Relationships between Observations of Elementary Mathematics Instruction and Student Achievement: Exploring Variability across Districts

KATHLEEN LYNCH and MARK CHIN
Harvard Graduate School of Education

DAVID BLAZAR
University of Maryland College Park

Much debate surrounding teacher quality has focused on students' standardized test scores, but recent federal and state initiatives have emphasized the use of multiple measures to evaluate teacher quality, including classroom observations. In this study, we explore differences across school districts in the relationship between student achievement outcomes and the observed quality of teachers' instruction. Using data from 298 elementary mathematics teachers in five urban US districts, we examine relationships between teachers' performance on the Mathematical Quality of Instruction observation instrument and their students' scores on both state standardized and researcher-developed tests. We find that these relationships differ across school districts. We explore the extent to which differences in skills and expectations for students across tests may explain this variability. An improved understanding of the relationship between classroom observations and student tests may help districts to better support teachers in developing their instructional effectiveness.

Spurred in part by federal initiatives such as Race to the Top and No Child Left Behind, school districts have accelerated development of new teacher evaluation systems that aim to improve student achievement and to increase teachers' effectiveness (e.g., Firestone 2014; Harris et al. 2014). Although much of the policy debate regarding teacher quality has focused on student standardized assessments (Konstantopoulos 2014), recent federal initiatives emphasize the role

Electronically published June 19, 2017
American Journal of Education 123 (August 2017)

0195-6744/2017/12304-0004$10.00

that classroom observations should play as a measure of teachers' professional practice (e.g., Coggshall et al. 2012).

To date, however, the evidence is mixed on the extent to which student achievement outcomes and classroom observation scores align to identify a common set of highly effective or ineffective teachers (e.g., Kane et al. 2011; Kane and Staiger 2012). Correlations between classroom observation scores and student achievement outcomes generally vary across studies, from weak to moderate (e.g., Bell et al. 2012; Daley and Kim 2010; Milanowski 2004, 2011), although some research has found stronger relationships (Schacter and Thum 2004).

Several theoretical and practical considerations underscore the need to understand why relationships between student achievement and teacher observation scores vary across contexts. On the one hand, perfect alignment between test scores and teacher observations is not expected; indeed, part of the rationale for including multiple measures in teacher evaluations is that these may provide broader information about instruction (Grossman et al. 2014). On the other hand, strong relationships between student test scores and teacher observation scores corroborate the logical expectation that better teaching will lead to increased student learning and, subsequently, improved test scores. Weak relationships challenge this theory and may impede the use of teacher evaluations in practice. Schools often use student-achievement and teacher-observation scores to inform important personnel decisions, such as merit pay or professional development (Podgursky and Springer 2007). Misalignment between student test scores and teacher observation scores complicates the decision of which teachers to target for such interventions. For teachers, misalignment could potentially result in conflicting feedback on how to improve practice. For example, if a teacher focused instruction relatively narrowly on basic skills, the teacher's students might show improvement on a basic skills–oriented state test, but the teacher might simultaneously receive low ratings on observational rubrics that emphasize more varied mathematics and instructional practices (Grossman et al. 2014; Polikoff 2014).

KATHLEEN LYNCH is a doctoral student at the Harvard Graduate School of Education. Her research interests include education policy and strategies to reduce educational inequality, particularly in mathematics. MARK CHIN is a doctoral student in education policy and program evaluation at the Harvard Graduate School of Education. His research interests center on how race, racism, and assimilative pressures affect the experiences and outcomes of students of color and early generation immigrant students in US K–12 contexts. DAVID BLAZAR is an assistant professor at the University of Maryland College Park. His research examines factors that affect teacher and teaching quality, with a focus on professional learning, the organizational context of schools and districts, and accountability policy.

In the current study, we take up the questions of whether differences in the relationships between teachers' observation scores and students' achievement outcomes exist across district contexts and why these differences may exist. We explore an explanation for differences in relationships that, until recently, has received relatively little attention: the extent to which classroom observation instruments value teaching practices that are not well aligned to the skills expected of students on tests, resulting in test–observation misalignment.

To explore this issue, we use data from a sample of 298 elementary teachers teaching 6,780 students in five urban school districts, which were located in four different states. We examine the relationship between teachers' scores on the Mathematical Quality of Instruction (MQI), a classroom observation instrument developed to assess elementary mathematics teaching, and these teachers' students' achievement on two assessments: (1) a researcher-developed test, which was uniform across the five study districts, and (2) state standardized math tests. Instructional quality in all classrooms was measured using the MQI, meaning that teachers across districts were observed using the same metric.

Results from this study can provide guidance on features that may affect the alignment of observation instruments and student assessments used in new teacher evaluation systems.

Background and Research Context

Relationships between Classroom Observations and Student Achievement Outcomes

There is a long history of research attempting to link characteristics of teachers and teaching to student achievement. Building from a broad tradition in the education production function literature of relating teacher characteristics (e.g., education, training, years of experience) to student outcomes, process-product studies beginning in the 1970s focused specifically on the relationship between teachers' classroom practices and student outcomes. Research found, for example, relationships of teachers' confidence, efficient use of class time, and group management to student achievement outcomes (for reviews, see, e.g., Good and Brophy 2007; Konstantopoulos 2014). Research on opportunities to learn has found the amount and quality of students' exposure to new knowledge to be related to learning outcomes (e.g., Stevens 1993). More recently, research using student surveys has found relationships between students' perceptions of teacher qualities such as care and challenge and student achievement (Kane and Cantrell 2010). Critiques of this literature have noted the largely correlational nature of findings and the lack of focus on subject-specific instructional practices (Hill et al. 2005).

In recent years, researchers and practitioners have developed a variety of new classroom observation instruments to assess the quality of teachers' instructional practices. The instruments were developed for various purposes ranging from the evaluation of early childhood instructional interactions and federally funded interventions (e.g., Gamse et al. 2008; Pianta et al. 2010) to instrument validation (e.g., Hill et al. 2008). From a policy standpoint, the competition between states for federal Race to the Top grants and the Measures of Effective Teaching (MET) project's research emphasis on rigor in teacher evaluation further supported the reform and development of new teacher observation instruments (Hill, Charalambous, and Kraft 2012). Classroom observation protocols were developed for several reasons, but they shared the goal of providing observers across contexts standardized metrics to evaluate instruction.

Relationships between teachers' scores on observation instruments and student achievement outcomes vary across studies. With some exceptions (e.g., Schacter and Thum 2004), these generally range from weak to moderate. Studies observing moderate relationships between these measures found correlations or effect sizes in the range of 0.3 and 0.4 (Daley and Kim 2010; Hill et al. 2011; Kane et al. 2011; Milanowski 2004). However, several other studies have found relatively smaller relationships (e.g., Bell et al. 2012; Kane and Staiger 2012; Milanowski 2011). As an example of what was being related in these studies, in Daley and Kim (2010), the classroom observation tool assessed a mix of process-oriented (e.g., grouping students, lesson pacing) and content-oriented (e.g., presenting instructional content, problem solving) elements, and state standardized tests measured student achievement.

Exploring Variability in the Relationships between Teacher Observation and Student Achievement Scores

Extant research suggests a handful of possible explanations for varying relationships between teacher observation scores and student test scores. One category of explanations relates to errors in measuring such scores. For example, researchers have found that teachers' value-added scores are inconsistent across years and sensitive to test timing (McCaffrey et al. 2009; Papay 2011) yet gain precision when data from multiple years are used (Koedel and Betts 2011). Observation scores also suffer from measurement error due to, for example, variability in rater use of the instrument and the specific lessons observed (Bell et al. 2012; Hill, Charalambous, and Kraft 2012; Kane and Staiger 2012; Milanowski 2011). Even before relating observation scores and student achievement, theory would indicate that measurement error in either or both of these measures would attenuate the strength of relationships (Spearman 1904). Prior research has thus documented the impact of measurement error on teacher

observation and student test scores; this suggests that it may also play a role in explaining cross-site differences in the relationships between these scores.

Beyond measurement error, however, relatively little research has explored which other factors might contribute to cross-district differences in the relationships between student test scores and teacher observation scores. In the current study, we explore the potential impact of misalignment between student tests and observation instruments in explaining these differences.

Classroom observation instruments may value certain teaching practices that are not well aligned with the skills expected of students on tests, resulting in test–observation misalignment. In this framework, different studies may return markedly different correlations between teacher observation scores and student achievement scores because of the differential sensitivity of the tests to varying classroom practices. Polikoff (2014) has suggested that state tests differ in their instructional sensitivity or the extent to which they reflect the content or quality of teachers' instruction. Low correlations between student test scores and classroom observation scores may reflect a discrepancy between the types of learning goals valued in classroom observations and the learning goals valued in standardized tests (Grossman et al. 2014; Polikoff 2014). In particular, a core dimension along which standardized tests and classroom observations may differ is the cognitive demand of the learning activities they measure (Doyle 1988). In Doyle's (1988) framework, low-cognitive-demand activities include memorizing facts and applying formulas and procedures without attention to meaning. Stein et al. (1996) argue that classrooms dominated by these activities "do not provide the conditions necessary for the development of students' capacity to think and reason mathematically" (457). By contrast, high-cognitive-demand activities generally involve comprehension, interpretation, or synthesis and include solving problems using multiple solution strategies, comparing representations, and explaining and justifying ideas (Doyle 1988).

Previous research has found that state standardized tests often assess relatively low-cognitive-demand skills, such as using procedures and applying formulas (e.g., Resnick et al. 2004; Webb 1999). For example, in a study analyzing state assessments in mathematics and language arts, Resnick et al. (2004) found that state assessment items were generally inappropriately easy relative to the levels of cognitive demand indicated in state curriculum standards. In addition, state tests are frequently composed of mostly multiple-choice items. Although open-ended items do not automatically indicate a higher level of cognitive demand than multiple-choice items, they are nonetheless theorized to provide the potential to pose higher levels of cognitive demand because they cannot be solved by working backward from a list of answer choices and because open-ended writing may offer opportunities to assess distinct logical abilities (Bridgeman 1992).

In contrast, as discussed in the section on the theory of instruction underpinning the MQI, classroom observation instruments often measure the extent to

which instruction facilitates high-cognitive-demand activities, such as comparing, explaining, and justifying. This suggests that on observation measures, teachers may be rewarded for imbuing content with meaning or providing cognitively demanding activities; alternatively, on some standardized tests, students may be asked primarily to demonstrate relatively low-level skills. In this example, the relationship between teachers' observation scores and their students' achievement may be weak because these measures are poorly aligned. Indeed, this problem has received attention from policy makers; with the increased interest in more rigorous curriculum frameworks such as the Common Core State Standards for Mathematics (National Governors Association Center for Best Practices, Council of Chief State School Officers 2010), this issue of misalignment between the content of what is taught and what is assessed has spurred work on the development of more conceptually oriented, rigorous assessments designed to better assess higher level competencies (Grossman et al. 2011).

To date, we have been able to identify only three prior studies that have explored differences in the types of skills expected of teachers and students in an observation instrument versus in student assessments, all using data from the MET project. MET researchers (Grossman et al. 2014; Kane and Staiger 2012) observed that the relationships between teacher scores on observation instruments used in the MET study and student achievement were stronger when student achievement was measured using the Stanford 9 Open-Ended (SAT-9 OE) reading assessment than when measured using state standardized tests. Kane and Staiger (2012) speculated that this may have been because of differences between the two tests. They suggested that most state tests do not ask students to write about what they have read, a key skill that is both measured on the SAT-9 OE and a major focus of reading teachers' instruction.

Also using MET data, Grossman et al. (2014) examined how the relationship between teachers' scores on the Protocol for Language Arts Teaching Observations (PLATO) instrument and their value-added scores differed depending on whether student learning was assessed using state tests or the SAT-9 OE. They found that the relationship between teachers' PLATO and value-added scores did differ depending on the student assessment used to construct the value-added scores. Teachers' PLATO scores were more correlated with their value-added scores when estimated from the SAT-9 OE test than from state standardized tests. Grossman et al. suggested that higher correlations between students' SAT-9 OE performance and teachers' PLATO scores could have resulted because the SAT-9 OE was designed to capture students' skills in argumentation and in developing ideas—skills also valued by the PLATO observation protocol—whereas the authors presumed that the state tests tended to emphasize lower level comprehension skills and multiple-choice questions.

In addition, Polikoff (2014) examined MET data disaggregated by state. He found that the correlations between value-added scores and teacher scores on

several observation instruments varied substantially across the six MET states. Polikoff concluded that some states' tests were more correlated with teacher performance on some classroom observation instruments than others; however, "the reasons for these differences are not well known, but they should be an intense focus of study in the coming years" (301). Polikoff suggested several hypotheses, including that variability in correlations may be due to characteristics of the test items or content (e.g., low-level or procedural versus higher-level cognitive demand, or multiple-choice versus open-response item format). We explore these hypotheses in the current study.

Theory of Instruction and Student Learning Underpinning the MQI

In this article, we use the MQI as an example of a mathematics-specific observation instrument emphasizing conceptually oriented instruction to illustrate the potential contrasts in the teaching and learning goals valued on this type of instrument versus those valued on standardized tests. We describe the theory of instruction and student learning underpinning the MQI.

Underlying a theory of instruction is a vision of what constitutes high-quality student learning (Grossman et al. 2014). Although there is a history of debate in mathematics education regarding the relative importance of students' conceptual versus procedural learning (e.g., Schoenfeld 2004), in recent years scholars have argued that procedural knowledge can be superficial or deep (Star 2005, 2007), and influential policy documents have expressed the view that high-quality student learning encompasses both rich conceptual understanding and procedural fluency and skill development. Moreover, these are complementary and intertwined (e.g., Common Core, National Research Council's *Adding It Up* [2001]).

Prior research has examined which elements of instruction support the development of high-quality student learning in mathematics. In their influential article, Hiebert and Grouws (2007) reviewed the literature on the effects of classroom mathematics teaching on student learning in the areas of conceptual understanding, defined as "mental connections among mathematical facts, procedures, and ideas," and skill efficiency, defined as "the accurate, smooth, and rapid execution of mathematical procedures" (380). A limitation of this conceptualization is that it leaves out other important goals, such as the ability to apply procedural skills flexibly and strategically in novel situations (e.g., Star 2005, 2007; Star et al. 2015); however, it nonetheless emphasizes two important and longstanding goals of school mathematics learning. Hiebert and Grouws (2007) identified two related sets of instructional interactions, which we refer to as "conceptually oriented" and "skill-efficiency oriented." Two key elements of conceptually oriented instruction were found to support students' conceptual

learning: (1) teaching attending explicitly to concepts or connections among facts, procedures, and ideas and (2) students' efforts to make sense of important mathematics (e.g., Boaler 1998; Carpenter et al. 1989; Stein and Lane 1996).

Regarding whether conceptually oriented or skill-efficiency-oriented instruction better supported students' skill learning, Hiebert and Grouws's (2007) findings were less conclusive. They write, "One place in which the complex nature of teaching and learning becomes apparent is in the effects of conceptually oriented teaching on skill learning. Many of the reports on the conceptual development of students also indicate that their skills increased at a level equal to or greater than students in the control groups. . . . Apparently, it is not the case that only one set of teaching features facilitates skill learning and another set facilitates conceptual learning. In this case, two quite different kinds of features both seem to promote skill learning" (390). The research suggested that skill development could also be bolstered under a quite different set of instructional conditions, those that are skill-efficiency oriented. Skill-efficiency-oriented instruction involved rapid-pace, short-answer, targeted questions from teachers and students completing large numbers of practice problems (Hiebert and Grouws 2007).

This lack of certainty about whether conceptually oriented teaching is more effective than or equally effective as skill-efficiency-oriented teaching at promoting skill learning makes the issue of alignment more complicated. Consider the (common) scenario in which teacher evaluations are based predominantly on classroom observation scores and relatively rote standardized tests. If conceptual teaching is, in fact, more effective than skill-efficiency teaching at improving skills, then teachers using conceptual teaching should receive scores on both classroom observations and student tests that are high relative to those using skill-efficiency teaching (although perhaps not as high as they should be, as noted below). However, if conceptually oriented teaching is actually equally effective as skill-efficiency-oriented teaching at improving skills, then a teacher using conceptually oriented teaching could potentially receive a conflicting evaluation report, with high classroom observation scores but average test score gains similar to those of his or her peers using skill-efficiency-oriented teaching methods. In either case, if we hypothesize that conceptually oriented teaching methods also have greater affordances for students' conceptual learning than do skill-efficiency-oriented teaching methods, then conceptually oriented teachers' students' test score gains on skills-oriented tests may underestimate the impacts that these teachers have on student learning because conceptual outcomes are not measured on the tests. Related to this, a second potential misalignment issue under this scenario is that skills-oriented test content is misaligned to standards advanced by the National Council of Teachers of Mathematics (NCTM) and the Common Core State Standards initiative (National Governors Association Center for Best Practices, Council of Chief State School Officers 2010; NCTM

2000). See figure 1 for a graphical depiction of the potential areas of test–observation misalignment under these hypotheses.

In contrast with standardized tests, which often measure low-level skills (Resnick et al. 2004), classroom observation instruments are often grounded in a theory of instruction aligned with conceptually oriented instruction, emphasizing the quality of teacher–student interactions and the development of content understanding via engagement in cognitively demanding activities (Grossman et al. 2014; Seidel and Shavelson 2007). Content-generic instruments (e.g., Framework for Teaching [Danielson 2011]; CLASS [Pianta et al. 2010]) are designed to assess instructional interactions across subject areas, whereas content-specific instruments assess subject-specific instructional quality (Charalambous et al. 2014). Examples of mathematics-content-specific instruments include the Instructional Quality Assessment (IQA; Boston 2012), which is rooted in research on challenging mathematical tasks and maintaining cognitive demand (e.g., Stein and Lane 1996). The Reformed Teaching Observation Protocol (RTOP; Sawada et al. 2002) was developed to measure instructional alignment with mathematics reforms that emphasized a problem-solving approach. (e.g., NCTM 1995, 2000).

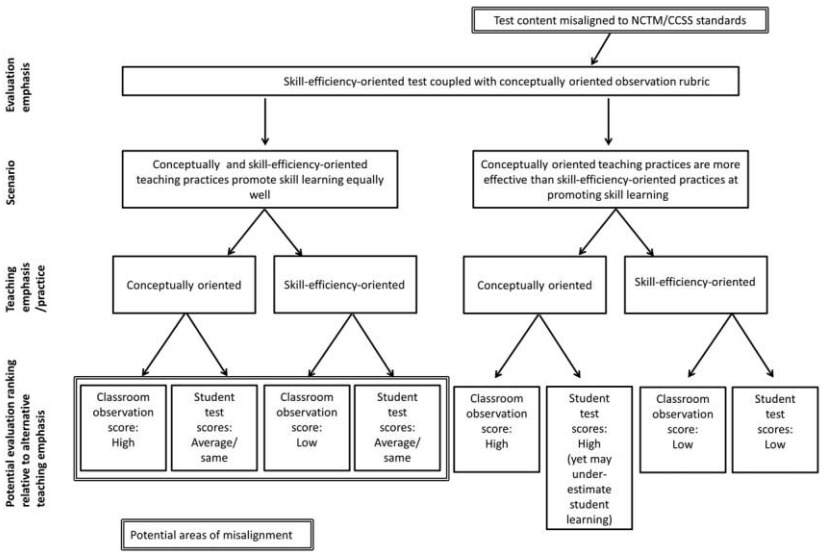

FIG. 1.—Logic model for hypothesized relationship between instructional practices captured by classroom observation and student learning as measured by assessments.

The MQI grew out of research on the importance of teachers' mathematical knowledge for teaching and precision in mathematical discussions and explanations. For the MQI, high-quality student learning involves reasoning and making meaning about the mathematical content, with the goal of developing conceptual understanding, adaptive reasoning, and procedural fluency. Note that we do not argue that the MQI instrument reflects all aspects of strong teaching as reflected in the mathematics education literature. Indeed, there are many alternative perspectives beyond those reflected in the MQI that could be used to evaluate the quality of teachers' instruction, such as gauging student discourse (e.g., O'Connor 1998) and charting in detail the cognitive demand of unfolding tasks (e.g., Stein and Lane 1996). It was not possible for us to measure all of the relevant constructs, and important ones are certainly left out. Nonetheless, we argue that the MQI is appropriate for our analyses because it measures several constructs that the mathematics education literature has identified as related to high-quality instruction in mathematics, as noted below.

Specifically, the MQI is grounded in a theory of ambitious instruction that emphasizes the importance of three core areas of instructional interactions: (1) teacher–students, (2) teacher–content, and (3) students–content (see figure 2; Hill et al. 2008). *Ambitious instruction* refers to instructional practice that is intellectually demanding and attentive to students' work (Cohen 2011). See table 1 for a description of the MQI dimensions and items. In interaction area 1, teacher–students, the MQI captures teachers' skill in responding to students'

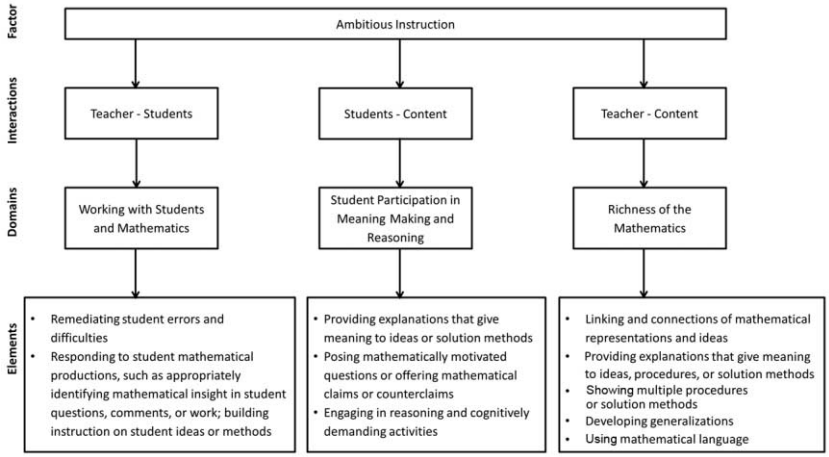

FIG. 2.—MQI instructional dimensions for the ambitious instruction factor

TABLE 1

Elements Included in the Mathematical Quality of Instruction (MQI), Ambitious Instruction Factor

Dimension	Description of Teaching Practices	Example Items
Working with students and mathematics	Captures the extent to which teachers interpret and respond to students' mathematical productions, including utterances, written work, and mathematical errors.	Responding to student mathematical productions, such as by building instruction around students' strategies or ideas or identifying and distilling mathematical points from students' utterances or work Remediating student errors and difficulties: substantially addressing students' misconceptions and difficulties with the mathematical content
Student participation in meaning making and reasoning	Captures the extent to which there is evidence in the classroom of students' involvement in cognitively activating work.	Engaging in reasoning, such as by identifying and explaining patterns or making connections across multiple representations, concepts, or solution strategies Making mathematical claims or counterclaims or asking mathematically motivated questions Providing explanations
Richness of the mathematics	Captures the depth of the mathematics students are provided the opportunity to learn, with attention to the meaning of facts, procedures, and concepts or to important mathematical practices.	Explanations: providing mathematical meaning to facts, procedures, concepts, or solution strategies. Multiple procedures or solution methods: comparing or considering multiple solution strategies for a problem Linking and connections: drawing explicit links among concepts, representations, and procedures. Developing generalizations: building generalizations of mathematical facts, procedures, or concepts using specific examples Mathematical language: using dense and precise mathematical language

NOTE.—For a detailed discussion of the instrument and items, see Hill et al. (2008).

mathematically substantive productions, including utterances, written work, and mathematical errors. For example, teaching practices in this MQI domain include the skill with which teachers identify mathematical insight in specific student questions, comments, or work and with which they build instruction on student ideas or methods. The elements in this domain are grounded in research on the importance of teachers' responses to students' mathematical questions and ideas (e.g., Borko et al. 1992; Cohen 1990; Stein et al. 1990), for example, the argument of Stein et al. (2008) for the importance of teachers helping students to draw out the connections between the mathematical ideas underpinning the strategies they use.

Interaction area 2, teacher–content, captures the depth and richness of the mathematics offered to students, in which rich mathematics focus on the meaning of facts, procedures, and practices. The elements in this domain are drawn from research in the mathematics education literature that highlights the affordances of linking and connecting mathematical representations, ideas, and procedures; considering multiple solution methods; and giving mathematical meaning to ideas or solution methods (e.g., Fennema and Franke 1992; Lloyd and Wilson 1998; Ma 1999; Stigler and Hiebert 1999). For example, one teaching practice evaluated in this dimension is the extent to which teachers draw explicit links among mathematical representations, ideas, and procedures using utterances, written work, and gestures (e.g., Richland 2015). Another practice measured in this domain is whether the teacher facilitates students considering and comparing multiple solution strategies or procedures for a single problem (e.g., Star and Rittle-Johnson 2009).

Interaction area 3, students–content, captures students' participation in cognitively activating work. Practices captured in this dimension include the extent to which students provide explanations (e.g., Henningsen and Stein 1997; Stein et al. 1996); pose mathematically motivated questions or advance mathematical claims or counterclaims (Yackel and Cobb 1996); draw connections among representations, concepts, or solution methods (Star et al. 2015); and engage in reasoning and other cognitively demanding activities, such as identifying and explaining patterns.

In work drawing on a subset of the data used in this analysis, Blazar (2015) found that the ambitious instruction measure from the MQI predicted gains in students' mathematics achievement on a researcher-developed test when teachers in multiple districts were pooled. An increase of 1.0 standard deviation (SD) in teachers' ambitious instruction scores was associated with a roughly 0.1-SD increase in students' mathematics achievement. That work did not explore differential relationships between ambitious instruction and student achievement across standardized assessments. For a detailed discussion of the MQI dimensions and supporting research, also see Blazar et al. (forthcoming) and Hill et al. (2008).

Contribution of the Current Study

Our work builds on prior studies in the following ways. First, our data allow us to investigate relationships between teacher observation scores, as measured by the MQI instrument, and student achievement across districts where students take different state standardized math tests but also a researcher-developed math test common to all students. Second, we conduct a formal coding analysis of tests used in the study to explore more systematically the hypothesis that tests' cognitive demand and item and format characteristics are related to test–observation misalignment.

Because we explore these possible explanations in a single study, we compile more evidence regarding certain explanations for varying relationships between student assessments and teacher observations than currently exists in the literature. However, our study design does not permit us to disentangle the specific impact of test–observation misalignment. In addition to the variables we have measured, educational outcomes necessarily reflect the complex interplay of students and teachers, resources, and the broader social, cultural, and historical context (Cohen et al. 2003). We do not address the contextual issues in the current study but focus specifically on the issue of test–observation misalignment. As such, the current study is a preliminary examination of an important yet limited set of constructs that may bear on relationships between teacher observation scores and student achievement scores. We return to this issue in the discussion.

Research Question

We ask the following research question: Does teacher performance on observation instruments predict student achievement equally well across district contexts? To the extent that relationships between teacher observation scores and student achievement scores vary across districts, we explore possible factors that may contribute to this variability.

Method

Sample

Data for this study come from two research projects that spanned the 2010–11 and 2011–12 school years and focused on fourth- and fifth-grade mathematics teachers. The first study is a large-scale project examining instructional quality

in four anonymous districts (henceforth, districts 1 through 4) from three states. The second is a randomized controlled trial of a mathematics professional development program in one anonymous district (henceforth, district 5); this study collected data on teachers and students similar to the first project. In the first project, schools were recruited based on district referrals and size (with a minimum of two teachers in each school in each sampled grade). Of eligible teachers in these schools, roughly 55% agreed to participate. In the second study, we include only the treatment teachers, as MQI data were not collected for the control-group teachers during these school years. We restrict this sample to the teachers for whom we have observation data, state test score data, and researcher-developed assessment score data, for a total sample of 298 teachers and 6,780 students. From these 298 teachers, we collected 1,560 videotaped lessons of mathematics instruction across both school years.

In table 2, we present descriptive statistics for the sample teachers and their students. Relative to other students in the study, students in district 1 scored around average (0.01 SD), whereas those in districts 4 and 5 scored above average (0.21 and 0.07 SD, respectively), and those in districts 2 and 3 scored below average (-0.25 and -0.22 SD, respectively) on the researcher-developed baseline mathematics assessment.

As we report elsewhere (Hill et al. 2015), instructional policies also differed across districts. Districts 1 and 2, which were located in the same state, both used the same set of National Science Foundation–funded curriculum materials, *Investigations in Number, Data, and Space* (TERC 2008), which were developed to be aligned with the mathematics reforms promulgated in policy documents such as NCTM's (2000) *Principles and Standards of School Mathematics*. Prior to the study, district 1 had also been engaged in an intensive, decade-long instructional improvement process centered on demanding mathematical pedagogy, featuring sustained teacher professional development and coaching. District 3 used a more traditional set of curriculum materials published by Harcourt Brace and offered teachers relatively few opportunities for professional development. District 4 was engaged in implementing a new high-stakes teacher evaluation program during the study period. Although district 4 also used a National Science Foundation–funded curriculum, *Everyday Mathematics* (UCSMP 2007), and had a district mathematics coordinator, the district lacked a coordinated plan to improve elementary mathematics instruction. Finally, district 5 used the curriculum *Math Expressions*, generally considered to be more aligned with the reforms recommended in the NCTM standards documents than the curriculum in district 3 but less so than the curricula used in districts 1, 2, and 4. Over the course of study, district 5 experienced a high degree of turnover in leadership, including the superintendent and the math coordinator. At the same time, there were multiple efforts to provide reform-oriented professional development to teachers both during and prior to this study.

TABLE 2

Sample Descriptive Statistics

	DISTRICT				
	1	2	3	4	5
Teachers:					
Male	29.23	13.46	13.64	9.20	0
African American	20.00	1.92	59.09	18.68	11.11
White	63.08	80.77	29.55	70.33	74.07
Other (Asian, Hispanic)	15.15	9.43	7.14	4.35	14.81
Experience	10.64	10.44	8.72	11.69	8.26
Traditionally certified	74.24	86.79	47.62	90.22	96.30
Alternatively certified	6.06	0	23.81	4.35	3.70
No certification	10.61	7.55	16.67	1.09	0
Observations	66	53	44	92	27
Students:					
Male	47.79	51.80	46.48	49.97	49.73
African American	37.46	51.95	66.51	31.22	57.85
Asian	13.27	3.42	2.38	9.41	3.35
Hispanic	37.35	12.90	11.80	21.69	6.52
White	6.70	27.44	17.64	33.79	23.81
Eligible for free or reduced-price lunch	82.19	72.70	61.26	49.07	72.08
Special education status	12.90	11.65	9.65	9.41	9.17
Limited English proficiency	36.43	23.21	10.01	11.74	0.71
Researcher-developed assessment baseline achievement	.01	−.25	−.22	.21	.07
Observations	1,628	2,077	839	1,669	567

NOTE.—Teacher characteristics generated from a subsample of those who completed a baseline survey.

Data

Mathematical quality of instruction.—The MQI instrument (Hill et al. 2008), described earlier, evaluates teachers' instructional practices across two dimensions: "ambitious instruction" and "errors and imprecision." Instrument developers originally envisioned four MQI dimensions, based largely on theory (Hill et al. 2008); however, factor analyses and more recent substantive interpretations by instrument developers (Blazar et al., forthcoming) have collapsed these to two. For our analyses, we focus on teachers' ambitious instruction scores as opposed to teachers' errors and imprecision scores because the former dimension measures the inquiry-oriented instruction and activities that occur in the classroom (e.g., linking multiple representations, solving a problem in mul-

tiple ways, student and teacher explanations, teachers' use of student ideas; see fig. 2).

Teacher ambitious instruction scores were generated from videotaped lessons of instruction captured over the course of 2 years. Teachers averaged 5.23 videotaped lessons (mode = 6), allowing for sufficient levels of predictive reliability (Hill, Charalambous, Blazar, et al. 2012).[1] Raters had a background in mathematics or mathematics education, passed a certification exam, and completed ongoing calibrations. Two raters scored teachers' instruction on each MQI item for each 7.5-minute lesson segment on a scale from 1 (low) to 3 (high). We estimate two reliability statistics for ambitious instruction. First, we calculate the percentage of all instances in which the two raters assigned a teacher the same score on an item in a 7.5-minute lesson segment; this statistic gives us a general measure of interrater reliability. Raters demonstrated exact agreement on 74% of all possible scoring instances. Second, we calculate the amount of variance in teacher scores attributable to the teacher (i.e., the intraclass correlation) as opposed to other sources of variation (i.e., lessons), adjusted for the modal number of lessons observed per teacher. Our estimate of 0.69 approximates conventionally acceptable levels (0.7) and is higher than those generated from similar studies (Bell et al. 2012; Kane and Staiger 2012).

Given that teachers have differing numbers of lessons from which to construct ambitious instruction scores, we use empirical Bayes estimation to shrink scores back toward the mean based on their reliability (Raudenbush and Bryk 2002). We calculate teacher-level scores by first averaging teacher performance on the MQI items within the ambitious instruction domain across segments and items. We then specify the following hierarchical linear model, in which lessons are nested within teachers:

$$\text{AMBITIOUS\_INSTRUCTION}_{lj} = \mu_j + \varepsilon_{lj}, \quad (1)$$

where μ_j is a random effect for teacher j, and ε_{lj} is a residual for each teacher's lesson. We use in analyses standardized estimates of the teacher-level random effect as the final MQI classroom-observation score.

Student demographic and test-score data.—Most student data come from district administrative records, including student–teacher links from verified classroom rosters, student demographic information, and end-of-year mathematics and reading scores for standardized state tests completed in 2009, 2010, and 2011.

In addition, students completed a researcher-developed mathematics assessment, the Upper-Elementary Mathematics Assessment Modules (Hickman et al. 2012), at the beginning and end of the school year.[2] This assessment was developed through a joint venture between Harvard University and the Educational Testing Service and is designed to be aligned with the Common Core State Standards for Mathematics. We describe these tests in more detail later.

Analyses

Our analysis is organized as follows. First, we conducted a quantitative analysis of the relationships between teachers' MQI scores and student achievement scores across districts. As we discuss below, the quantitative modeling revealed substantial differences among districts in the relationships between MQI and student achievement outcomes when using state standardized tests but not when using the common researcher-developed assessment.

Next, we examined whether features of the various state standardized tests might contribute to the observed differences. Specifically, we compared all of the state tests with the researcher-developed assessment on several key dimensions of potential importance for the relationship with MQI scores, which we describe below. We present results concerning whether cross-test differences in test demand may contribute to the differences in the relationships between teacher observation scores and student achievement outcomes.

Our method for exploring possible reasons for the observed cross-district differences is similar to that of Papay (2011) in that we are able to evaluate whether the pattern of results in our data is logically consistent with our hypotheses. With the limited number of districts, however, our design does not allow us to definitively test for the impact of each element. Given this limitation, we consider our findings suggestive, rather than definitive, of possible explanations for the variability in relationships between student test scores and teacher observation scores.

Quantitative modeling.—Other similar studies quantitatively test the relationship between teacher observation scores and student achievement scores by calculating a value-added score for each teacher and then correlating this with the teacher's observation score. We take a slightly different approach by including the observation score in our value-added model. This allows us to test formally for differences in the relationship between MQI and student achievement across districts. We estimate the following equation:

$$\begin{aligned} A_{jt} = {} & \textstyle\sum_{d=1}^{5} \beta_d \mathrm{MQI}_k \times D_d + \zeta f(A_{jt-1}) + \sigma \mathrm{X}_{jt} + \theta C_{ct} + \alpha S_{sgt} \\ & + \pi G_{gt} + \varphi D_d + \mu_k + \epsilon_{jt}, \end{aligned} \tag{2}$$

where the outcome of interest, A_{jt}, represents math score for student j at time t. We regress our outcome of interest on interactions between each teacher k's MQI score on ambitious instruction and district d, ($\mathrm{MQI}_k \times D_d$); a function of student's prior achievement, $f(A_{jt-1})$; a vector of student demographic variables, X_{jt}, including gender, race, free or reduced-price lunch eligibility, special education status, and limited English proficiency; vectors of student demographic and test score variables aggregated to the class, C_{ct}, and school grade, S_{sgt}, levels; grade-by-year fixed effects, G_{gt}, that account for different scaling of tests across

grades and years; and district fixed effects, φD_d.[3] Given the nested structure of the data, we include a random effect for teachers, μ_k, as well as a student-level error term, ϵ_{jt}. We fit models using all years of available test-score data to increase the precision of our estimates (e.g., Goldhaber and Hansen 2012; Schochet and Chiang 2013). We also limit the sample to those classes in which less than 50% of students have special education status, where 50% or less are missing scores for the prior achievement vector, and after all other restrictions, where there are at least five students. We perform these restrictions to exclude atypical classrooms from our value-added model; the restriction removes 5% of classrooms and results in an average class size of approximately 18 students.

Our parameters of interest are in the vector β_d, which estimates the relationship between teachers' MQI scores and student achievement for each district. To test whether these parameters differ across districts, we conduct a series of post hoc general linear hypothesis (GLH) tests that compare all pairwise relationships. We fit this model separately for the researcher-developed assessment and the state standardized math tests.

Exploring variability in MQI/achievement relationships across districts.—To explore whether our data are consistent with the hypothesis that cross-test differences may contribute to the cross-district differences in the relationships between teacher observation scores and student achievement outcomes, we coded state standardized math tests from each of the sampled districts and grades (fourth and fifth) and the researcher-developed test. We completed the test coding as follows.

First, we gathered information about the state test administered in each district from publicly accessible websites. Because districts 1 and 2 are in the same state, test information and materials are the same. As we were able to recover test information from multiple years, and investigation of state test blueprints suggested that characteristics of each test largely remained constant from year to year, we analyzed all test items from a single randomly selected school year. Of the four state tests, only those from district 5 were publicly available in their complete and original form. For the other three state tests, we included in our analyses all available publicly released items in the school year analyzed and cross-referenced these against state test blueprints.[4] We examined 95 items, on average, from each state test, across both grades. For the researcher-developed assessment, we randomly selected 1 year and one form from each of the sampled grades and coded these tests in their entirety.

The authors jointly coded these tests on two dimensions: item format (AERA, APA, and NCME 1999) and alignment with the MQI instrument (see online app. A for a complete description of these coding schemes and procedures). To operationalize degree of alignment with the MQI, for each test, we assessed whether each test item demanded a high, medium, or low level of student engagement with MQI task cognitive activation, which captures the degree of

difficulty and challenge associated with a task. Items rated "low" are those that asked students to recall and apply procedures or to reproduce known facts or formulas. Items rated "high" asked students to engage with content at a high level of cognitive activation, such as by determining the meaning of mathematical concepts or relationships or drawing connections among representations or concepts. To measure item format, we used the test item format categories described by the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME 1999) to code the proportion of items on each test that were multiple choice, short answer (i.e., constructed response items; items that asked students to bubble in an answer), and open ended (i.e., items requesting longer responses, short essays, or explanations of answers). To summarize these analyses, we created a summary metric for overall test demand. Each assessment was assigned a score of high, mid, or low on each dimension, based on our coding. The overall test demand codes reflected our holistic judgment of the overall levels of expected skills placed by the student achievement tests on the dimensions.

Before beginning this process, we individually reviewed coding manuals, scored a subset of items from an external test, and held meetings to calibrate discrepant scores. In the final coding process, we coded all items jointly. We examined all items as a group and only moved to the next item once we agreed on all codes. During the coding process, we were not blind to the state test from which items were derived. To minimize bias that may have resulted from this knowledge, we frequently compared our scores for specific items with the scores we assigned to similar items on other tests, checking for alignment.

Results

We begin by presenting results from our quantitative analysis, in which we examine cross-district differences in the relationships between MQI and student achievement on both the researcher-developed and state standardized tests. Next, we present results from our exploratory analyses, examining potential factors contributing to the observed cross-district variability in relationships, including cross-district variability in the distribution of observed instructional quality and reported test prep behaviors and differences in expected skills across student achievement tests.

Quantitative Models

In table 3, we present estimates of the relationships between MQI and student achievement on both the researcher-developed and state standardized tests for each district. In the first column, we present results for the researcher-developed

TABLE 3

Estimates of the Relationship between Instructional Quality on the Measure of Ambitious Instruction and Student Achievement across Districts

District	Researcher-Developed Mathematics Assessment	State Standardized Tests
1	.035 (.024)	.081** (.031)
2	.095* (.039)	.126** (.041)
3	.046 (.040)	−.004 (.039)
4	.021 (.032)	−.053 (.040)
5	−.015 (.030)	.028 (.038)

NOTE.—Each column represents a separate regression model. Robust standard errors in parentheses. Sample contains 298 teachers and 6,780 students.

* $p < .05$.

** $p < .01$.

assessment. We observe that the relationships between MQI ambitious instruction and student achievement on the researcher-developed assessment are generally small and not statistically significant.

In the second column of table 3, we present results for the state standardized assessments. Comparing the second column with the first column, we observe that, compared with the results for the researcher-developed assessment, results for the relationship between MQI observation scores and student performance on state standardized assessments are more varied across districts—with estimates ranging from −0.053 in district 4 to 0.126 in district 2. The relationship between MQI ambitious instruction scores and student achievement on state standardized tests was significantly different from zero only in districts 1 and 2 (which are in the same state).

In table 4, we present the results of a series of post hoc GLH tests comparing the coefficients from the models in table 3 across districts. As seen in the first column, on the researcher-developed mathematics test, we observe only one statistically significant difference in the MQI and student achievement relationship across districts—between districts 2 and 5 ($p = .03$).

By contrast, examining the second column, we observe more cross-district differences using the state standardized assessments. Note that we do not find differences in the relationship between MQI scores and student achievement outcomes on the state test in district 1 versus in district 2. Recall that these two districts are in the same state, so students take the same state test. By contrast, the relationships between MQI and student achievement on the state standardized tests are stronger in district 1 than in districts 3 and 4 ($p = .08$ and $p = .01$, respectively). In addition, the relationships between MQI and student

TABLE 4

Estimates of the Relationship between Instructional Quality on the Measure of Ambitious Instruction and Student Achievement across Districts

p-Value on Post Hoc GLH Tests	Researcher-Developed Mathematics Assessment	State Standardized Tests
All district coefficients equal	.40	.02[+]
District 1 = district 2	.20	.39
District 1 = district 3	.82	.08[+]
District 1 = district 4	.73	.01[+]
District 1 = district 5	.20	.30
District 2 = district 3	.39	.02[+]
District 2 = district 4	.14	<.01[+]
District 2 = district 5	.03[+]	.08[+]
District 3 = district 4	.64	.38
District 3 = district 5	.21	.57
District 4 = district 5	.40	.15

[+] Statistically significant difference.

achievement are stronger in district 2 than in districts 3, 4, and 5 ($p = .02$, $p < .01$, and $p = .08$, respectively).

Differences in Expected Skills across Student Achievement Tests

To summarize the results from our test coding and analyses of the differences in expected skills across student achievement tests, we present results from a summary metric for overall test demand (see table 5). The results from our coding of the dimensions that contributed to the summary measure are found in online appendix B.

TABLE 5

Overall Test Demand

	TEST CHARACTERISTICS		
	Item Format	Alignment to MQI: Task Cognitive Activation	OVERALL DEMAND
State tests:			
Districts 1 and 2	High	High	High
District 3	Mid	Mid	Mid
District 4	Low	Mid	Mid/low
District 5	Low	Mid	Mid/low
Common assessment:			
All Districts	Mid	High	High/mid

Overall, we judged the state test used in districts 1 and 2 to be highest in demand, followed by the researcher-developed assessment. We judged the state test in district 3 to be of moderate demand and those in districts 4 and 5 to be of moderate to low demand. These results seem somewhat consistent with the hypothesis that stronger relationships between teacher observation scores and student test scores may be observed when characteristics of the tests used to measure student achievement are more aligned with characteristics of the observation instruments. As noted, we observed the strongest relationships between MQI scores and student achievement in districts 1 and 2, where the demand of the state test was relatively high. In districts 3 through 5, where the overall demand of the state tests was lower, we generally observed weaker relationships.

However, when student achievement is measured using the researcher-developed assessment, the relationship between the MQI and student achievement in district 2 is significantly different from that of district 5, despite the fact that the researcher-developed assessment presented the same demands to all students in all districts. Consequently, variation in test characteristics does not appear to be solely responsible for the observed differences.

Discussion

In summary, we found that relationships between teachers' MQI observation scores and their students' achievement on state mathematics tests differed by district. We explored whether differential characteristics of the tests, or test–observation misalignment, might contribute to this variability. If alignment between test content and instructional quality drives these relationships, then we would expect to see two things: (1) variability in the relationship between MQI scores and student achievement across districts on state standardized tests, with stronger relationships in districts where state tests are better aligned to the MQI, and (2) no variability in the relationship between MQI scores and student achievement across districts on the researcher-developed assessment.

These analyses are exploratory in nature, but our findings do suggest some support for the test–observation misalignment hypothesis. We found that the relationships between teachers' MQI scores and their students' achievement on state standardized tests were the strongest in districts 1 and 2. These two districts also had the most demanding state standardized test, as rated on the summary metric of overall test demand. In particular, this test had the highest proportion of open-ended items in the sample, with nearly a quarter of all items coded as open ended. This compared with few or no open-ended items on the other states' tests. In addition, although the average levels of cognitive demand for all state tests were relatively low, with most items coded as "perform procedures,"

the state test for districts 1 and 2 was distinguished by the highest mean level of cognitive demand of the state tests in the sample.

By contrast, in districts 3, 4, and 5, relationships between teachers' MQI scores and their students' state standardized test scores were smaller and not statistically significant. These districts also used state tests whose overall demand was judged to be middling to low. The state tests used in districts 3, 4, and 5 demonstrated middling cognitive demand and alignment with the MQI instrument and relied heavily on multiple-choice items. The tests in districts 4 and 5 were entirely multiple choice.

At the same time, we also observed one cross-district difference in the relationship between MQI and student achievement on the researcher-developed assessment. Because the researcher-developed assessment presents the same demands for all students in all districts, and the MQI presents the same demands for all teachers in all districts, this data point is inconsistent with the notion that test characteristics could be the sole reason for variability in relationships.

Cross-District Differences in Relationships between Teacher Observation Scores and Student Achievement: Potential Alternative Explanations

Considering the current findings in the broader district policy context, we speculate that the contributing factors we discuss are likely intertwined: cross-district differences in relationships between teacher observation and student achievement scores may arise in part from the interplay between characteristics of state tests and the cross-district differences in teacher instructional quality, potentially spurred by how instruction is measured and implemented in school districts.

Elsewhere, we describe how measures of teachers' instructional quality, including ambitious mathematics instruction, differ substantively across the districts assessed in this study (Hill et al. 2015). Blazar et al. (2016) also found that teachers' value-added scores, or contributions to their students' math performance as measured on state standardized assessments, signaled different sets of mathematics-specific instructional practices across these districts. These patterns were not explained away by observable background characteristics of teachers, including their math content knowledge, prior course taking in mathematics, and certification pathway, suggesting that factors beyond labor-market sorting and, instead, internal to district context likely played a key role.

These findings are consistent with prior research that has emphasized the important roles that local instructional policies, including testing policy, curriculum, and professional development opportunities, play in influencing teachers' instruction (e.g., Booher-Jennings 2005; Coburn 2005; Cohen and Hill 2000).

With regard to the relationship between high-stakes testing and instructional practices, for example, Booher-Jennings (2005) found that when confronted with a new high-stakes testing policy, teachers in one urban elementary school altered their instructional practices markedly to attempt to improve test scores, such as by providing targeted tutoring to students believed to be on the border of passing the test while requiring the rest of the class to perform seatwork.

As noted, districts in the current study varied in their instructional policies, which may have both reflected and shaped teachers' differential responses to high-stakes testing across districts. Recall that districts 1 and 2, which were located in the same state, used the same set of inquiry-oriented curriculum materials and provided teachers with aligned professional development (especially so in district 1). In this context, we speculate that perhaps in these districts, teachers may have been incented to provide ambitious instruction aligned both to their state's high-demand test and to the MQI instrument and may have been provided with curricular and professional development resources to support this challenging pedagogy. This interplay between facets of the test and the way in which instruction was implemented may have supported alignment between student achievement and teacher observation outcomes. Alternatively, in district 3, where a new high-stakes teacher-evaluation program was being implemented, teachers may have had incentives to engage in more test-prep instruction that was well aligned to their state's relatively basic skills-oriented test but poorly aligned to the ambitious instruction goals valued in the MQI observation instrument. Here, properties of the state test may have interacted with the district policy climate to hinder alignment between student test and teacher observation scores.

Another possibility is that factors beyond the classroom, such as parent involvement, may both respond to instructional quality and influence test scores while also varying by district. For example, it is possible that parents could respond to weak instructional quality in their children's classrooms by supplementing with extra tutoring or after-school programs (Hill et al. 2008). Such practices could attenuate the observed relationship between observed instructional quality and student achievement outcomes and could also vary by district if the availability or quality of supplemental programs varies across districts. As noted, however, these observations are speculative. Additional research is needed to determine how specific factors may be contributing to variability in alignment.

Limitations

This study presents several limitations that point to potential directions for future research. First, we are unable to empirically model and test the strength

of test–observation misalignment to variability in the relationship between MQI and student achievement because the dimensions we examine are perfectly collinear with each district. Second, the current study is limited to the case of the MQI. As districts overhaul their teacher evaluation systems, ushering in new assessments and observation instruments, future research is needed to explore whether the observed relationships between characteristics of tests and observation scores are replicated for other observation instruments and contexts. In addition, further work is needed to parse in detail the district-level factors that mediate the relationships between test scores and observation outcomes, which are likely overlapping and intertwined. For example, future studies should analyze districts' adopted curricula in depth for their alignment with inquiry-oriented instructional practices, as valued in observation instruments, and for their cognitive demand, to determine how curriculum may moderate the relationship between classroom observations and test scores. In addition, complementary ethnographic studies that analyze cross-district differences in how math instruction and student achievement play out when embedded in policy tensions could shed light on these issues.

Conclusion and Future Directions

In the current study, we find evidence that the relationship between instructional quality, as judged by classroom observations, and student achievement on state tests is differs by district. We suggest that one factor that might contribute to variability in these relationships may be the sensitivity of the observation instrument according to state test. When observation instruments and student assessments are more aligned, stronger relationships between instructional quality and student achievement may result.

In a broader theoretical context, the current study raises questions about the goals of classroom observation instruments versus standardized achievement tests, what each type of assessment deems important, and what each offers and does not offer as we work to understand and assess teacher quality. Specifically, the type of learning promoted in classroom observations versus on standardized tests may be in discord. The MQI and many other classroom observation instruments begin with the goals of fostering students' conceptual understanding and skill learning (Grossman et al. 2014). To assess teachers' effectiveness in providing students opportunities to meet these learning goals, classroom observation instruments often measure instructional practices that promote students' conceptual learning, in line with the conceptually oriented instructional practices identified in the mathematics education literature (Hiebert and Gruows 2007). By contrast, whereas state standardized tests are also often intended to

evaluate students' progress toward both conceptual and skill-efficiency learning goals, as envisioned in state standards (Webb 1999), in reality, they often disproportionately measure lower level skills (Resnick et al. 2004). Evaluating teachers using student standardized test scores may thus result in high evaluation scores for teachers using skill-efficiency-oriented instruction but not reward teachers using more conceptually oriented instruction for the relative affordances they offer to students' conceptual learning and richer procedural understanding.

How might the current study's findings be applicable to the work of states and districts as they seek to implement more rigorous curriculum standards, such as the Common Core? The implementation of more rigorous curriculum standards such as the Common Core seems poised to streamline the content of student assessments as states and districts move to adopt new assessments (e.g., Partnership for Assessment of Readiness in College and Careers, Smarter Balanced Assessment Consortium) designed to be aligned with these more rigorous standards (Grossman et al. 2011). These assessments are indeed being designed with the goal of improving some of the shortcomings of the current state tests that we noted in the current study, such as disproportionate emphasis on lower level skills and the multiple-choice format. To the extent that these new assessments succeed in their goals, a problem that we detect in the current study—namely, the misalignment between the lower level skills often demanded on student tests and the higher level skills teachers are expected to foster on new observation instruments—may be minimized. However, it is not clear to date the extent to or speed with which states will move to replace their existing local tests with the new, more conceptually oriented assessments. Indeed, a report from the National Governors Association Center for Best Practices expresses concern that states may simply tack on new assessments to their existing testing regimens, resulting in overtesting while retaining some of the problems associated with the current tests (Grossman et al. 2011).

In either case, concerns about the alignment of test-based measures of effectiveness and teacher observations remain pressing for policy. Although perfect alignment between these measures is not expected, varying relationships may prove problematic for teacher-evaluation policy and practice. For example, under current conditions, a teacher who focuses instruction on lower level tasks in anticipation of a basic-skills-oriented state standardized test could potentially receive a conflicting evaluation report indicating high value added on a state standardized test coupled with a low teacher observation score. In such a case, the teacher may face unclear guidance on how to improve practice. The current results suggest that districts may need to examine their classroom observation instruments for alignment with their high-stakes student assessments to aid both teachers and districts in better supporting teachers for instructional effectiveness.

Notes

1. Teachers were allowed to select the dates for videotaping in advance. Project managers required that teachers select a typical lesson and exclude days on which students were taking a test. It is possible that these videotaped lessons are unique from teachers' general instruction, but prior research suggests that when teachers were given discretion to choose their best classroom videos from a set, the chosen videos provided essentially similar information about teachers' instructional quality as the videos that were not chosen (Ho and Kane 2013).

2. In district 4, teachers began the overarching study in the second semester of the 2010–11 school year. Because characteristics of the tests themselves are a major component of our analyses, we exclude student-level administrative data from district 4 in this school year. We use complete student-level administrative data from the district in the 2011–12 school year.

3. We also consider a fully interacted model that allows all parameters on covariates to vary by district. However, we find that results follow the same pattern and, therefore, use this more parsimonious model.

4. The exception to this was the test from district 4, for which we included all publically available items that we could recover, regardless of year, because of the lower number of released items.

References

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

Bell, Courtney, Drew Gitomer, Daniel McCaffrey, Bridget Hamre, Robert Pianta, and Yi Qi. 2012. "An Argument Approach to Observation Protocol Validity." *Educational Assessment* 17 (2–3): 62–87.

Blazar, David. 2015. "Effective Teaching in Elementary Mathematics: Identifying Classroom Practices That Support Student Achievement." *Economics of Education Review* 48:16–29.

Blazar, David, David Braslow, Charalambos Y. Charalambous, and Heather C. Hill. Forthcoming. "Attending to General and Mathematics-Specific Dimensions of Teaching: Exploring Factors across Two Observation Instruments." *Educational Assessment.*

Blazar, David, Erica Litke, and Johanna Barmore. 2016. "What Does It Mean to Be Ranked a 'High' or 'Low' Value-Added Teacher? Observing Differences in Instructional Quality across Districts." *American Educational Research Journal* 53 (2): 324–59.

Boaler, Jo. 1998. "Open and Closed Mathematics: Student Experiences and Understandings." *Journal for Research in Mathematics Education* 29 (1): 41–62. doi:10.2307/749717.

Booher-Jennings, Jennifer. 2005. "Below the Bubble: 'Educational Triage' and the Texas Accountability System." *American Educational Research Journal* 42 (2): 231–68.

Borko, Hilda, Margaret Eisenhart, Catherine A. Brown, Robert G. Underhill, Doug Jones, and Patricia Agard. 1992. "Learning to Teach Hard Mathematics: Do Novice

Teachers and Their Instructors Give Up Too Easily?" *Journal for Research in Mathematics Education* 23 (3): 194–222.
Boston, Melissa. 2012. "Assessing Instructional Quality in Mathematics." *Elementary School Journal* 113 (1): 76–104.
Bridgeman, Brent. 1992. "A Comparison of Quantitative Questions in Open-Ended and Multiple-Choice Formats." *Journal of Educational Measurement* 29 (3): 253–71.
Carpenter, Thomas P., Elizabeth Fennema, Penelope Peterson, Chi-Pang Chiang, and Megan Loef. 1989. "Using Knowledge of Children's Mathematics Thinking in Classroom Teaching: An Experimental Study." *American Educational Research Journal* 26 (4): 499–531.
Charalambous, Charalambos Y., Andreas Komitis, Maria Papacharalambous, and Afroditi Stefanou. 2014. "Using Generic and Content-Specific Teaching Practices in Teacher Evaluation: An Exploratory Study of Teachers' Perceptions." *Teaching and Teacher Education* 41:22–33.
Coburn, Cynthia E. 2005. "The Role of Nonsystem Actors in the Relationship between Policy and Practice: The Case of Reading Instruction in California." *Educational Evaluation and Policy Analysis* 27:23–52.
Coggshall, Jane G., Claudette Rasmussen, Amy Colton, Jessica Milton, and Catherine Jacques. 2012. *Generating Teacher Effectiveness: The Role of Job-Embedded Professional Learning in Teacher Evaluation.* Washington, DC: National Comprehensive Center for Teacher Quality.
Cohen, David K. 1990. "A Revolution in One Classroom: The Case of Mrs. Oublier." *Educational Evaluation and Policy Analysis* 12:311–30.
Cohen, David K. 2011. *Teaching and Its Predicaments*. Cambridge, MA: Harvard University Press.
Cohen, David K., and Heather C. Hill. 2000. "Instructional Policy and Classroom Performance: The Mathematics Reform in California." *Teachers College Record* 102 (2): 294–343.
Cohen, David K., Stephen W. Raudenbush, and Deborah Loewenberg Ball. 2003. "Resources, Instruction, and Research." *Educational Evaluation and Policy Analysis* 25 (2): 119–42.
Daley, Glenn, and Lydia Kim. 2010. *A Teacher Evaluation System That Works*. Santa Monica, CA: National Institute for Excellence in Teaching.
Danielson, Charlotte. 2011. *Enhancing Professional Practice: A Framework for Teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.
Doyle, Walter. 1988. "Work in Mathematics Classes: The Context of Students' Thinking during Instruction." *Educational Psychologist* 23 (2): 167–80.
Fennema, Elizabeth, and Megan Loef Franke. 1992. "Teachers' Knowledge and Its Impact." In *Handbook of Research on Mathematics Teaching and Learning*, ed. Douglas A. Grouws. New York: Macmillan.
Firestone, William A. 2014. "Teacher Evaluation Policy and Conflicting Theories of Motivation." *Educational Researcher* 42 (2): 100–7.
Gamse, Beth C., Robin T. Jacob, Megan Horst, Beth Boulay, and Faith Unlu. 2008. *Reading First Impact Study*. Washington, DC: US Department of Education, Institute of Education Sciences.
Goldhaber, Daniel, and Michael Hansen. 2012. "Is It Just a Bad Class? Assessing the Long-Term Stability of Estimated Teacher Performance." *Economica* 80 (319): 589–612.
Good, Thomas L., and Jere E. Brophy. 2007. *Looking in Classrooms*. New York: HarperCollins.

Grossman, Pam, Julie Cohen, Matthew Ronfeldt, and Lindsay Brown. 2014. "The Test Matters: The Relationship between Classroom Observation Scores and Teacher Value Added on Multiple Types of Assessment." *Educational Researcher* 43 (6): 293–303.

Grossman, Tabitha, Ryan Reyna, and Stephanie Shipton. 2011. *Realizing the Potential: How Governors Can Lead Effective Implementation of the Common Core State Standards*. Washington, DC: National Governors Association Center for Best Practices.

Harris, Douglas N., William K. Ingle, and Stacey A. Rutledge. 2014. "How Teacher Evaluation Methods Matter for Accountability: A Comparative Analysis of Teacher Effectiveness Ratings by Principals and Teacher Value-Added Measures." *American Educational Research Journal* 51 (1): 73–112.

Henningsen, Marjorie, and Mary Kay Stein. 1997. "Mathematical Tasks and Student Cognition: Classroom-Based Factors That Support and Inhibit High-Level Mathematical Thinking and Reasoning." *Journal for Research in Mathematics Education* 28 (5): 524–49.

Hickman, Judy J., Jianbin Fu, and Heather Hill. 2012. *Technical Report: Creation and Dissemination of Upper-Elementary Mathematics Assessment Modules*. Princeton, NJ: Educational Testing Service.

Hiebert, James, and Douglas A. Grouws. 2007. "The Effects of Classroom Mathematics Teaching on Students' Learning." In *Second Handbook of Research on Mathematics Teaching and Learning*, ed. F. K. Lester Jr. Greenwich, CT: Information Age.

Hill, Heather C., David Blazar, and Kathleen Lynch. 2015. "Resources for Teaching: Examining Personal and Institutional Predictors of High-Quality Instruction." *AERA Open* 1 (4): 1–23.

Hill, Heather C., Merrie L. Blunk, Charalambos Y. Charalambous, Jennifer M. Lewis, Geoffrey C. Phelps, Laurie Sleep, and Deborah Loewenberg Ball. 2008. "Mathematical Knowledge for Teaching and the Mathematical Quality of Instruction: An Exploratory Study." *Cognition and Instruction* 26 (4): 430–511.

Hill, Heather C., Charalambos Y. Charalambous, David Blazar, Daniel McGinn, Mary Beisiegel, Andrea Humez, Matthew Kraft, Erica Litke, and Kathleen Lynch. 2012. "Validating Arguments for Observational Instruments: Attending to Multiple Sources of Variation." *Educational Assessment* 17 (2–3): 88–106.

Hill, Heather C., Charalambos Y. Charalambous, and Matthew A. Kraft. 2012. "When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study." *Educational Researcher* 41 (2): 56–64.

Hill, Heather C., Laura Kapitula, and Kristin Umland. 2011. "A Validity Argument Approach to Evaluating Teacher Value-Added Scores." *American Educational Research Journal* 48 (3): 794–831.

Hill, Heather C., Brian Rowan, and Deborah Loewenberg Ball. 2005. "Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement." *American Educational Research Association* 42 (2): 371–406.

Ho, Andrew D., and Thomas J. Kane. 2013. *The Reliability of Classroom Observations by School Personnel*. Seattle: Gates Foundation.

Kane, Thomas J., and Steven Cantrell. 2010. *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project*. Measures of Effective Teaching Project Research Paper 9, Gates Foundation, Seattle.

Kane, Thomas J., and Douglas O. Staiger. 2012. *Gathering Feedback for Teaching: Combining High-Quality Observations, Student Surveys, and Achievement Gains*. Research Paper, Measures of Effective Teaching Project, Gates Foundation, Seattle.

Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten. 2011. "Identifying Effective Classroom Practices Using Student Achievement Data." *Journal of Human Resources* 46 (3): 587–613.

Koedel, Cory, and Julian Betts. 2011. "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." *Education* 6 (1): 18–42.

Konstantopoulos, Spyros. 2014. "Teacher Effects, Value-Added Models, and Accountability." *Teachers College Record* 116 (1): 1–21.

Lloyd, Gwendolyn M., and Melvin Wilson. 1998. "Supporting Innovation: The Impact of a Teacher's Conceptions of Functions on His Implementation of a Reform Curriculum." *Journal for Research in Mathematics Education* 29 (3): 248–74.

Ma, Liping. 1999. *Knowing and Teaching Elementary Mathematics*. Mahwah, NJ: Erlbaum.

McCaffrey, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly. 2009. "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy* 4 (4): 572–606.

Milanowski, Anthony. 2004. "The Relationship between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati." *Peabody Journal of Education* 79 (4): 33–53.

Milanowski, Anthony. 2011. *Validity Research on Teacher Evaluation Systems Based on the Framework for Teaching*. Madison, WI: Consortium for Policy Research in Education.

National Governors Association Center for Best Practices, Council of Chief State School Officers. 2010. *Common Core State Standards for Mathematics*. Washington, DC: National Governors Association Center for Best Practices, Council of Chief State School Officers.

National Research Council. 2001. *Adding It Up: Helping Children Learn Mathematics*. Washington, DC: National Academy Press.

NCTM (National Council of Teachers of Mathematics). 1995. *Assessment Standards for School Mathematics*. Reston, VA: NCTM.

NCTM (National Council of Teachers of Mathematics). 2000. *Principles and Standards for School Mathematics*. Reston, VA: NCTM.

O'Connor, Mary Catherine. 1998. "Language Socialization in the Mathematics Classroom: Discourse Practices and Mathematical Thinking." In *Talking Mathematics in School: Studies of Teaching and Learning*, ed. Magdalene Lampert and Merrie L. Blunk. New York: Cambridge University Press.

Papay, John P. 2011. "Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates across Outcome Measures." *American Educational Research Journal* 48 (1): 163–93.

Pianta, Robert C., Bridget K. Hamre, and Susan Mintz. 2010. *Classroom Assessment Scoring System (CLASS) Manual: Upper Elementary*. Charlottesville, VA: Teachstone.

Podgursky, Michael J., and Matthew G. Springer. 2007. "Teacher Performance Pay: A Review." *Journal of Policy Analysis and Management* 26 (4): 909.

Polikoff, Morgan S. 2014. "Does the Test Matter? Evaluating Teachers When Tests Differ in Their Sensitivity to Instruction." In *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*, ed. Thomas J. Kane, Kerri A. Kerr, and Robert C. Pianta. San Francisco: Jossey-Bass.

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2nd ed. Thousand Oaks, CA: Sage.

Resnick, Lauren B., Robert Rothman, Jean B. Slattery, and Jennifer L. Vranek. 2004. "Benchmarking and Alignment of Standards and Testing." *Educational Assessment* 9 (1–2): 1–27.

Richland, Lindsey E. 2015. "Linking Gestures: Cross-Cultural Variation during Instructional Analogies." *Cognition and Instruction* 33 (4): 295–321.
Sawada, Daiyo, Michael D. Piburn, Eugene Judson, Jeff Turley, Kathleen Falconer, Russell Benford, and Irene Bloom. 2002. "Measuring Reform Practices in Science and Mathematics Classrooms: The Reformed Teaching Observation Protocol." *School Science and Mathematics* 102 (6): 245–53.
Schacter, John, and Yeow Meng Thum. 2004. "Paying for High- and Low-Quality Teaching." *Economics of Education Review* 23 (4): 411–30.
Schochet, Peter Z., and Hanley S. Chiang. 2013. "What Are Error Rates for Classifying Teacher and School Performance Using Value-Added Models?" *Journal of Educational and Behavioral Statistics* 38 (2): 142–71.
Schoenfeld, Alan H. 2004. "The Math Wars." *Educational Policy* 18 (1): 253–86.
Seidel, Tina, and Richard J. Shavelson. 2007. "Teaching Effectiveness Research in the Past Decade: The Role of Theory and Research Design in Disentangling Meta-analysis Results." *Review of Educational Research* 77 (4): 454–99.
Spearman, Charles. 1904. "The Proof and Measurement of Association between Two Things." *American Journal of Psychology* 15:72–101.
Star, Jon R. 2005. "Reconceptualizing Procedural Knowledge." *Journal for Research in Mathematics Education* 36 (5): 404–11.
Star, Jon R. 2007. "Foregrounding Procedural Knowledge." *Journal for Research in Mathematics Education* 38 (2): 132–35.
Star, Jon R., Courtney Pollack, Kelley Durkin, Bethany Rittle-Johnson, Kathleen Lynch, Kristie Newton, and Claire Gogolen. 2015. "Learning from Comparison in Algebra." *Contemporary Educational Psychology* 40:41–54. doi:10.1016/j.cedpsych.2014.05.005.
Star, Jon R., and Bethany Rittle-Johnson. 2009. "It Pays to Compare: An Experimental Study on Computational Estimation." *Journal of Experimental Child Psychology* 102 (4): 408–26.
Stein, Mary Kay, Juliet A. Baxter, and Gaea Leinhardt. 1990. "Subject-Matter Knowledge and Elementary Instruction: A Case from Functions and Graphing." *American Educational Research Journal* 27 (4): 639–63.
Stein, Mary Kay, Randi A. Engle, Margaret S. Smith, and Elizabeth K. Hughes. 2008. "Orchestrating Productive Mathematical Discussions: Five Practices for Helping Teachers Move Beyond Show and Tell." *Mathematical Thinking and Learning* 10 (4): 313–40.
Stein, Mary Kay, Barbara W. Grover, and Marjorie Henningsen. 1996. "Building Student Capacity for Mathematical Thinking and Reasoning: An Analysis of Mathematical Tasks Used in Reform Classrooms." *American Educational Research Journal* 33 (2): 455–88.
Stein, Mary Kay, and Suzanne Lane. 1996. "Instructional Tasks and the Development of Student Capacity to Think and Reason: An Analysis of the Relationship between Teaching and Learning in a Reform Mathematics Project." *Educational Research and Evaluation* 2:50–80.
Stevens, Floraline I. 1993. "Applying an Opportunity-to-Learn Conceptual Framework to the Investigation of the Effects of Teaching Practices via Secondary Analyses of Multiple-Case-Study Summary Data." *Journal of Negro Education* 62 (3): 232–48.
Stigler, James W., and James Hiebert. 1999. *The Teaching Gap: Best Ideas from the World's Teachers for Improving Education in the Classroom*. New York: Free Press.
TERC. 2008. *Investigations in Number, Data, and Space*. Salt Lake City: Pearson.
UCSMP (University of Chicago School Mathematics Project). 2007. *Everyday Mathematics*. New York: McGraw-Hill.

Webb, Norman L. 1999. *Alignment of Science and Mathematics Standards and Assessments in Four States*. Research Monograph No. 18, National Institute for Science Education, Madison, WI.

Yackel, Erna, and Paul Cobb. 1996. "Sociomathematical Norms, Argumentation, and Autonomy in Mathematics." *Journal for Research in Mathematics Education* 27 (4): 458–77.