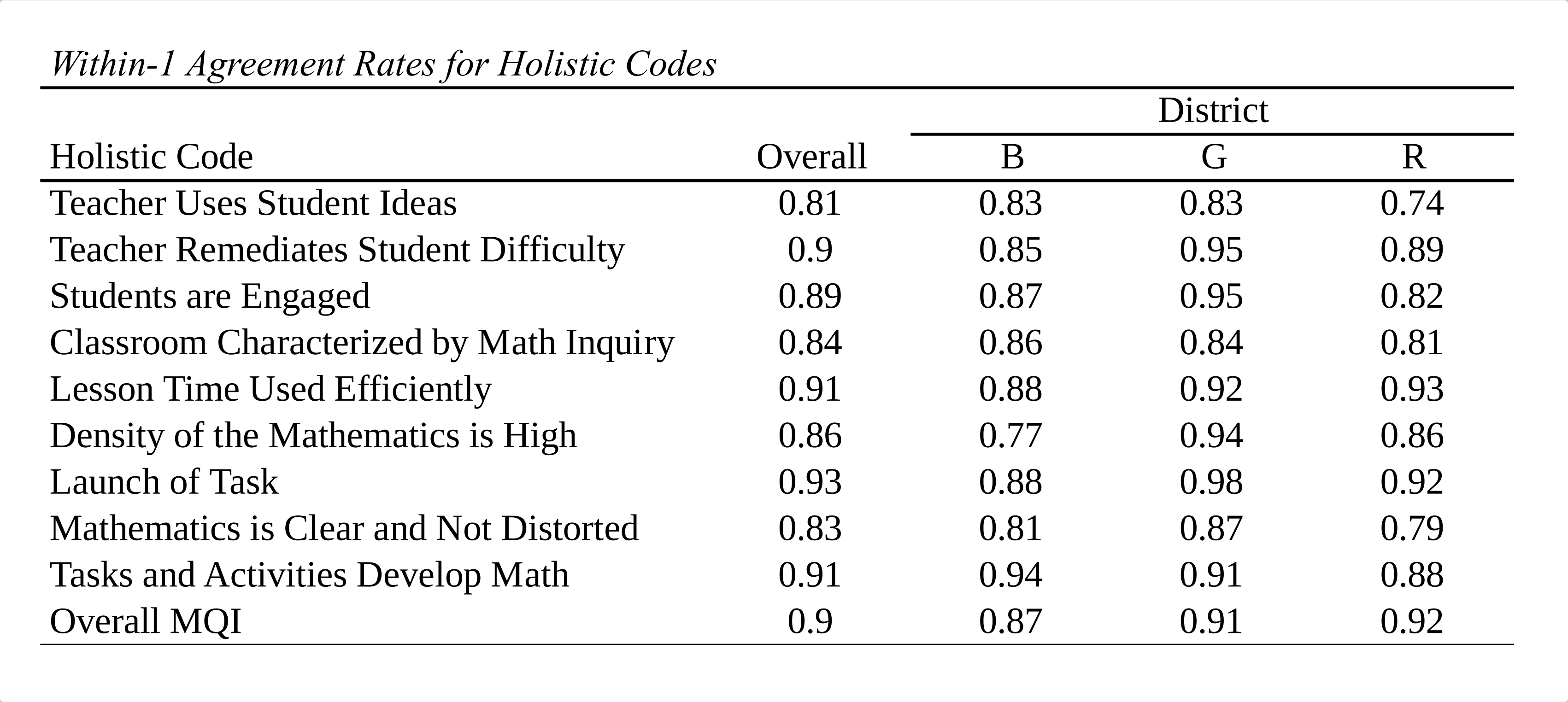
- What is the degree of convergence between observers' impressions of instruction and teachers' value-added scores?
  - Jacob & Lefgren, 2008; Stronge, Grant, & Ward, 2011
- Are there a set of instructional practices that consistently characterize high but not low value-added teachers' classrooms, and vice versa?
  - Bell et al, 2012; Grossman et al, 2010; Tyler, Taylor, Kane, & Wooten, 2010

DATA ANALYSIS

- Select sample
  - Rank all teachers in 3 districts ("B", "G", and "R") on a value-added model with three years of test-score data
  - Randomly select 3 teachers with video data from the top, middle, and bottom quintiles of value-added scores
- Watch instruction
  - Groups of 4 raters blind to value-added category watch ~6 lessons for each of 9 teachers in assigned district
  - Assess mathematics-specific and general instructional practices through memos and whole-lesson codes generated from exploratory analyses and memos
- Rank all teachers from low (1) to high (9)

RESULTS: RQ1 – CONVERGENCE BETWEEN OBSERVERS' IMPRESSIONS OF INSTRUCTION AND VALUE-ADDED SCORES?

- Raters tend to agree about the quality of instruction that they observe in lessons.

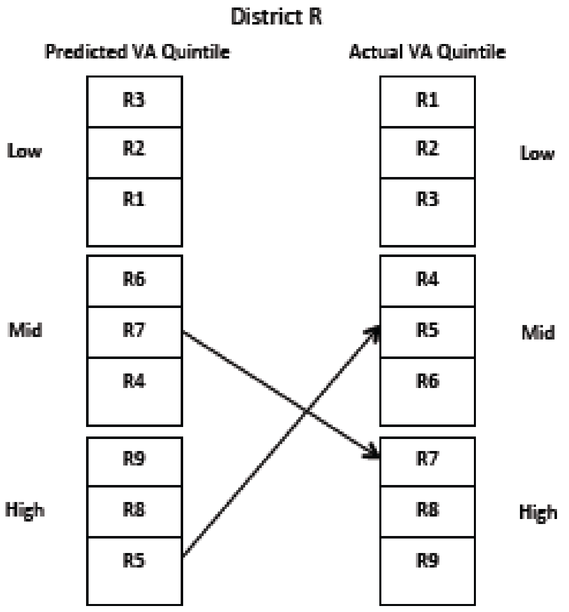


Within-1 Agreement Rates for Holistic Codes

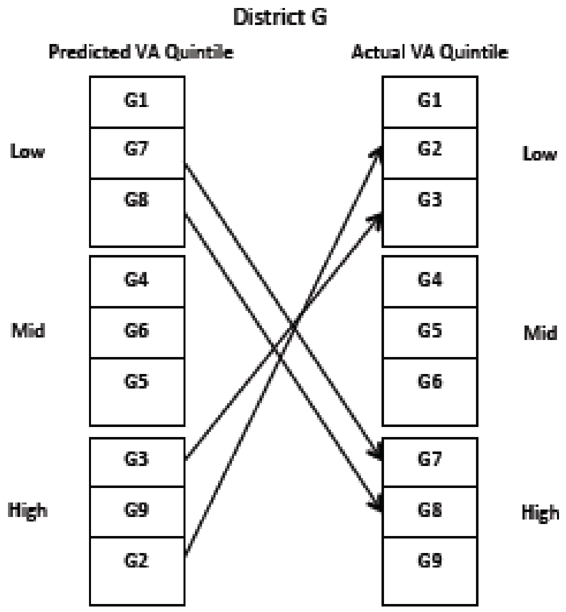
		District		
Holistic Code	Overall	B	G	R
Teacher Uses Student Ideas	0.81	0.83	0.83	0.74
Teacher Remediates Student Difficulty	0.9	0.85	0.95	0.89
Students are Engaged	0.89	0.87	0.95	0.82
Classroom Characterized by Math Inquiry	0.84	0.86	0.84	0.81
Lesson Time Used Efficiently	0.91	0.88	0.92	0.93
Density of the Mathematics is High	0.86	0.77	0.94	0.86
Launch of Task	0.93	0.88	0.98	0.92
Mathematics is Clear and Not Distorted	0.83	0.81	0.87	0.79
Tasks and Activities Develop Math	0.91	0.94	0.91	0.88
Overall MQI	0.9	0.87	0.91	0.92

RESULTS: RQ1 CONT'D

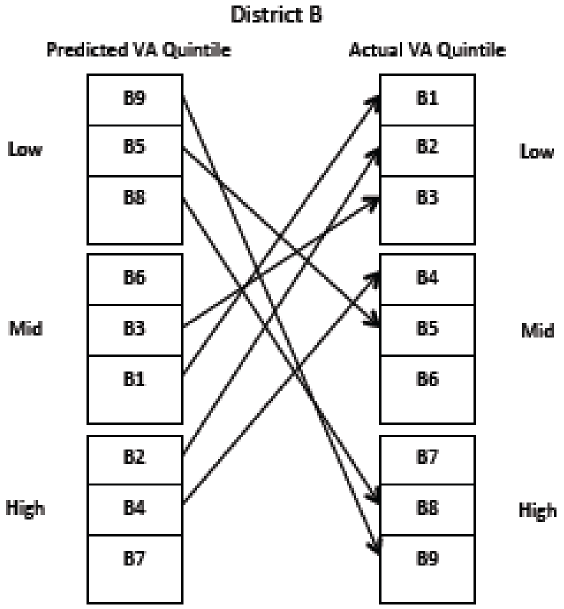
- However, ability to predict value-added from instructional quality varies by district.



Off by 1+ Quintile: 22% (of 9 predictions)
Off by 2 Quintile: 0% (of 6 predictions)



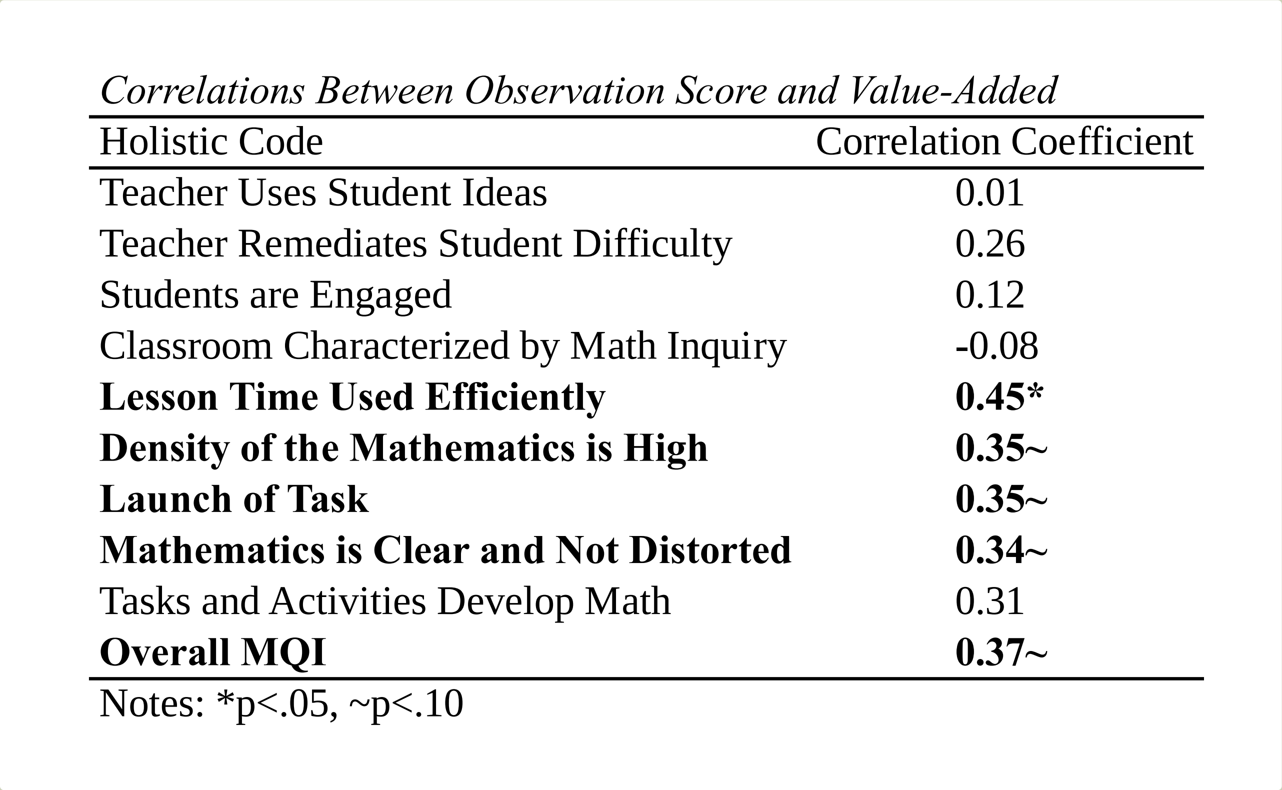
Off by 1+ Quintile: 44%
Off by 2 Quintile: 66%



Off by 1+ Quintile: 77%
Off by 2 Quintile: 50%

RESULTS: RQ2 – WHAT INSTRUCTIONAL FEATURES CHARACTERIZE HIGH- OR LOW-VAM TEACHERS

- Quantitatively, observe some associations of medium strength between value-added and instructional quality focused on classroom organization.



Correlations Between Observation Score and Value-Added

Holistic Code	Correlation Coefficient
Teacher Uses Student Ideas	0.01
Teacher Remediates Student Difficulty	0.26
Students are Engaged	0.12
Classroom Characterized by Math Inquiry	-0.08
Lesson Time Used Efficiently	**0.45***
Density of the Mathematics is High	**0.35~**
Launch of Task	**0.35~**
Mathematics is Clear and Not Distorted	**0.34~**
Tasks and Activities Develop Math	0.31
Overall MQI	**0.37~**

Notes: *p<.05, ~p<.10

EXPLAINING MISALIGNMENT

- Qualitatively, rater memos and synthesis after actual value-added rankings suggest:
  - Across district groups, raters often noted little variability in instructional quality, which made it difficult to differentiate teachers.
  - Multiple instructional features that characterized lessons or teachers made it difficult to translate instructional quality into value-added rankings.
  - In some cases, limited information available to observers.

CONCLUSIONS

- While other studies show that observers and school leaders can tell the difference between teachers in the tails, we find that this is not necessarily true.
- We cannot better explain “production function” that converts classroom teaching into value-added scores.
- May need to rethink the ways in which teacher practices translate into desired student outcomes.

TAKE-AWAYS ACROSS THE STUDIES

- Consider using 3-year test-based accountability metrics TBAM) instead of 1-year metrics
- Consider the alignment of TBAM with alternative measures (like observations) to help inform model choice

- Consider the alignment of the components of teacher evaluation systems, in particularly how classroom observation instruments compare student assessments
  - Observation protocols may be better predictors of the type of rigor expected with CCSS than some state value-added scores

- Know that creating alignment will be an ongoing process.
  - We cannot better explain the "production function" that converts classroom teaching into value-added scores.

LESSONS FROM OUR WORK WITH CLASSROOM OBSERVATION

- Learn from early implementers
  - Observers who certify well, still have trouble rating teachers they know
  - Monitor during the school year; don’t wait until end of year
    - Compliance and quality
  - Consider independent observers
  - Use an established rubric
- Train and certify on the rubric with actual scoring
- Monitor
  - Co-observe in person or by video
  - Compare to mater ratings
- Train on how to give feedback based on the rubric, as well as on scoring with the rubric
  - get feedback from teachers on what is helpful and actionable