How Well Do Teacher Observations of Elementary Mathematics Instruction Predict Value-Added? Exploring Variability Across Districts

Kathleen Lynch, Mark Chin, & David Blazar

Harvard Graduate School of Education

Preliminary working paper. Please do not cite without authors' permission.

Abstract

In this study we ask: Do observational instruments predict teachers' value-added equally well across different state tests and district/state contexts? And, to what extent are differences in these correlations a function of the match between the observation instrument and tested content? We use data from the Gates Foundation-funded Measures of Effective Teaching (MET) Project(N=1,333) study of elementary and middle school teachers from six large public school districts, and from a smaller (N=250) study of fourth- and fifth-grade math teachers from four large public school districts. Early results indicate that estimates of the relationship between teachers' value-added scores and their observed classroom instructional quality differ considerably by district.

How Well Do Teacher Observations of Elementary Mathematics Instruction Predict Value-Added? Exploring Variability Across Districts

Kathleen Lynch, Mark Chin, & David Blazar

Introduction

School districts have rapidly adopted value-added models as a mechanism for measuring teachers' effectiveness (Chetty, Friedman, & Rockoff, 2011). Yet to date, the evidence is mixed on how value-added scores relate to expert observers' ratings of classroom instruction, with different studies returning markedly different correlations (Hill, 2009). One reason for this variability may be the differential sensitivity of the observational instrument to different tests of student achievement used to generate teacher value-added scores. When observational instruments and student assessments are more aligned on certain facets, stronger relationships may result. However, differences in achievement-by-observation score correlations may simply be stochastic, the result of unreliably estimated indicators or weak underlying relationships.

In the current study, we explore the relationship between teachers' instructional quality, as rated by trained observers who scored lesson videotapes using a single instrument and who were blind to teachers' districts, and the student achievement of these teachers' students, on (1) a standardized assessment that was uniform across the five study districts, and (2) districts' own state tests. Because instructional quality in classrooms across the five districts was measured using observational scores from a single, uniform metric, instructional quality scores should provide a stable measure across districts. Analogously, children participating in the study took a standardized mathematics assessment that was uniform across the five study districts; this allowed us to compare children's achievement in different districts on the same scale. Since instructional quality scores and children's scores on this standardized assessment were measured on the same metric across districts, we hypothesized that the relationship between teachers' instructional quality scores and their students' performances on this standardized assessment should be similar in different districts.

By contrast, during the course of the study, participating children also took their states' annual mathematics exams. Because study participants lived in four different states, they took four different state exams. Since instructional quality scores were still measured on the same metric across districts, but children's scores on their state exams likely did not, we hypothesized that the relationship between teachers' instructional quality scores and their students' achievement on state exams might differ depending on the district.

One possible factor which might contribute to district variability in the relationship between teachers' instructional quality and student achievement on state tests may be different sensitivity of the observational instrument to different state tests. When observational instruments and student assessments are more aligned, stronger relationships may result. One reason this may matter is that if variability in relationships is a function of the 'match' between the mathematical practices highlighted or the cognitive demand of the observation instrument and student assessments, districts may need to consider the alignment between their observational instruments and state tests when designing teacher observation plans.

Research Questions

We ask the following two research questions:

RQ1: Do observational instruments predict student achievement equally well across different state tests and district/state contexts?

RQ2: To the extent that these relationships vary across districts, can we identify factors that explain this variability?

We will explore the extent to which differences in these relationships may be a function of the ‘match’ between the observation instrument and tested content. We will also explore the possible impact of test cognitive demand, the alignment of test content with the state’s standards, test item format, and different levels of teacher ‘coaching’ to high-stakes tests across districts.

We hypothesize that in districts whose state tests (1) assess content at a higher level of cognitive demand, (2) pose more open-ended assessment items (which are designed with the goal of requiring more complex student thinking), (3) are more closely aligned with the MQI classroom observation instrument, and (4) more consistently assess their own content standards, the relationship between MQI and value-added will be stronger. We also hypothesized that the correlations between MQI and student achievement would be weaker in districts where teachers reported higher levels of ‘test prep’ and coaching in response to high-stakes testing; we hypothesized that in these districts, student achievement might increase in response to these teacher behaviors which are not captured by the MQI instrument.

Data

We use a subsample of data from two larger studies of fourth- and fifth- grade mathematics teachers from five large public school districts in the eastern United States, for a final sample including 298 teachers teaching a total of 6,780 students across two school years, 2009-2010 and 2010-2011. See Table 1 for a breakdown of the sample by district. For this study, we utilize administrative data gathered from districts on these students and videotapes of instruction from these teachers.

Table 1

Sample Breakdown of Teachers and Students by District

District	Teachers (N)	Students (N)
B	68	1628
D	92	1669
G	46	839
N	39	567
R	53	2077
Total	298	6780

Administrative data included student-teacher links from verified classroom rosters, student demographic information, and two sets of student test scores: (1) end-of-year mathematics and reading scores for standardized state tests completed in 2009, 2010 and 2011, encompassing in total six unique state test scores for each student (with 2009 scores serving as baseline prior scores), and (2) students' end-of-year scores on an alternative, external mathematics assessment, aligned with the Common Core State Standards for Mathematics, which students completed once at the beginning of the school year and once at the end of the school year.

Trained, certified raters watched and scored videotapes of each teacher's instruction using the Mathematical Quality of Instruction classroom observation instrument. Raters scored a total of 1,560 videos, with 93% of teachers being scored on at least three videos, and 55% of teachers being scored on at least six.

Research Question 1

In our first research question, we ask: Do observational instruments predict student achievement equally well across different state tests and district/state contexts?

Method

Teacher MQI scores

For this study, we focused on teachers' scores on six different codes of the MQI: Richness of Mathematics, Ability to Work With Students on Mathematics, Mathematical Errors and Imprecision, Common Core Aligned Student Practices, Lesson-level MQI, and Guess at Typical MQI. These codes were specifically designed to capture valued components of mathematics instruction, such as the teacher's development of mathematical generalizations, ability to remediate student mathematics, precision in use of mathematical language, and students' cognitive engagement with the mathematics beyond a procedural level. For a more in- depth description of each code, see Hill et al., 2012.

Each teacher received a score for each of these six MQI codes on each of his or her lessons. To generate teacher scores from this lesson-level data, we estimate the following equation:

$$(1)\ MQI_{jk} = \beta_0 + \mu_k + \epsilon_{jk}$$

Where the outcome of interest, MQI_{jk}, represents lesson k's score on the appropriate MQI code for teacher j. In this unconditional model, β_0 represents the grand mean of scores for the particular MQI code across all lessons. We use hierarchical linear modeling (HLM) to estimate equation (1), with nested random effects, μ_k, for each teacher *k*.

HLM provides empirical Bayes estimates of the teacher random effect, $\widehat{\mu_k}$, that are the best linear unbiased predictions. These empirical Bayes estimates are "shrunken" estimates, which account for differences in the reliability of the estimates due to differences in number of lessons scored from teacher to teacher by shrinking less reliable estimates toward the mean (Raudenbush & Bryk, 2002).

Analyses

To address our first research question of whether observational instruments predict student achievement equally well across different state tests and districts/state contexts, we estimate the following equation:

(~~2~~3) $a_{jcksgdt} = \beta_0 + \beta_1 A_{jt-1} + \beta_2 X_{jt} + \beta_3 C_{jct} + \beta_4 S_{jsgt} + G_{gt} + \mu_k + \epsilon_{jcksgt}$

(~~3~~4) $\mu_k = \beta_6 (D_d \times MQI_k) + \tau_k$

Where the outcome of interest, a_{jcksgt}, represents student j's standardized score on either (1) the state mathematics exam, or (2) the alternative mathematics exam;

A_{jt-1} represents a vector of prior achievement for student *j* in time *t-1*, including a linear, quadratic, and cubic term for student j's mathematics exam score at time *t-1*, and a linear term for student *j*'s score on the reading exam from time *t-1*;

X_{jt} represents a vector of student demographic indicators for student *j* at time *t*, including gender, race, free- or reduced-price lunch eligibility, special education status, and limited English proficiency; and

C_{jct} and S_{jsgt}represent the aggregate of these two vectors for each student's (1) class *c*, and (2) school *s* and grade *g*.

Also included in the model is a vector of grade-by-year fixed effects, G_{gt}, to account for differences across grades and school years. To be included in the model, student j's tested grade at time t must follow sequence with regards to his or her tested grade at time t-1. Furthermore, class c must have fewer than 50% of students having special education status, fewer than 50% of students missing scores for the prior achievement vector, and, after all other restrictions, must have a sample of at least five students.

In our model, students are nested within teachers; thus, we include a random effect μ_k for teacher. From equation (4), we see that part of the random effect of teacher t on student achievement is due to the interaction of teacher t's MQI score and the district d he or she teaches in. If the MQI predicts student achievement equally well across districts, we would expect that the coefficient for each district d's interaction to not be statistically significant different from one another.

Results

Results from Wald tests, examining whether or not MQI-interaction regression parameters on student achievement significantly differed comparing districts, indicate that the relationships between value-added scores estimated using the alternative mathematics assessment and classroom quality are similar across districts, for the population. By contrast, estimates of the relationship between teachers' MQI scores and student achievement using state mathematics tests differ considerably by district. See Figures 1 and 2.

Figure 1. The relationship between teachers' value-added scores estimated using the NCTE test (top panel), and state tests (bottom panel), with teachers' observed classroom instructional quality

Regression Parameters on Student NCTE Test Achievement

	Overall		District B		District D		District G		District N		District R	
MQI Code	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
Richness	0.02	0.01	0.02	0.02	0.07	0.05	0.02	0.03	-0.02	0.04	0.05	0.03
Working with Students	0.04**	0.01	0.05*	0.02	0.03	0.04	0.03	0.03	-0.02	0.03	0.09*	0.04
Errors and Imprecision	-0.02	0.01	-0.04*	0.02	-0.03	0.03	0.01	0.03	0.00	0.03	0.00	0.03
Common Core Student Practices	0.03*	0.01	0.04*	0.02	0.02	0.04	0.00	0.03	0.00	0.04	0.06	0.04
Lesson-Level MQI	0.02	0.01	0.04	0.02	0.04	0.03	0.00	0.03	-0.02	0.04	0.07	0.04
Guess at Typical MQI	0.03*	0.01	0.03	0.02	0.04	0.03	0.01	0.03	-0.01	0.04	0.07	0.04

Regression Parameters on Student State Test Achievement

	Overall		District B		District D		District G		District N		District R	
MQI Code	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
Richness	0.03	0.02	0.06**	0.03	-0.01	0.05	-0.06	0.03	0.02	0.05	0.06	0.04
Working with Students	0.05**	0.02	0.08***	0.02	-0.05	0.05	-0.03	0.04	0.03	0.04	0.12*	0.05
Errors and Imprecision	-0.03	0.02	-0.07**	0.03	0.02	0.03	0.01	0.03	-0.04	0.04	-0.02	0.05
Common Core Student Practices	0.06**	0.02	0.08***	0.02	0.02	0.05	0.05	0.04	0.03	0.04	0.05	0.05
Lesson-Level MQI	0.02	0.02	0.07**	0.02	-0.01	0.04	-0.08*	0.03	0.00	0.04	0.08	0.05
Guess at Typical MQI	0.03	0.02	0.07**	0.02	-0.02	0.04	-0.05	0.04	0.04	0.04	0.07	0.05

**p<.05 **p<.01 ***p<.001*

Figure 2. Estimates of the statistical significance of district differences in the relationship between teachers' value-added scores estimated using the NCTE test (top panel), and state tests (bottom panel), with teachers' observed classroom instructional quality.

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	

From the top panel of the Figure 2 above, we see that, in the population, the interaction of teacher MQI scores and each district on student NCTE test achievement does not significantly differ from the main effect of teacher MQI on student NCTE test achievement (see column, All=Beta). Similarly, when testing the statistical significances of differences between these interaction regression parameters from district to district, we see only one significant difference, despite seeing in the top panel of Figure 1 statistically significant relationships of MQI and student achievement on the NCTE assessment for certain districts. We see that the relationship of teacher ability to work with students and student achievement on the NCTE assessment is significantly different from district N and district R.

Conversely, we see many more statistically significant differences between districts regarding the relationship of MQI and student achievement on state tests. Looking at the bottom panels of Figure 1 and Figure 2 above, a pattern emerges. Teacher MQI scores often do relate to student achievement on state tests for districts B and R, and the stronger relationship of teacher MQI to student achievement in these two districts generally differs significantly from that of districts D and G, in the population. As expected, no statistically significant differences emerged between districts whose students took the same state test (districts B and R).

We focus primarily on these differences between districts B and R to districts D and G in our subsequent analyses.

Research Question 2

In our second research question, we ask: To the extent that the relationships between observational instruments and student achievement vary across districts, can we identify factors that explain this variability? To explore this question, we examined six possible factors that might contribute to cross-district differences: (1) alignment of test items with states' curriculum standards; (2) tests' cognitive demand; (3) tests' item formats; (4) tests' alignment with MQI instrument; (5) content coverage; and (6) cross-district variations in teachers' responses to high-stakes testing.

Method

Characteristics of Mathematics Tests

We first gathered information about the state mathematics tests administered to fourth- and fifth-graders during the study years in each of our five study districts. For one district, we were able to obtain complete versions of the relevant tests; for the four districts where complete versions of the test were not available, we considered all available publicly-released test items for each test, and cross-referenced the released items against each year's relevant state test blueprints in order to check that the released items were reasonably representative of the administered test as a whole. Two districts were from the same state and therefore utilized the same state test. We recovered complete versions of all forms of the external, alternative mathematics assessment that was distributed to all students in the study.

Of the complete set of test information gathered, we used a subset of items to characterize each test: (1) the complete set of released items from three state mathematics tests from the school year 2009 to 2010, (2) all available released items from one state mathematics test (for which the complete set of items was not available), and (3) the complete set of items from a randomly selected form for the external, alternative mathematics test in the school year 2011 to

2012, chosen for better item functioning. In total, 235 items from grade 4 assessments and 240 items from grade 5 assessments were scored. On average, 95 items from each test was scored. We coded these items and tests on the dimensions below.

Alignment of Test Items with States' Curriculum Standards

First, we examined each test for the consistency with which it assessed its own state's content standards. We adapted the Achieve framework (Resnick, Rothman, Slattery, & Vranek, 2004)[1], which has been utilized in prior research to examine the extent to which state tests are aligned to state standards. Raters utilized the Achieve protocol to score each test item on the following dimensions: (1) *content centrality*, which measured the quality of the match between the content of each test question and the content of its associated state standard; (2) *performance centrality*, which assessed the quality of the match between the cognitive demand made by each item and the cognitive demand level of the performance specified in the associated target state standard; and (3) *source of challenge*, which evaluated whether or not each item was 'fairly constructed,' in the sense that the challenge posed by the item was rooted in the subject matter and performance delineated in the target standard (and not in an unfair source of challenge, such as an unfair appeal to students' background knowledge). Content centrality and performance centrality were scored on a scale of 0 (inconsistent) to 2 (clearly consistent). Each item was scored for source of challenge on a 0 to 1 scale, with 0 indicating an inappropriate source of challenge, and 1 indicating an appropriate source of challenge.

Tests' Cognitive Demand

[1] Because the alternative mathematics assessment was developed indirectly tied to a set of standards, we did not score items on the assessment on the Achieve framework.

Second, to evaluate the level of cognitive demand made by each test, we drew on the Surveys of Enacted Curriculum (SEC) framework (Porter, 2002) to code each test item based on its expectations for student performance, categorized using five levels of cognitive demand: (1) *memorize*, (2) *perform procedures*, (3) *communicate understanding*, (4) *solve non-routine problems*, and (5) *conjecture/generalize/prove*. For a detailed description of the codes for each of these categories, see Porter, 2002. Raters assigned each item one score for this dimension, representing their judgment of the category that best matched the type of cognitive demand that the item posed to students. We consider these codes as representing a 1 to 5 scale with higher values representing higher levels of cognitive demand, with *memorize* items at the low end and *conjecture/generalize/prove* items at the high end. Each score point on the SEC scale was thus given a numeric value such that the scale is ordinal. Therefore, higher average SEC scores for a test might suggest a test with fewer 'memorize' items and more 'demonstrate understanding' items.

Tests' Item Formats

Third, we sought to develop a picture of each test's distribution of item formats. We utilized the test item format categories described in the AERA/APA/NCME *Standards for Educational and Psychological Testing* (1999) to code the proportion of items on each test that were multiple choice, short answer (including constructed response items and items in which students were asked to 'bubble-in' an answer), and open-ended (such as items requesting longer responses, short essays, or explanations of answers). We used complete tests or associated test blueprints and documentation in order to generate codes for this category.

Tests' Alignment with MQI Instrument

Fourth, we examined the extent of the alignment between the MQI instrument and the skills and competencies measured by the various student assessments. We asked: Did the skills and competencies recognized on the MQI observation instrument (such as recognizing and utilizing multiple procedures or providing mathematical explanations) align with the skills and competencies that the test items demanded?

To measure the degree of alignment, for each of the four state tests, we assessed whether each test item demanded a high, medium, or low level of student engagement with two MQI elements: Overall Richness of Mathematics, and Enacted Task Cognitive Activation. We selected these two dimensions because they represent summary measures of students' opportunities to engage with rich and cognitively activating mathematics, and due to our hypothesis that these were the two MQI dimensions which could be captured in a student assessment. For example, the Overall Richness dimension on the MQI includes sub-domains such as the extent to which the teacher exposes students to multiple strategies for solving problems. If a test item also asked students to solve a problem using multiple strategies, we would capture this alignment using the Overall Richness test code category. Other MQI dimensions, such as the number and severity of mathematical errors that teachers make during instruction, are important but do not map readily onto skills and competencies demanded on a student assessment, and thus were not coded for the current analysis.

Tests' Content Coverage

To allow for comparisons of tests' relative coverage of different domains of mathematics content, reviewers evaluated the proportion of items on each test that were dedicated to geometry; number and operations; patterns, relations, and algebra; measurement; and data analysis, statistics, and probability.

Cross-District Variations in Teachers' Responses to High-Stakes Testing

We explored the possibility that differences between districts in the relationships between student achievement as measured on state tests and instructional quality might be linked to the differences in the prevalence of 'coaching' practices across districts. We hypothesized that in districts where teachers reported engaging in a high degree of test prep 'coaching' practices, such as test-targeted drill, students might improve significantly on state tests, while teachers might simultaneously earn low scores for instructional quality on an instrument that valued cognitively activating tasks and conceptual learning. To assess this possibility, we examined items from teacher questionnaires administered in the fall of each study year that asked teachers about their usage of 'test prep' and 'test coaching' practices.

From the items on the survey, two metrics of test prep/coaching were derived for each teacher. The first measure captured the frequency of test prep activities engaged in by the teacher, including: use of state test items or practice test materials in the classroom; incorporating formats of state tests in instruction; using class time to review state test material; focusing instruction on 'bubble' students; and teaching state-test-taking specific strategies. The internal reliability of this test prep 'activities' measure was high ($\alpha_{t1} = .73;\ \alpha_{t2} = .78$).

The second measure captured the extent to which preparation for state tests altered a teacher's instruction, including: changing topic coverage based on tested items of the state test; spending less time discussing mathematical concepts in depth; lowering the number of special projects or demanding mathematics problems; and sequencing mathematical topics such that state test content is covered earlier. The internal reliability of this measure of change in instruction due to test prep was high ($\alpha_{t1} = .86, \alpha_{t2} = .78$).

To estimate a teacher score for each of these test prep measures, we used the following equation:

$$(2)\ TP_{jk} = \beta_0 + \mu_k + \epsilon_{jk}$$

Where the outcome of interest, TP_{jk}, represents teacher j's response to item *k* of questions of each test prep measure. We use hierarchical linear modeling (HLM) to estimate equation (2), with nested random effects, μ_k, for each teacher *k*. HLM provides empirical Bayes estimates of the teacher random effect, $\widehat{\mu_k}$, that are the best linear unbiased predictions. These empirical Bayes estimates are “shrunken” estimates, which account for differences in the reliability of the estimates due to differences in number of items responded to from teacher to teacher by shrinking less reliable estimates toward the mean.

Each teacher's score was recovered from the equation, and then standardized with mean zero, standard deviation one, across the entire population of teachers. District-level test prep behavior was generated by collapsing the scores of all teachers within a given district. Because teachers in District N were not given this survey, we were unable to recover test prep behavior scores for this district.

Scoring Test Items and Scale Reliability

To ensure the reliability of item scores on the scales, three raters first scored a random subsample of the items, equally distributed across 4th grade tests, to establish common scoring practices and to refine the coding scheme. Reliability was then calculated from rater scores on a different subset of items, randomly selected to be equally distributed across 5th grade tests. Inter-rater reliability, as determined by Cohen's Kappa, varied from high to weak, as Kappa values for each of the scales ranged from .25 to .85. Because of the wide range in reliability, scores for each item were ultimately reconciled by and agreed upon by all three raters.

Results

Alignment of Test Items with States' Curriculum Standards

We hypothesized that if in some states, content on the state tests was simply not aligned with the standards that teachers were asked to cover or the items simply were simply unfair, the relationship between MQI and student state test achievement would be weaker than in states where teachers had a fair chance to teach the tested content. To evaluate this hypothesis, we coded state test items using three dimensions of the Achieve Framework Protocol, (1) *content centrality*, (2) *performance centrality*, and (3) *source of challenge*, described above.

Results of this analysis are presented below.

Table 2

Test Breakdown on Achieve Framework Measure of Match of Test Items to Curriculum Standards

	Content Centrality		Performance Centrality		Source of Challenge	
Test	Mean	SD	Mean	SD	Mean	SD
Districts B & R, N=25	1.76	0.52	1.68	0.56	0.96	0.20
District D, N=91	1.89	0.43	1.66	0.56	0.95	0.23
District G, N=24	1.13	0.99	0.79	0.83	0.50	0.50
District N, N=50	1.94	0.24	1.76	0.43	1.00	0.00

Our results show, with the exception of the state test for District G, items on these tests generally assessed the content standards and performance standards that they were matched to consistently, and that the source of challenge of these items was fair.

While the standard deviations of our estimates are large due to the relatively small number of test items, in examining the point estimates we find that the results only partially support our hypothesis. Tests from Districts B, R, D, and N appear to score similarly on each of the measures of the Achieve Framework Protocol, though student state test achievement did weakly relate to teacher MQI scores in District G, where test items seem to be unfair in their source of challenge and are commonly misaligned to the performance and content standards purported.

Tests' Cognitive Demand

We hypothesized that in states where the state tests assess content at a higher level of cognitive demand, the relationship between MQI and student state test achievement would be stronger than it is in states where state tests assess content at a lower level of cognitive demand.

If this were true, we would expect that Districts B and R, which were in the same state and had relatively stronger correlations between teachers' MQI and value-added, would have the state test with the highest level of cognitive demand. By contrast, we would expect districts D and G, which had relatively weaker correlations between teachers' MQI and value-added, to have state tests presenting lower levels of cognitive demand.

The results of this analysis are presented in Table 3 below. We first note that across all four tests, the overall average level of cognitive demand presented by state tests was relatively low. Most state test items asked students only to memorize vocabulary terms or to perform procedures. Very few asked students to engage with the content at higher levels of cognitive demand by solving non-routine problems, and fewer still asked students to make conjectures, arrive at generalizations, or prove.

While the standard deviations of our estimates are large due to the relatively small number of test items, in examining the point estimates we do find that the results are generally in the direction of our hypothesis. Districts B and R's state test presented the highest level of cognitive demand. Districts D and G had state tests that assessed content at a somewhat lower level of cognitive demand.

Table 3. Average Levels of Cognitive Demand Presented by Items on Four State Tests, as Measured Using Scale Scores from the Surveys of Enacted Curriculum Protocol (Porter, 2002)

	Summary Statistics	
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

We hypothesized that the correlation between MQI and value-added would be stronger in states whose test posed more open-ended assessment items, since we hypothesized that these might require more complex student thinking of the type that the MQI instrument is designed to capture. We found that District B/R's state test did indeed look different from those of District D and District G on this dimension. The District B/R state test devoted more than a quarter of its math assessment to non-multiple-choice items, and over 20% of the test was comprised of longer, open-ended items. By contrast, District D dedicated only 14% of its assessment to short-answer items, and included no longer, open-ended items; District G's test was comprised of multiple-choice items only.

However, if item formats were the sole reason for district differences in the relationship between MQI and value-added on state tests, then we would also have expected to observe differences on this measure between Districts B/R and District N, since, like District G's, District N's test was also exclusively multiple choice. However, we do not observe differences between Districts N or R and District N in the relationship between MQI and value-added on the state test. Therefore, item format composition appears not to be the sole driver of the between-district differences that we observe.

Table 4.
Breakdown of Test by Item Format

	Percent of Items		
Test	Multiple Choice	Short Answer	Open-Ended
Districts B & R, N=25	64	12	24
District D, N=91	86	12	2
District G, N=24	100	0	0
District N, N=50	100	0	0

Tests' Alignment with MQI Instrument

We hypothesized that in districts where the state tests were more closely aligned with the MQI classroom observational scoring instrument, the relationship between MQI and student state test achievement would be stronger than in districts whose state tests assess were more weakly aligned with the MQI.

We hypothesized that Districts B and R, which were in the same state and had relatively stronger correlations between teachers' MQI and value-added, would have the state test that was most closely aligned with the MQI observation instrument. By contrast, we expected districts D and G, which had relatively weaker correlations between teachers' MQI and value-added, to have state tests that were less aligned with the MQI instrument.

We present results from these analyses in Tables 5 and 6 below. We find that in general, the results do seem to support our hypotheses. Although again the standard errors were relatively large due to the small number of test items, in examining the point estimates we find that the state tests for Districts B and R was more closely aligned with the MQI observational instrument than the tests for Districts D and G on both measured dimensions.

Table 5. Average Levels of State Tests' Alignment with the MQI Observational Instrument on the MQI Overall Richness Dimension

	Summary Statistics	
Test	Mean	SD
Districts B & R, N=25	1.40	0.58
District D, N=91	1.22	0.47
District G, N=24	1.13	0.34
District N, N=50	1.06	0.24

Table 6. Average Levels of State Tests' Alignment with the MQI Observational Instrument on the MQI Enacted Task Cognitive Activation Potential Dimension

	Summary Statistics	
Test	Mean	SD
Districts B & R, N=25	1.72	0.94
District D, N=91	1.25	0.63
District G, N=24	1.21	0.59
District N, N=50	1.30	0.58

Tests' Content Coverage

To allow for comparisons of tests' relative coverage of different domains of mathematics content, reviewers evaluated the proportion of items on each test that were dedicated to geometry; number and operations; patterns, relations, and algebra; measurement; and data analysis, statistics, and probability.

We present the results of this analysis in Table 7. Based on this analysis, tests' differential content coverage did not appear to be a likely cause of the differences we observed between districts in the MQI/state test value-added relationship. The distribution of content coverage on different state tests was generally quite similar, perhaps with the exception of District G's assessment which contained fewer items in the areas of patterns, relations, and algebra and data, statistics and probability, and more measurement items in the fifth grade and geometry items in the fourth grade.

Table 7. Percent of Items on Each State Test (2009-2010) that Corresponded to Five Mathematical Domains.

Test	Grade	Percent Geometry	Percent Number Sense and Operations	Percent Patterns, Relations, & Algebra	Percent Measurement	Percent Data Analysis, Statistics, & Probability
Districts 1 and 2	4	0.125	0.35	0.2	0.125	0.2
	5	0.13	0.33	0.26	0.13	0.15
District 3	4	0.2	0.43	0.1	0.17	0.1
	5	0.1	0.38	0.1	0.32	0.1
District 4	4	0.13	0.32	0.2	0.15	0.2
	5	0.15	0.3	0.25	0.15	0.15
District 5	4	0.24	0.4	0.2	(Categorized with geometry)	0.16
	5	0.24	0.4	0.2	(Categorized with geometry)	0.16

Cross-District Variations in Teachers' Responses to High-Stakes Testing

We hypothesized that the relationships between MQI and student state test achievement would be weaker in districts where teachers reported higher levels of 'test prep' and coaching, with the idea that in such districts, student achievement might increase even as instructional quality was weak. We examined teachers' responses to two sets of questionnaire items designed to capture the extent to which they engaged in test prep activities and test prep instruction.

We present the results of this analysis in Table 8, below. Overall, the standard errors for our estimates are large, but in inspecting the point estimates the results do not appear to be clearly consistent with our hypothesis. While we do observe that teachers in Districts B and R reported that they engaged in lower levels of test prep activities and instruction than teachers in Districts D, teachers in District R reported engaging in similar and even slightly higher levels of test prep activities and instruction than those in District G.

Table 8. Average Teacher Test Prep Activities and Changes to Instruction by District

	Test Prep Activities		Test Prep Instruction	
Test	Mean	SD	Mean	SD
District B, N=71	-0.21	0.92	-0.25	1.11
District R, N=56	0.08	0.93	0.01	0.86
District D, N=49	0.32	1.16	0.34	1.05
District G, N=125	-0.04	0.99	0.00	0.04

Discussion

We begin by summarizing our findings. First, we found that relationships between student achievement on the NCTE mathematics assessment and classroom quality are similar across districts. By contrast, estimates of the relationship between teachers' MQI scores and student achievement using state mathematics tests differ considerably by district.

What factors might explain this variability across districts in the relationship between instructional quality, as judged by external observers, and teachers' value-added scores? At the outset, we hypothesized that in districts where the state tests (1) assessed content at a higher level of cognitive demand, (2) posed more open-ended assessment items, (3) were more closely aligned with the MQI observation instrument, and (4) more consistently assessed their own content standards, the observed relationship between MQI and student state test achievement would be stronger. We also hypothesized that the relationships between MQI and student achievement would be weaker in districts where teachers reported higher levels of 'test prep' and coaching activities. Specifically, we hypothesized that that state test for Districts B and R would exhibit a higher level of content cognitive demand, more open-ended assessment items, closer alignment to the MQI instrument, and a greater degree of alignment to the state's content standards than the state tests for Districts D and G.

In general, our analyses appear to suggest some support for these hypotheses. Although the standard errors for many of our estimates were large, in examining the point estimates, we found that the state test for Districts B and R had higher average item scores on the SEC Cognitive Demand protocol, more open-ended assessment items, closer alignment to the MQI instrument as captured in the Overall Richness and Enacted Task Cognitive Activation Potential measures, and a greater degree of alignment to the state's content standards as measured using a subset of items from the Achieve Protocol than the state tests for Districts D and G.

This study presents multiple important limitations. First, a key limitation of the current study is that we have been unable to identify an appropriate framework in which to empirically model and test the strength of these multiple contributing factors to variability in the relationship between MQI and student achievement, as the dimensions we examine are perfectly collinear with district. We hope to continue to explore this modeling possibility in future research. Second, as noted above, data are missing for some analyses (such as from the teacher questionnaire in District N). Third, the current study is clearly observational and exploratory in nature.

Conclusions and Future Directions

School districts and states have moved rapidly to implement teacher evaluation systems that heavily weight classroom observation and value-added scores. In the current study, we find evidence that the relationship between instructional quality and teacher value-added as calculated using state tests is different in different districts. As a result, we suggest that districts may need to examine their classroom observational instruments for alignment with their high-stakes student assessments.

References

American Educational Research Association (AERA)/American Psychological Association (APA)/National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics*, *126*(4), 1593-1660. doi: 10.1093/qje/qjr041

Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management*, *28*(4), 700-709. doi: 10.1002/pam.20463

Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., ... & Lynch, K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment*, *17*(2-3), 88-106.

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, *48*(1), 163-193. doi: 10.3102/0002831210362589

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, *31*(7), 3-14. doi: 10.3102/0013189X031007003

Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2004). Benchmarking and alignment of standards and testing. *Educational Assessment*, *9*(1-2), 1-27. doi: 10.1080/10627197.2004.9652957