

# Does Teaching Quality Cross Subjects? Exploring Consistency in Elementary Teacher Practice Across Subjects

Julie Cohen  
Erik Ruzek

University of Virginia

Lia Sandilos

Temple University

*Teacher evaluation systems treat instructional quality as generic. Principals often observe elementary teachers in one subject and generalize assessments to all subjects. However, there is little empirical work to justify these decisions. This study provides needed evidence about whether elementary teachers engage in comparable instruction across the school day. We draw on data from the Measures of Effective Teaching (MET) project, including student survey and classroom observational data from more than 500 elementary teachers. Findings indicate that there is moderate within-teacher, cross-subject consistency on the Tripod and Classroom Observation Scoring System (CLASS). Cross-subject correlations are higher on the Tripod scales ( $r$  values from 0.55 to 0.73) than the CLASS dimensions and domains ( $r$  values from 0.25 to 0.55). These findings suggest that teaching quality is not a uniform construct across subjects, even though current teaching evaluation systems largely treat it as such. Implications for elementary teacher preparation, professional development, and evaluation are discussed.*

**Keywords:** elementary schools, hierarchical linear modeling, observational research, school/teacher effectiveness, secondary data analysis, teacher assessment

## Introduction

A principal walks into a fourth-grade classroom and observes students engaged in rich, mathematically complex tasks. The teacher circulates around the room, providing detailed feedback to students. Management issues are seemingly invisible, and students are respectful and warm with each other and their teacher. The principal might assume this snapshot of instructional quality should be attributed to the teacher and hypothesize that a writing lesson in that same classroom run by the same teacher would feature similar characteristics. In fact, most current teacher evaluation systems rest on that assumption: that teaching quality is stable across content areas. Principals often observe elementary teachers in a single subject and generalize those assessments of instructional quality across subjects (J. Cohen & Goldhaber, 2016). In other words, evaluation systems treat measures of teaching as measures of teachers without recognizing the ways in which instructional quality may vary across subjects, even when the same teacher is working with the same students (Bell et al., 2012; Gitomer, 2009).

Decades of research on teaching suggests such assumptions might be misguided, that teaching activities and corresponding quality are situated in the particular content of instruction (D. K. Cohen, Raudenbush, & Ball, 2003; Stodolsky, 1988). Based on this, one would hypothesize that the teacher of the high-quality mathematics lesson described

previously might not demonstrate the same teaching quality in a writing lesson because of various factors, from content knowledge in the two subjects, to the curricular materials made available in different subjects, to the students' prior histories in mathematics and writing instruction (Graybeal & Stodolsky, 1986). Indeed, the cross-subject stability of measures of teaching quality has been a recent topic in the value-added literature. Asking whether "a good elementary teacher is always good" (Goldhaber, Cowan, & Walch, 2013), researchers have found moderate within-teacher, cross-subject correlations on value-added models (VAMs), ranging from 0.35 to 0.65 (Goldhaber et al., 2013; Loeb & Candelaria, 2012; Loeb, Kalgorides, & Bêteille, 2012; Teh, Resch, Walsh, Isenberg, & Hock, 2013). These findings indicate that some, but not all, individuals are comparably effective at influencing student achievement on standardized tests in different subjects. However, given that less than 30% of teachers are assessed with VAMs (Watson, Kraemer, & Thorn, 2009), districts designing evaluation systems would benefit from comparable analyses of other measures of teaching quality, including classroom observations and student surveys.

Observational measures are used to assess nearly every teacher in America (Goldring et al., 2015), and student surveys have been administered to more than a million students over the past 15 years (Ferguson, 2012). Both types of measures are used in consequential evaluation systems and



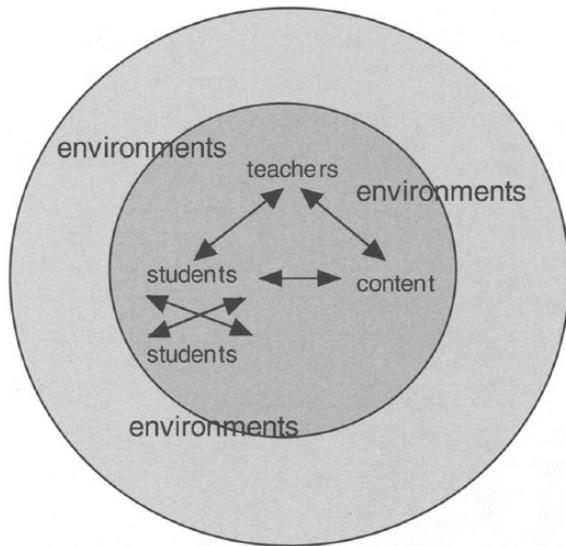


FIGURE 1. *Instruction as interaction* (from D. K. Cohen, Raudenbush, & Ball, 2003).

teacher professional development (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Downer, Stuhlman, Schweig, Martinez, & Ruzek, 2014). Given that nearly 90% of all elementary teachers instruct in multiple content areas, it would be helpful for districts and school leaders to know more about the stability of these measures across subject areas. Information about within-teacher, cross-subject variability in instructional quality could inform decisions about teaching assignments (e.g., subject specialists vs. generalists) and allow for more targeted professional development (e.g., extra support in mathematics instruction). Allocating resources based on teachers' relative strengths and weaknesses is predicated on evaluation systems that parse the measurement of instructional quality in different subjects. Current evaluation systems rarely provide this kind of subject-specific information about teaching (J. Cohen & Goldhaber, 2016).

This paper leverages data from the largest study of teaching ever conducted, the Gates Foundation's Measures of Effective Teaching (MET) project (Kane & Staiger, 2012), to analyze the degree to which the same elementary teachers demonstrate the same levels of instructional quality in mathematics and English language arts (ELA) based on the ratings of outside observers and their own students. In doing so, we raise a number of hypotheses about the factors that may contribute to within-teacher, cross-subject variation in such measures. Some of these hypotheses we can test with the MET data, and many we cannot. In surfacing these issues, we make a broader case for more research analyzing cross-subject instructional quality for elementary school teaching and studying cross-subject variability in instructional quality in elementary teacher preparation and professional development.

## Background Literature and Framework

D. K. Cohen et al. (2003) theorized that teaching was represented by "an instructional triangle" reflecting the interactions between the teacher, students, and content (see Figure 1). The theory suggests that shifting one corner of the triangle likely shifts the shape as a whole. A key contribution of the MET study was to randomly assign students to teachers to empirically test whether changing the "students" corner of the triangle changed teachers' value-added estimates, which, on average, it did not (Kane, McCaffrey, Miller, & Staiger, 2013). However, in elementary classrooms, even when teacher and students are the same, the content varies across the school day. Thus, this paper explores whether shifting the "content" corner of the instructional triangle is associated with observable changes in teaching quality. Though our methods do not allow for the causal claims afforded by the student random assignment experiment, our study is the first to use the MET data to look at the question of within-teacher, cross-subject variation in instructional quality across a large and diverse population of elementary school teachers.

In doing so, we draw on Bell and colleagues' (2012) notion that teaching quality, "the quality of interactions between students and teachers," is a characteristic of classrooms rather than individual teachers (p. 64). Those interactions are contingent on teacher and student knowledge, practices, and beliefs, which would likely vary by the content of instruction. Curricular materials are largely content specific, and teacher guides for textbooks are differentially elaborated in different subjects (Remillard, 2005; Reutzell, Child, Jones, & Clark, 2014). Principals may be more skilled at providing support for teaching language arts than mathematics (Neumerski, 2013). Districts might have more robust professional development resources for particular subjects (Guskey & Yoon, 2009; Van Driel & Berry, 2012). Research suggests elementary teachers may have differential beliefs (Kagan, 1992; Mewborn, 2001) and knowledge (Ball, 1990; Hill et al., 2008; Leinhardt, Putnam, Stein, & Baxter, 1991) about teaching mathematics and ELA. These all represent content-specific resources for teaching that could shift classroom interactions and corresponding measures of teaching quality.

Given these differences, we would expect within-teacher measures of instructional quality to vary across subjects. There is indeed some empirical evidence that elementary school teachers do not employ the same instructional techniques and activity structures across subjects and that quality correspondingly varies. Stodolsky (1988) used structured, qualitative observations of 20 teachers and analyzed activity structure, pacing, cognitive level, and student involvement in elementary social studies and mathematics instruction. Detailed analysis of these teachers' instruction led Stodolsky to conclude that the "subject matters" in instructional quality and that "what [teachers] are teaching shapes the way they

teach” (p. 74). Wood, Cobb, and Yackel (1990) also focused on instructional activities with a detailed, qualitative case study of a single teacher who provided direct, teacher-led instruction in ELA but allowed students to grapple with open-ended tasks and construct meaning independently during mathematics lessons.

Knapp, Shields, and Turnbull (1995) analyzed the extent to which 140 elementary teachers “taught for meaning” in mathematics, reading, and writing instruction. Using qualitative classroom observations, teacher interviews, analysis of curricular materials, and teacher logs of instructional activities, they found that “what teachers in our sample did in one subject area reveals relatively little about what they did in another” (p. 9). Few teachers seemed to have the resources—knowledge, beliefs, curricular materials, instructional support—to teach for meaning across subjects. Graeber, Newton, and Chambliss (2012) also analyzed cognitive demand for 69 elementary teachers. Drawing on data from classroom observations, teacher logs, and interviews, they also find notable differences in instructional quality based on the content of instruction. Only a small percentage of teachers demonstrated demanding instruction across subjects. These studies do not suggest that instructional quality *should* vary when teachers teach different subjects, but they provide empirical support that some features of teaching *do* vary, based on the instructional content.

What is less clear from the extant literature is the degree to which other instructional practices would similarly vary across subjects, particularly those that, on the surface, focus less directly on the teaching of academic content. For example, studies suggest that classroom management is the most temporally stable instructional practice across a range of different classroom observation measures, though no one to our knowledge has analyzed whether such practices are similarly stable across subjects (Gitomer et al., 2014; Polikoff, 2014). However, if students and teacher have different resources, knowledge, and beliefs about different subjects, then following D. K. Cohen et al.’s (2003) theory, we might also expect that *all* teaching practices would be influenced by such differences. Lower levels of self-efficacy for teaching mathematics than ELA, for example, might contribute to weaker behavior management and a less positive climate in mathematics classrooms (Pajares, 1996).

A new generation of classroom observation instruments allows for these cross-subject comparisons with larger samples of teachers and more facets of instructional quality. The MET study affords the possibility of looking at the cross-subject consistency of instructional quality using multiple measures, including the Classroom Observation Scoring System (CLASS; Pianta, Hamre, & Mintz, 2012) and the Tripod student survey (Ferguson, 2012). Both tools are designed to assess instructional quality and student engagement using classrooms, rather than teachers, as the unit of analysis. CLASS and Tripod have been used reliably across

thousands of classrooms in multiple subjects and grade levels (Ferguson, 2012; Pianta & Hamre, 2009).

All measures of instructional quality rest on a theory of teaching and learning. CLASS is based primarily on developmental theory, which suggests the interactions children have with adults and peers drive both learning and social development (Pianta et al., 2012). The tool’s developers suggest children are more likely to be engaged in learning when they have frequent, warm, and supportive exchanges with their classmates and teachers (Pianta & Hamre, 2009). Tripod also assesses classrooms as learning environments in which interactions and student-teacher relationships serve as a primary mechanism for students’ emotional and academic development (Ferguson, 2008; Ferguson & Danielson, 2014). The authors of both tools describe them as measures of the quality of classroom processes and student engagement and a proximal outcome of those processes<sup>1</sup> (Ferguson, 2010; Pianta & Hamre, 2009). Both measures are explicitly designed as content-neutral tools that assess aspects of classrooms that should exist regardless of the subject taught (Ferguson, 2012; Pianta & Hamre, 2009). That is, the instruments themselves are not designed to differentially privilege practices more common in a particular subject. Neither tool assesses features of classrooms that would or should occur more or less during the teaching of mathematics or language arts. Instead, they assess features of high-quality classroom spaces regardless of the content taught. Evidence suggests CLASS and Tripod scores are associated with student outcomes in both mathematics and language arts, indicating they are indeed valid measures of instructional quality across subjects (Allen et al., 2011; Bell et al., 2012; Ferguson, 2012; Ferguson & Danielson, 2014; Kane & Staiger, 2012).

There is also inevitable measurement error associated with any tool designed to assess teaching quality (Ho & Kane, 2013), and evidence suggests that raters are often the largest sources of error (Curby et al., 2011; Ho & Kane, 2013). Different instruments have developed different safeguards for mitigating rater error. CLASS relies on highly trained raters who undergo extensive certification and ongoing calibration exercises to promote consistent scoring, though most studies using CLASS still report moderate intraclass correlation coefficients (ICCs) among teams of scorers, ranging from .15 to .4 (Curby et al., 2011; Hamre et al., 2013; Mashburn, Downer, Rivers, Brackett, & Martinez, 2014). Tripod relies on the perspective of student raters, who are untrained in using the instrument in consistent ways. Unlike CLASS raters, who must demonstrate their skill at scoring according to established norms, Tripod assumes students are able to assess classroom quality as a function of their consistent and ongoing interactions with teachers (Wallace, Kelcey, & Ruzek, 2016). Perhaps because of a lack of training or because students have different classroom experiences, past studies have demonstrated substantial within-teacher variability in Tripod ratings (Raudenbush

& Jean, 2013; Schweig, 2014). However, Tripod surveys do aggregate information from the varied perspectives of multiple students, increasing the overall reliability of the tool (Kane & Staiger, 2012).

Despite these newer tools that allow for cross-subject comparisons, there is little empirical evidence about the stability in an individual teacher's practice across different subjects. Research has focused primarily on other sources of variability, especially temporal variations in instructional quality. Studies have demonstrated within-teacher variability in quality by time of year, occasion (lesson), and time within a lesson (Hill, Charalambous, & Kraft, 2012; Joe, McClellan, & Holtzman, 2014; Pianta & Hamre, 2009). Evidence using CLASS suggests moderate to strong within-year stability (Polikoff, 2014). To enhance the likelihood of comprehensive and stable estimates of practice, most observational measures require multiple observations, separated by several weeks or months (Hill et al., 2012; Ho & Kane, 2013; Pianta & Hamre, 2009). Cross-subject, within-teacher variability in scores is comparatively understudied. Given that so many elementary school teachers engage in both formative and summative assessment with such tools, we need more empirical evidence about the degree to which teaching quality is comparably stable across content areas. This is the need this study directly addresses.

Our purpose with this paper is not to make claims about whether "good instruction" is content-neutral or content-specific, nor is our goal to analyze the relative strengths of content-generic versus subject-specific measures of teaching. Others have written extensively on these topics (e.g., Hill & Grossman, 2013). Our aim is to provide empirical evidence about the degree to which elementary teachers demonstrate similar profiles of teaching quality depending on the content of instruction, using well-established measures that have been demonstrated to be valid and reliable across multiple subjects (Allen et al., 2011; Ferguson, 2012). The fact that we can reliably use these measures across subjects enables these cross-subject comparisons. We ask the following research questions:

*Research Question 1:* How much within-teacher variation is there on different measures of instructional quality?

*Research Question 2:* What is the relationship between teachers' instructional quality in mathematics and ELA?

## Method

### *Participants*

Data were collected during the first year of a two-year observational study of teaching, the MET Project, whose primary aim was to examine existing measures of instructional quality and teacher effectiveness (Kane & Staiger, 2012). The current study includes Tripod survey data from

572 fourth- ( $n = 288$ ) and fifth-grade ( $n = 284$ ) classrooms as well as observational data from a smaller subsample of teachers ( $n = 338$ ).<sup>2</sup> Classrooms were located in five large districts from Colorado, Florida, New York, North Carolina, and Tennessee. Elementary teachers included in the current study were regarded as *generalists* because they taught all major subjects to their students. The MET study also included 383 elementary *specialist* teachers, who only taught one subject area (ELA or mathematics) but were excluded from this study because of our interest in cross-subject consistency.

The following descriptive statistics are reported for the sample of 572 fourth- and fifth-grade generalist teachers. The majority of teachers were female (91%), and about half had a master's degree or higher (55%). Teachers reported an average of 6 years teaching in their current district ( $SD = 5.65$ , range, <1–34). The ethnic composition of teachers was 55% White, 39% Black, 5% Hispanic, and <1% other race. Across classrooms, an average of 14% of students were classified English language learner (ELL), 9% were eligible for special education, and 7% identified as gifted. Average ethnic composition in classrooms was 18% White, 51.2% Black, 23%, Hispanic, and 8% other race.

A total of 11,674 students completed ratings of their teachers' ELA or math instruction. On average, the sample of students was 49% male, 49% Black, 19% White, 24% Hispanic, 6% Asian, and 2% other race. An average of 8% of students qualified for special education, 7% of student qualified for gifted education, 14% were ELLs, and 44% received free or reduced price lunch (FRPL status was only provided for a subset of districts).

### *Measures*

*Classroom Assessment Scoring System, Upper Elementary.* The CLASS Upper Elementary (Pianta et al., 2012) is designed to assess teacher-child interactions in fourth through sixth grades. It is a structured observation system in which trained observers rate lessons on the emotional tone and climate in a classroom, management of behavior, presence of negativity, amount of time devoted to learning, and facilitation of a deeper understanding of content and analytical thinking. The dimensions consist of positive climate, teacher sensitivity, regard for student perspectives, behavior management, productivity, negative climate, instructional learning formats, content understanding, analysis and problem solving, quality of feedback, instructional dialogue, and student engagement (see online Supplemental Table S2). All dimensions are scored in 15-minute observation cycles using a 7-point scale ranging from *low* (1–2), to *middle* (3–5), to *high* (6–7). In a given cycle, a teacher will receive one score on each dimension.

The CLASS dimensions are then aggregated to form three domains per observation cycle. The emotional support

domain consists of the mean of positive climate, regard for student perspectives, and teacher sensitivity. The classroom management domain consists of the mean of negative climate, behavior management, and teacher sensitivity. The instructional support domain is an aggregate of the instructional learning formats, content understanding, analysis and problem solving, quality of feedback, and instructional dialogue dimensions.

Raters were randomly assigned to observation cycles across all participating teachers to mitigate potential effects of rater error. All raters also engaged in calibration activities designed to minimize rater drift each time they began scoring (Harik et al., 2009). Five percent of the videos in the sample were double-coded for interrater agreement (White & Rowan, 2013). The study's lead researchers deemed the interrater agreement within 1 point (i.e., adjacent agreement ranging from 68%–86% for CLASS MET data) to be acceptable for all the dimensions (Pianta et al., 2012).

*Tripod 7Cs Student Perceptions Survey.* The Tripod 7C's survey (Ferguson, 2008) examines students' perspectives on instructional quality and classroom processes using seven scales: clarify, challenge, captivate, confer, consolidate, control, and care. All 36 items are scored from 1 (*no, never*) to 5 (*yes, always*), and each scale is calculated by aggregating the item-level scores. The Tripod total score represents the mean of the seven subscale scores on a given survey. The Tripod overall score demonstrated high internal consistency across ELA and mathematics (36 items;  $\alpha_{\text{ELA}} = .96$ ,  $\alpha_{\text{mathematics}} = .95$ ).

In the current sample, scale-level Cronbach's alphas ranged from .61 to .91, also suggesting high internal consistency: Clarify (eight items;  $\alpha_{\text{ELA}} = .88$ ,  $\alpha_{\text{mathematics}} = .86$ ) measures the teachers' ability to help students gain a better understanding of the content being taught. Challenge (four items;  $\alpha_{\text{ELA}} = .73$ ,  $\alpha_{\text{mathematics}} = .72$ ) measures the rigor of instruction and effort required of students. Captivate (four items;  $\alpha_{\text{ELA}} = .81$ ,  $\alpha_{\text{mathematics}} = .78$ ) measures the teachers' skill at cultivating students' interest in academic content. Confer (seven items;  $\alpha_{\text{ELA}} = .85$ ,  $\alpha_{\text{mathematics}} = .83$ ) assesses teachers' skill at taking students' perspectives into account. Consolidate (two items;  $\alpha_{\text{ELA}} = .64$ ,  $\alpha_{\text{mathematics}} = .61$ ) measures teachers' skill at connecting different curriculum topics. Control (four items;  $\alpha_{\text{ELA}} = .76$ ,  $\alpha_{\text{mathematics}} = .78$ ) measures a teacher's ability to manage behavior in a classroom. Care (seven items;  $\alpha_{\text{ELA}} = .91$ ,  $\alpha_{\text{mathematics}} = .91$ ) inquires about a teacher's emotional supportiveness in the classroom (White & Rowan, 2013; see online Supplemental Table S3).

### Recruitment and Sampling

Recruitment was conducted through an opportunity sampling procedure; MET researchers targeted large urban districts that had previously worked with the Gates Foundation.

The sampling procedure resulted in participation from six large districts. The present study does not include one district, which did not have participating fourth- and fifth-grade classrooms. Within each participating district, school principals identified eligible teachers but did not require teachers to participate. To incentivize participation, each participating teacher received \$1,500. Participating schools also received a total of \$1,500 as well as \$500 per year to pay for a project coordinator. Unique educational settings, such as alternative schools and team-teaching situations in which it would be difficult to link students to one teacher, were excluded.

The present study analyzed data collected during the first year of the MET study (spring 2010). Students were administered the Tripod 7Cs survey between February and June 2010, depending on the schools' scheduling needs. It is worth noting that students rated their teacher's instruction in either mathematics or ELA at a single time point in the academic year. The scales measured students' perceptions of instruction during *either* ELA or mathematics instruction. Students were randomly assigned to recollect and rate instructional quality in one of the two subjects. The randomization of students to subject ratings was designed to reduce potential bias associated with students' preferences toward a specific content area and yield an independent set of ratings for each teacher in each subject. One group of students within each teacher's classroom rated ELA instruction, and a *different* group of students within the same classroom rated math instruction. On average, 20 students per classroom consented to participate in the study, which resulted in an average of 10 students per classroom completing ELA ratings and 10 students per classroom completing math ratings. An examination of demographic differences in the samples of student raters in mathematics and ELA indicated a slightly higher percentage of students who received special education in the mathematics sample compared to the ELA sample ( $\chi^2 = 6.18$ ,  $p < .05$ ). No other demographic differences were identified between student rater groups (see online Supplemental Table S1 for full demographics).

Classroom video recordings occurred over several months, between February and June 2010. Unlike the Tripod scores, which represent point-in-time perceptions of instructional quality, CLASS scores come from multiple lessons collected over a substantial portion of the school year. During the data collection window, generalist teachers recorded themselves teaching four ELA lessons and four mathematics lessons. For each lesson, only the first 30 minutes of the video were scored (15 minutes per cycle), totaling eight scored CLASS cycles for one subject. Thus, for the 11 dimensions on CLASS, teachers had approximately eight scores on each dimension in ELA and eight scores on each dimension in mathematics, which were analyzed in the current study to determine consistency in quality within and across subjects.

The MET data set is restricted because it contains sensitive and potentially identifying information about school districts. The findings reported in this paper follow the MET secure data reporting policies (Kane & Staiger, 2012).

### Data Analysis

The larger sample of participating generalists with Tripod data in the Year 1 elementary sample consisted of 593 teachers. However, 21 of those teachers had incomplete data (i.e., missing mathematics or ELA ratings). Logistic regression analyses were used to determine if classroom and teacher demographics predicted missingness ( $1 = \text{complete data}$ ,  $0 = \text{incomplete data}$ ). Findings indicated that teachers with incomplete data were more likely to have higher numbers of English language learners in their classrooms ( $B = -4.70$ ,  $p < .03$ ) and teachers with complete data were more likely to have students with higher state test scores in mathematics the year prior ( $B = 5.40$ ,  $p < .01$ ). No other classroom or teacher demographic variables were predictive of missing Tripod data. Given that the focus of analysis was cross-subject comparison and teachers with incomplete data made up less than 5% of the sample, these teachers were removed from all analyses, bringing the final sample to 572.

To assess within-teacher variability on the Tripod scales and CLASS dimensions in our sample, we first examine ICCs using unconditional models within and across subject areas. Next, given that each teacher received multiple ratings by both students (Tripod) and expert raters (CLASS) in each subject, we estimated a series of multilevel models to predict a teacher's mathematics and ELA ratings in each CLASS dimension and domain separately for math and ELA and similarly for each of the 7Cs and overall Tripod score separately for math and ELA. The model analyzed is shown in the following:

$$\text{Level 1: } y_{ij} = \beta_{0j} + e_{ij}$$

$$\begin{aligned} \text{Level 2: Intercept: } \beta_{0j} = & \gamma_{00} \\ & + \gamma_{01}(\text{classroom characteristics}) \\ & + \gamma_{02-05}(\text{district}_j) + U_{0j}. \end{aligned} \quad (1)$$

In this model, a single Tripod 7C or CLASS dimension score at Level 1 from student or observer  $i$  in the classroom of teacher  $j$  ( $y_{ij}$ ) was a function of a residual term. The Level 2 random intercept ( $\beta_{0j}$ ) for a teacher's mean level of the 7C or CLASS dimensions in mathematics or ELA was predicted by a set of teacher-level controls, including aggregate characteristics of students in the classroom—prior achievement levels in mathematics and ELA, racial composition, and gender composition—( $\gamma_{01}$ ), a set of four indicators for the district the teacher was in ( $\gamma_{02-05}$ ), and an error term ( $U_{0j}$ ) assumed to be normally distributed with a mean of 0 and an

estimated variance. Separate models were run for each of the CLASS and Tripod dimensions/subscales in each subject area.

To address Research Question 1, we ran unconditional models without predictors from which we estimated the ICC for each CLASS/Tripod dimension. Next, we added the Level 2 predictors indicated in Equation 1 and predicted values of the Level 2 random intercept for each teacher using empirical Bayes prediction. This employs information about the Level 1 and Level 2 error variances and the number of observations within a cluster to predict a teacher's Level 2 intercept value. In so doing, it corrects for unreliability in the prediction by shrinking predictions toward 0 when the Level 1 residual variance is large, the Level 2 random intercept variance is small, and/or a teacher has a small number of Level 1 observations (CLASS) or student ratings (Tripod).

To address Research Question 2, we correlated the empirical Bayes prediction for math and ELA for a given instrument and dimension (e.g., CLASS Productivity mathematics prediction with the ELA Productivity prediction, Tripod Control mathematics prediction with ELA Control prediction, etc.).

## Results

The findings are organized by research question, looking first at the results for the Tripod and then CLASS.

### Variation on Measures of Teaching Quality

*Tripod scales.* Means, ranges, and standard deviations for teachers' average scores on the Tripod scales are reported by subject area in Table 1. There were comparable distributions of scores in the two subjects on all seven scales. This suggests that *across* the MET sample of fourth- and fifth-grade generalist teachers, there was similar instructional quality in mathematics and ELA as measured by the Tripod.

Table 1 also shows the ICCs for each of the Tripod subscales in each subject. The ICCs indicate the average amount of variance between teachers in student ratings of a given Tripod subscale. In other words, these metrics demonstrate whether the teachers' students rate instructional quality consistently in *either* mathematics or ELA.<sup>3</sup> The ICCs suggest that in mathematics lessons, between 11% and 27% of variation in Tripod subscale scores exists between teachers being rated by students. In ELA observations, between 12% and 26% of subscale score variation exists between teachers being rated by students. Students are most consistent in their ratings of control, which describes how a teacher manages his or her classroom and maintains students' focus on the topic of instruction (Ferguson, 2012). The comparably high ICC of the Control scale is consistent with prior research using the Tripod; students seem to have more convergent perceptions of the level of control in their classrooms

TABLE 1  
Descriptive Statistics for Each Tripod Scale by Subject

Scale	Mathematics				English language arts				Across subjects
	Range	Mean	SD	Intraclass Coefficient Correlation	Range	Mean	SD	Intraclass Coefficient Correlation	Intraclass Coefficient Correlation
Care	2.59–4.95	4.16	.40	.19	2.68–4.99	4.14	.40	.18	.18
Control	2.16–5.00	3.49	.44	.27	2.10–4.92	3.47	.44	.26	.26
Challenge	3.00–5.00	4.18	.34	.11	2.75–5.00	4.17	.35	.14	.13
Confer	3.33–5.00	4.23	.28	.12	2.97–5.00	4.21	.31	.15	.14
Captivate	2.29–5.00	3.67	.40	.12	2.20–5.00	3.63	.43	.14	.13
Consolidate	2.25–5.00	3.81	.46	.11	1.90–5.00	3.81	.46	.12	.12
Clarify	2.73–4.98	4.24	.29	.15	3.09–4.95	4.21	.31	.18	.16

Note.  $N = 572$ .

TABLE 2  
Bivariate Correlations Among Tripod Scales in Mathematics (Above the Diagonal) and English Language Arts (Below the Diagonal)

	1	2	3	4	5	6	7
1. Care	—	.57**	.38**	.71**	.55**	.60**	.76**
2. Control	.58**	—	.40**	.52**	.40**	.40**	.60**
3. Challenge	.43**	.38**	—	.52**	.36**	.44**	.52**
4. Confer	.73**	.53**	.55**	—	.54**	.69**	.75**
5. Captivate	.57**	.42**	.41**	.59**	—	.55**	.56**
6. Consolidate	.57**	.39**	.48**	.69**	.54**	—	.62**
7. Clarify	.80**	.60**	.58**	.81**	.62**	.66**	—

Note.  $N = 572$ . These correlations reflect teacher-level averages on each of the 7Cs in mathematics and English language arts.  
\*\* $p < .01$ .

(Ferguson, 2012; Wallace et al., 2016). The care dimension had the next highest ICC, followed by clarify. The other Tripod scales all had ICCs between .11 and .15, which suggest a moderate degree of congruence among the student Tripod subscale ratings and align with other research using the tool.

Table 2 reports within-subject, across-teacher bivariate correlations among Tripod scales, again using teachers' average scores on each of the seven scales. Across the sample, average within-subject correlations fell within the moderate to high range ( $r_{\text{math}} = .36-.76$ ;  $r_{\text{ELA}} = .38-.81$ ). These scales are correlated with each other, suggesting the distinct Tripod scales capture related constructs within a particular subject. This is also consistent with prior research on the Tripod (Kuhfeld, 2017; Schweig, 2014; Wallace et al., 2016). As such, we also explore the within-teacher cross subject relationship at the overall Tripod score level in the following (see Table 5).

*CLASS dimensions.* Means, ranges, and standard deviations for the CLASS dimensions are reported in Table 3.

Here too, there are comparable distributions of scores in the two subjects across the MET sample of fourth- and fifth-grade generalist teachers. The descriptive statistics are also comparable to other studies using CLASS (e.g., Allen et al., 2013; Goble & Pianta, 2017).

The ICCs for the CLASS dimensions suggest that in mathematics observations, between 9% and 23% of the variation in dimension scores exists between teachers. In ELA observations, between 5% and 22% of the variation in dimension scores exists between teachers.<sup>4</sup> In both subjects, the highest ICC was for behavior management, which is the closest CLASS analog to the Control subscale of the Tripod. Across the two instruments, the largest amount of teacher-level variation is found in those measures that assess a teacher's efforts to keep student behavior appropriately focused on learning objectives. Positive and negative climate dimensions had the next largest ICCs, suggesting that a reasonable amount of variation in classroom emotional climate is attributable to a particular teacher in a given subject. These ICCs are comparable to other studies using CLASS, albeit on the lower end of the typical range (e.g., Mashburn et al., 2014)

TABLE 3  
*Descriptive Statistics for Each CLASS Dimension by Subject*

Dimension	Mathematics				English language arts				Across subjects
	Range	Mean	SD	Intraclass correlation coefficient	Range	Mean	SD	Intraclass correlation coefficient	Intraclass correlation coefficient
Positive climate	3.00–6.33	4.61	.65	.16	3.25–6.50	4.75	.61	.13	.15
Regard for student perspectives	1.50–6.00	3.20	.61	.12	2.25–5.50	3.52	.59	.09	.09
Teacher sensitivity	2.50–6.00	4.23	.59	.13	2.63–6.17	4.22	.55	.09	.11
Negative climate	1.00–3.38	1.34	.36	.16	1.00–2.67	1.29	.34	.16	.15
Behavior management	3.33–6.83	5.87	.58	.23	3.83–6.88	5.95	.54	.22	.21
Productivity	4.38–7.00	5.84	.45	.11	4.17–6.75	5.87	.43	.10	.10
Instructional learning formats	2.63–5.83	4.44	.53	.11	2.83–6.50	4.42	.48	.06	.09
Content understanding	2.50–5.75	4.08	.52	.09	2.50–5.50	4.12	.51	.06	.08
Quality of feedback	2.25–5.33	3.74	.61	.12	2.00–6.50	3.80	.61	.10	.11
Instructional dialogue	1.50–5.50	3.48	.61	.12	2.13–6.00	3.69	.62	.12	.11
Analysis and problem solving	1.50–5.00	2.76	.59	.11	1.00–4.57	2.88	.53	.05	.07
Student engagement	3.63–6.33	5.15	.50	.13	3.75–6.43	5.14	.47	.09	.10

Note.  $N = 338$  teachers who had math and English language arts ratings. These correlations reflect teacher-level averages on each of the CLASS dimensions in mathematics and English language arts.

TABLE 4  
*Bivariate Correlations Among CLASS Dimensions in Mathematics (Above the Diagonal) and English Language Arts (Below the Diagonal)*

	1	2	3	4	5	6	7	8	9	10	11	12
1. Positive climate	—	-.49**	.61**	.75**	.33**	.29**	.67**	.55**	.51**	.73**	.63**	.63**
2. Negative climate	-.45**	—	-.26**	-.48**	-.61**	-.45**	-.38**	-.35**	-.24**	-.41**	-.31**	-.40**
3. Regard for student perspectives	.63**	-.25**	—	.60**	.16**	.18**	.59**	.53**	.67**	.58**	.71**	.51**
4. Teacher sensitivity	.72**	-.39**	.63**	—	.43**	.43**	.63**	.58**	.51**	.74**	.64**	.62**
5. Behavior management	.38**	-.65**	.15**	.35**	—	.72**	.41**	.39**	.27**	.36**	.28**	.45**
6. Productivity	.32**	-.49**	.17**	.37**	.71**	—	.46**	.48**	.31**	.39**	.32**	.46**
7. Instructional learning formats	.62**	-.40**	.59**	.66**	.42**	.43**	—	.71**	.57**	.65**	.64**	.65**
8. Content understanding	.55**	-.28**	.56**	.65**	.37**	.41**	.70**	—	.63**	.69**	.66**	.53**
9. Analysis and problem solving	.47**	-.18**	.57**	.59**	.22**	.27**	.53**	.66**	—	.63**	.74**	.52**
10. Quality of feedback	.67**	-.28**	.62**	.74**	.28**	.32**	.63**	.73**	.68**	—	.78**	.57**
11. Instructional dialogue	.63**	-.27**	.68**	.69**	.24**	.32**	.64**	.73**	.70**	.84**	—	.60**
12. Student engagement	.62**	-.41**	.53**	.58**	.48**	.51**	.63**	.51**	.44**	.54**	.58**	—

Note.  $N = 338$ .  
 \*\* $p < .01$ .

Table 4 reports on the *between*-teacher bivariate correlations among average CLASS dimensions, which varied widely within subject area ( $r_{\text{mathematics}} = -.61$  to  $.78$ ;  $r_{\text{ELA}} = -.65$  to  $.84$ ), with correlations largely falling within the moderate to high ranges. This suggests that some of the CLASS

dimensions assess related aspects of the broader construct of instructional quality, which is consistent with a large body of literature using CLASS across thousands of classrooms (Hamre et al., 2013). Given these relatively high correlations, research studies using CLASS often report at the

TABLE 5  
*Empirical Bayes Estimates of Mathematics–English Language Arts Correlation for Tripod*

Tripod subscale	Empirical Bayes correlation
Care	0.60
Control	0.73
Challenge	0.59
Confer	0.58
Captivate	0.55
Consolidate	0.58
Clarify	0.60
Total Score	0.67

*Note.*  $N = 572$ . Covariates include: school district indicators, grade level of classroom, prior math and English language arts achievement of students, and percentage of students that are Black, Hispanic, receive free and reduced price lunch, and receive special education services.

domain level, which aggregates scores across conceptually and empirically related dimensions: emotional support, instructional support, and classroom management (Pianta & Hamre, 2009). As such, we also include domain-level scores in our findings that follow (see Table 6).

#### *Correlations Between Mathematics and ELA Instructional Quality by Instrument*

*Tripod scales.* Table 5 displays the correlations between teachers' empirical Bayes predictions of their mathematics and ELA student ratings for each of the seven Tripod subscales and the overall Tripod score. These predictions are based on the multilevel models in Equation 1 with a full set of Level 2 control variables for various classroom demographic and achievement characteristics. The correlations are all 0.55 or higher, suggesting that student ratings in mathematics and ELA tend to be moderately consistent across the two subjects a teacher teaches. The largest correlations between mathematics and ELA ratings were found for the Control subscale ( $r = 0.73$ ) and the Tripod total score ( $r = 0.67$ ), both of which have emerged as among the most salient aspects of the instrument in measurement research (Kuhfeld, 2017; Wallace et al., 2016).

*CLASS dimensions.* Table 6 displays the correlations between teachers' empirical Bayes predictions of their mathematics and ELA ratings for each of the CLASS dimensions and domains. The empirical Bayes predictions are based on multilevel models with the full set of classroom controls. Mathematics and ELA correlations tended to be lowest in the instructional support dimensions ( $r$  values between 0.25 for analysis and problem solving and 0.39 for instructional learning formats), next lowest in the emotional support dimensions, and highest in the classroom organization dimensions. That said, there were dimensions in the emotional

support and classroom organization domains with lower cross-subject correlations, including regard for student perspectives, which assesses the degree to which a teacher provides opportunities for autonomy and student leadership ( $r = 0.31$ ), and productivity, which assesses the degree to which instructional time is maximized ( $r = 0.36$ ). Along the same lines, at the domain level, mathematics and ELA correlations were lower in magnitude for instructional support ( $r = 0.42$ ) than for emotional support ( $r = 0.52$ ) and classroom organization ( $r = 0.55$ ).

## Discussion

Overall, the findings indicate that there is moderate within-teacher, cross-subject consistency on the Tripod and CLASS. Cross-subject correlations are higher on the Tripod scales ( $r$  values from 0.55 to 0.73) than the CLASS dimensions and domains ( $r$  values from 0.25 to 0.55). The range of cross-subject correlations on student surveys and observational measures are comparable to those highlighted in the teacher value-added literature (Goldhaber et al., 2013; Loeb & Candelaria, 2012; Loeb et al., 2012; Teh et al., 2013). These findings indicate that some teachers provide comparable instructional quality across subjects, but many do not.

These data also suggest that correlations between teachers' mathematics and ELA ratings depended in part on the aspect of teaching assessed, the rater scoring instructional quality, and scoring procedures used. Cross-subject correlations were, in general, lower when the raters were trained experts using the CLASS to score multiple lessons over a period of several months than when the raters were students who were asked to make point-in-time evaluations of instructional quality (Tripod). Only 2 of the 12 CLASS dimensions showed correlations between mathematics and ELA greater than 0.5, whereas *all* correlations between mathematics and ELA Tripod scores were greater than 0.5.

In particular, there were lower cross-subject correlations on the CLASS dimensions measuring features of instructional support, including analysis and problem solving, quality of feedback, content understanding, and instructional dialogue. Student engagement was also less consistent across the two subjects ( $r = 0.30$ ). Dimensions assessing the affective and organizational tenor of the classroom were more consistent across subjects, including measures of climate (both positive and negative) and behavior management. Little empirical research to date has examined whether these features of elementary classrooms look consistent across subjects, and our findings suggest more work in this area may be merited.

There were stronger cross-subject associations in Tripod ratings of the same teacher, though these too were in the moderate range. The higher correlations make conceptual sense given that in the data collection, students had the difficult task of rating teachers' practice entirely from recall

TABLE 6

*Empirical Bayes Estimates of Mathematics–English Language Arts Correlation for CLASS*

	CLASS Dimensions	Empirical Bayes Correlation
Emotional support	Positive climate	0.53
	Regard for student perspectives	0.31
	Teacher sensitivity	0.41
Classroom organization	Negative climate	0.45
	Behavior management	0.55
	Productivity	0.36
Instructional support	Instructional learning formats	0.39
	Content understanding	0.35
	Quality of feedback	0.38
	Instructional dialogue	0.36
	Analysis and problem solving	0.25
	Student engagement	0.30
CLASS domains	Emotional support	0.52
	Classroom organization	0.55
	Instructional support	0.42

*Note.*  $N = 338$ . Covariates include school district indicators, grade level of classroom, prior math and English language arts achievement of students, and percentage of students that are Black, Hispanic, receive free and reduced-price lunch, and receive special education services.

and mentally separating mathematics instruction from ELA instruction during this recall. The difficulty in making this separation, given students' exposure to instruction in both subjects, likely contributes to attenuation bias with respect to cross-subject differences in instructional quality. In contrast, CLASS raters had the benefit of rating instruction in real time and in only one subject while being effectively blinded to instructional practice in the other subject, which all but eliminates the potential for the raters' exposure to a teacher's mathematics instruction biasing their rating of ELA instruction or vice versa.

These data, particularly the CLASS data, raise questions about the common assumption that teachers who are skilled when teaching one subject are similarly skilled in other subjects. These findings also lend some empirical support to the theoretical ideas proposed by D. K. Cohen et al. (2003). Indeed, given these consistently moderate correlations, classrooms may be different places with distinct features depending on the content of the instruction. Content can indeed serve as context, even when actors are the same (Grossman & Stodolsky, 1994).

Why might we see moderate correlations between empirical Bayes predictions of instructional quality in mathematics and ELA? One hypothesis is that there is a stronger association between "true" instructional quality across subjects but we are not measuring quality well in either subject. All measures are imperfect and prone to measurement error, and observational measures are particularly susceptible to rater error (Park, Chen, & Holtzman, 2014). Indeed, there is considerable evidence in these data (see Tables 1 and 3) that

there is substantial within-teacher, within-subject variation across the CLASS dimensions and Tripod scales. The variability both within and across subjects lends support for a major conclusion drawn by the larger MET study: Instructional quality is best understood across multiple measures of instruction (e.g., observations, student ratings, VAMs) to mitigate the impact of measurement error (Kane & Staiger, 2012).

The MET study did attempt to address rater error for observational measures like CLASS in several ways: requiring all raters go through extensive training and certification procedures prior to scoring, mandating raters calibrate with "expert raters" before every scoring session, randomly assigning raters to lessons across teachers and subjects, and requiring a substantial percentage of lessons be double scored to ensure high levels of reliability (Kane & Staiger, 2012). In a practical context, observational measures may be used for consequential decision-making purposes with far fewer precautions in place to reduce systematic rater error (Whitehurst, Chingos, & Lindquist, 2015).

Tripod scores were collected from student raters who might be differentially biased toward the construct of interest, teaching quality, in mathematics or language arts. Students who dislike mathematics might be predisposed to rating instruction lower when recollecting a recent mathematics lesson, which would not reflect the actual quality of the instruction. That said, by randomly assigning a teacher's students to rate *either* mathematics or ELA and collecting student surveys from a large number of raters/students, the MET study attempted to minimize the likelihood of subject

preferences biasing students' responses. For subject preferences to bias the results, students who prefer a given subject would have had to be differentially assigned to one subject over the other across all teachers. However, the MET study does not include measures of students' attitudes or beliefs about particular subjects, making it impossible to test whether or not randomization worked as intended and the group of students rating mathematics instruction were in fact comparable to those rating ELA instruction. Across the two measures, the MET designers' efforts minimized but did not eliminate the influence of rater error on these findings.

Stable estimates of teaching quality are also predicated on appropriate sampling procedures. We need an adequate amount of information about instruction captured over sufficient time to make broader inferences about quality in a particular subject. In the MET study, only eight lessons were scored (four mathematics, four ELA), and all were captured during a four-month window, often during a timeframe that coincided with high-stakes testing. It is unclear if the instruction observed during this window would generalize across the school year. In addition, if an insufficient number of mathematics lessons or language arts lessons were captured or if different subject lessons were captured at systematically different times during that window, this could bias scores and misrepresent the actual relationship between instructional quality in the two subjects.

In addition, within the lessons captured, only the first 30 minutes of instruction were scored on CLASS. If, for example, mathematics and ELA lessons tend to develop differently, with more opportunities for "analysis and problem solving" or "instructional dialogue" in the beginning of ELA lessons but at the end of hour-long mathematics lessons, then a 30-minute sample would misrepresent the instructional quality in each subject as well as the relationship between the two subjects. Similarly, students rated instructional quality with Tripod surveys at a single time-point in the year. Given that temporal variation in instructional quality is well documented in prior studies of teaching quality, including those using CLASS (Pianta, Belsky, Vandergrift, Houts, & Morrison, 2008), these are real concerns. As a result, the MET study was designed based on a rigorous time sampling study, which indicated that the sampling parameters used for CLASS scoring would generate reasonably stable estimates of instructional quality (Kane & Staiger, 2012; McClellan, Donoghue, & Park, 2013), and the sampling is consistent with prior research using CLASS (Bell et al., 2012; Curby, Rimm-Kaufman, & Abry, 2013). Similarly, prior research examining student perception ratings have demonstrated high within-year stability (Ferguson, 2010). That said, the substantial within-teacher variability we observe in CLASS scores (see Table 3) suggests temporal issues likely played a role in score variability, along with differences in the subjective perspective of raters.

We know remarkably little about factors that might contribute to consistent practice for elementary school teachers, and these data do not allow us to understand *why* these elementary teachers demonstrate within- and across-subject differences on some CLASS dimensions and domains and Tripod scales. Individual characteristics such as years of experience (Kane, Rockoff, & Staiger, 2008; Papay & Kraft, 2015) or prior academic achievement (Wayne & Youngs, 2003) might be associated with higher quality in teaching in general. However, there is no evidence or theory to suggest that these returns to experience, for example, would cut across subjects. Elementary teachers may develop at different rates along different trajectories in different subjects contingent on the resources available to them.

Organizational theory and empirical studies suggest school-level factors from leadership to professional development opportunities can also promote or impede teaching quality in particular subjects (Bryk & Schneider, 2002; Ferguson & Hirsch, 2014; Johnson, Kraft, & Papay, 2012; Kanter, 2003; Kraft & Papay, 2014). However, here too, we lack evidence about how particular features of these resources foster development in certain subjects and/or promote consistent practice across subjects. "Resources for teaching" are often for teaching a particular subject (D. K. Cohen et al., 2003). Teacher and student knowledge and beliefs about instruction in a particular subject likely interact with contextual features of the school, shaping the development of instruction in a particular subject. It is important to understand if there are individual or institutional factors that contribute to consistency in teaching, because our goal is uniformly high-quality instruction across a school day. Therefore, continued research examining the relationship between school-level variables and consistency in high-quality practice is warranted.

### **Study Limitations**

In addition to the limitations of the measures themselves, the MET data do not allow for analyses of elementary instruction in subjects other than ELA and mathematics. It is not clear the degree to which these patterns would hold up in these teachers' science and/or history/social studies lessons. Second, this sample also only includes fourth- and fifth-grade teachers. It is also unclear whether upper-elementary teachers' practices represent patterns in cross-subject instruction across the elementary grades. Perhaps primary grade teachers, whose students take fewer standardized tests or are subject to fewer accountability pressures, are more likely to integrate curricula into cross-subject units and exhibit more consistent instructional profiles. Third, the districts, schools, and teachers in the MET study are a volunteer sample that may differ from teachers and students in districts nationwide in important ways. Understanding the full range of elementary teachers' practice in different grade levels,

subject areas, and school and district contexts will necessitate additional study.

All Tripod ratings and CLASS-scored videos were collected during a time in the school year that often coincided with the administration of the high-stakes student achievement tests. This may have shaped both the instructional practices used in CLASS-scored observations and student ratings of instructional quality. Teaching quality during this window, measured by students and outside raters alike, may not generalize across the school year. This is an important limitation of the MET database (J. Cohen, 2015; Grossman, Cohen, Ronfeldt, & Brown, 2014).

It is also worth noting that external raters, who do not know these teachers or their school contexts, scored the teachers on the observational measures used in these analyses. There is a growing body of research that suggests that in-school observers such as principals or coaches rate teaching practice differently than outside raters (Bell et al., 2016). Principals tend to rate teaching with more uniformly high scores and rely on organizational demands in scoring (Bell et al., 2016; Kraft & Gilmour, 2016). As a result, they may be less likely to rate teaching practices differently across subjects. Understanding more about the within-teacher, cross-subject consistency of measures of teaching quality in real school contexts is an important direction for future research on elementary teachers (J. Cohen, 2018).

Finally, these analyses do not allow for causal claims about the effects of content on teacher practice. Instead, this is a needed descriptive first step to better understand the degree to which a relatively large and diverse sample of elementary teachers exhibit consistent instructional quality across mathematics and ELA lessons, using well-established valid and reliable measures of teaching quality.

### **Implications and Directions for Future Research**

Policymakers have tended toward teacher evaluation policies that do not often rely on sampling across subjects for elementary teachers. These data suggest that might be misguided given cross-subject correlations on dimensions of CLASS ranging from 0.25 to 0.55. Domain-level CLASS scores are somewhat higher, but still below 0.55. While domain-level scores are more stable and less prone to measurement error, the cross-subject correlations at the dimension level may be more policy relevant because this is most often the level at which schools and districts collect information about practicing teachers. Given that our primary goal with these analyses is to explore whether schools, districts, and teachers would get comparable information about “teacher quality” across subjects using different kinds of measures of instructional quality, it was sensible to align our analytical approach to the ways in which these measures are typically used in formative and summative

observation-based evaluation systems around the country (J. Cohen & Goldhaber, 2016; Goldring et al., 2015).

Policymakers and districts alike need to ask whether assessments of teaching in one subject allow for inferences about the quality of teaching in others. If, for example, a teacher is observed and evaluated only or even primarily when teaching language arts, schools and districts may lose crucially important and distinct information about the instructional quality that teacher provides when teaching mathematics. Teachers may be rewarded or sanctioned for instructional quality that is specific to a particular subject rather than more generalizable to their teaching as a whole.

Purposive sampling of instruction in every subject may provide more accurate information about that teacher’s practice. That said, if districts simply average scores on observations conducted across subjects, they may still end up with a distorted portrait of teaching quality. For example, a teacher may score a low-level 2 on the CLASS dimension of instructional dialogue in mathematics lessons while scoring a 7 (the highest CLASS score) in ELA lessons. This teacher’s average would fall in the middle range of the CLASS rubric, a 4.5, which would not reflect the quality of his or her instructional dialogue in either subject. Elementary teachers would also lose important content-specific feedback on their instruction when observations are used in formative ways. Rather than treat elementary school teaching quality as a monolithic construct attributable to individual teachers, districts and schools might be well served to acknowledge and incorporate potential subject specificity and variability of teaching quality into both formative and summative evaluation systems as well as professional development efforts (Hiebert & Stigler, 2017).

Along the same lines, the moderate within-teacher, cross-subject correlations also have implications for research that utilizes student ratings and/or observational measures as outcomes of classroom-based interventions. These findings suggest that subject matter may need to be considered as a key variable in determining equivalence between treatment and control conditions in various randomized control trial studies. These data suggest we may come to different conclusions about the effects of particular interventions on instructional quality if we collect observational or student survey data in mathematics versus ELA.

This within-teacher, cross-subject variability in instructional quality, particularly when assessed over multiple lessons by trained outside raters, lends some empirical support for the current move toward departmentalized elementary classrooms, the structure traditionally used in middle and high schools (Goldhaber et al., 2013; Hess, 2009). If teaching quality is notably different by subject across a range of dimensions, it might be sensible to direct elementary teachers to focus on teaching the subject in which they demonstrate the highest quality instruction. For example, Baroody (2017) examined MET data and found that being an elementary

subject-specialist, a teacher who taught only one subject, had a small positive association with higher teaching effectiveness ratings in ELA classes but not mathematics.<sup>5</sup> That said, many elementary educators prefer a self-contained classroom where they work with the same students for the whole school day and have a richer sense of children's individual needs. There are good developmental reasons for teachers, parents, and students to prefer the self-contained model for young children. More research is needed to better understand the affordances and constraints of moving to a departmentalized model in elementary school.

We also need to understand more about how elementary teachers learn to engage in consistently high-quality teaching across the school day and how and why instructional quality may vary by subject. Theory suggests that less readily quantifiable features of schools, curricula, or teacher preparation might support high-quality instruction across the school day (D. K. Cohen et al., 2003). These are important directions for future research.

Elementary teachers are key resources for understanding the role of content in teaching. They teach multiple subjects but often have differential resources for teaching those subjects. It is not clear what extent divergent experiences with learning to teach different subjects contribute to some of the cross-subject variability in instructional quality we see in these data. Thus, by studying how those experiences and outcomes vary across the same teachers, we will better understand the role of content in the preparation and professional development of elementary school teachers. If some features of teacher education or in-service support are associated with consistently high-quality instruction across content areas, we will be better able to design pre- and in-service programs to capitalize on those commonalities. It could also be especially helpful for elementary teachers to analyze differences in how outside raters score their instructional quality in different subjects and in some cases, how students perceive them across subjects. These kinds of data could spur valuable reflection. Capitalizing on strengths in one subject could serve as motivation to improve on facets of instructional quality in another subject.

If we think it is sensible to emphasize cross-subject pedagogical linkages in elementary teacher preparation and professional development, then those who conduct research on teaching need to support those efforts by engaging in more comparative research analyzing elementary teaching across subjects. It is imperative to build a more robust research base about the factors that contribute to consistent high-quality practice. The subject-specific nature of extant research on teaching has limited opportunities for cross-subject comparisons like these, which are invaluable for understanding the variability of elementary teaching quality. We need to know more about what contributes to within-teacher, cross-subject variability in teaching so that we can improve the consistency of instruction in elementary classrooms.

## Notes

1. Engagement is measured as a proximal outcome of teaching quality in the CLASS dimension of student engagement and the Tripod scale of captivate. All other dimensions and scales are designed to measure classroom processes.

2. CLASS video data are only available for a subsample of generalist teachers because a portion of the teachers participating in the MET study did not provide consent for their video data to be analyzed by researchers (White & Rowan, 2013). The subsample of teachers with CLASS observation data was descriptively similar to the larger sample of generalist teachers. *t* tests and chi-square difference tests were used to determine if there were any significant differences in teacher background and classroom characteristics between teachers who had both Tripod and CLASS data and teachers who were missing CLASS observations. A significant difference was found ( $\chi^2 = 4.72, p = .03$ ) between the number of Hispanic teachers who had CLASS observation data ( $n = 10$ ) and those who did not ( $n = 17$ ). There were no other significant differences between the samples.

3. The last column of Table 1 includes the intraclass correlation coefficients (ICCs) from a model that is not restricted by subject. These ICCs are of a similar magnitude as the subject-specific ICCs.

4. The last column of Table 3 includes the ICCs from a model that is not restricted by subject. These ICCs are of a similar magnitude as the subject-specific ICCs.

5. English language arts and mathematics generalist classrooms in the MET sample served more students of color, served students with lower initial achievement, and had teachers with fewer years of teaching experience but more likely to have a master's degree. This suggests that there are likely meaningful differences between these two populations of teachers working with elementary-age students.

## References

- Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher-student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring system-secondary. *School Psychology Review, 42*(1), 76–98.
- Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*(6045), 1034–1037.
- Ball, D. L. (1990). The mathematical understandings that prospective teachers bring to teacher education. *The Elementary School Journal, 90*(4), 449–466.
- Baroody, A. (2017). Exploring the contribution of classroom formats on teaching effectiveness and achievement in upper elementary classrooms. *School Effectiveness and School Improvement, 28*(2), 314–335.
- Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2–3), 62–87.
- Bell, C., Jones, N., Qi, Y., Lewis, J., Witherspoon, M., Redash, A., & Kirui, D. (2016). *Administrators' roles in "valid" observation scores: Moving beyond a narrow measurement perspective. Paper presented at the annual meeting of the Association of Education Finance and Policy, Denver, CO.*

- Bryk, A. S., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. New York, NY: Russell Sage Foundation.
- Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25*(2), 119–142.
- Cohen, J. (2015). Challenges in identifying high-leverage practices. *Teachers College Record, 117*(8), 1–41.
- Cohen, J. (2018). Practices that cross disciplines?: Revisiting explicit instruction in elementary mathematics and language arts. *Teaching and Teacher Education, 69*(3), 324–335.
- Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*(6), 378–387.
- Curby, T. W., Rimm-Kaufman, S. E., & Abry, T. (2013). Do emotional support and classroom organization earlier in the year set the stage for higher quality instruction? *Journal of School Psychology, 51*(5), 557–569.
- Curby, T. W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Mooney, L., Downer, J., . . . Pianta, R. C. (2011). Within-day variability in the quality of classroom interactions during third and fifth grade. *The Elementary School Journal, 112*(1), 16–37.
- Downer, J. T., Stuhlman, M., Schweig, J., Martinez, J. F., & Ruzek, E. (2014). Measuring effective teacher-student interaction from a student perspective: A multi-level analysis. *Journal of Early Adolescence, 35*, 722–758.
- Ferguson, R. F. (2008). *The Tripod project framework*. Cambridge, MA: Harvard University.
- Ferguson, R. F. (2010). *Student perceptions of teaching effectiveness* (Discussion Brief). Cambridge, MA: National Center for Teacher Effectiveness and the Achievement Gap Initiative, Harvard University.
- Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan, 94*(3), 24–28.
- Ferguson, R. F., & Danielson, C. (2014). How framework for teaching and Tripod 7Cs distinguish key components of effective teaching. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 98–144). San Francisco, CA: Jossey-Bass.
- Ferguson, R. F., & Hirsch, E. (2014). How working conditions predict teaching quality and student outcomes. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 332–380). San Francisco, CA: Jossey-Bass.
- Gitomer, D. (Ed.). (2009). *Measurement issues and assessment for teaching quality*. Thousand Oaks, CA: Sage.
- Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record, 116*(6), 1–32.
- Goble, P., & Pianta, R. C. (2017). Teacher-Child interactions in free choice and teacher-directed activity settings: Prediction to school readiness. *Early Education and Development, 28*, 1035–1051.
- Goldhaber, D., Cowan, J., & Walch, J. (2013). Is a good elementary teacher always good? Assessing teacher performance estimates across subjects. *Economics of Education Review, 36*, 216–228.
- Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher, 44*(2), 96–104.
- Graeber, A. O., Newton, K. J., & Chambliss, M. J. (2012). Crossing the borders again: Challenges in comparing quality instruction in mathematics and reading. *Teachers College Record, 114*(4), 1–30.
- Graybeal, S. S., & Stodolsky, S. S. (1986). *Instructional practice in fifth-grade math and social studies: An analysis of teachers' guides*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (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, P., & Stodolsky, S. (1994). Content as context: The role of school subjects in secondary school teaching. *Educational Researcher, 24*(8), 5–23.
- Guskey, T. R., & Yoon, K. S. (2009). What works in professional development? *Phi Delta Kappan, 90*(7), 495–500.
- Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., & Brackett, M. A. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *The Elementary School Journal, 113*(4), 461–487.
- Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement, 46*(1), 43–58.
- Hess, F. (2009). How to get the teachers we want. *Education Next, 9*(3), 34–39.
- Hiebert, J., & Stigler, J. W. (2017). Teaching versus teachers as a lever for change: Comparing a Japanese and a US perspective on improving instruction. *Educational Researcher, 46*(4), 169–176.
- Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*(4), 430–511.
- Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56–64.
- Hill, H., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review, 83*(2), 371–384.
- Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Seattle, WA: Bill and Melinda Gates Foundation.
- Joe, J. M., McClellan, C. A., & Holtzman, S. L. (2014). Scoring design decisions: Reliability and the length and focus of classroom observations. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 415–444). San Francisco, CA: Jossey-Bass.

- Johnson, S. M., Kraft, M. A., & Papay, J. P. (2012). How context matters in high-need schools: The effects of teachers' working conditions on their professional satisfaction and their students' achievement. *Teachers College Record*, *114*, 1–39.
- Kagan, D. M. (1992). Implication of research on teacher belief. *Educational Psychologist*, *27*(1), 65–90.
- Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle, WA: Bill and Melinda Gates Foundation.
- Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, *27*(6), 615–631.
- Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teachers: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill and Melinda Gates Foundation.
- Kanter, R. M. (2003). *Challenge of organizational change: How companies experience it and leaders guide it*. New York, NY: Simon and Schuster.
- Knapp, M. S., Shields, P. M., & Turnbull, B. J. (1995). Academic challenge in high-poverty classrooms. *Phi Delta Kappan*, *76*(10), 770–776.
- Kraft, M. A., & Gilmour, A. F. (2016). *Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness*. Providence, RI: Brown University.
- Kraft, M. A., & Papay, J. P. (2014). Can professional environments in schools promote teacher development? Explaining heterogeneity in returns to teaching experience. *Educational Evaluation and Policy Analysis*, *36*(4), 476–500.
- Kuhfeld, M. (2017). When students grade their teachers: A validity analysis of the Tripod student survey. *Educational Assessment*, *22*(4), 253–274.
- Leinhardt, G., Putnam, R. T., Stein, M. K., & Baxter, J. (1991). Where subject knowledge matters. In J. Brophy (Ed.), *Advances in research on teaching* (pp. 87–113). Greenwich, CT: JAI Press.
- Loeb, S., & Candelaria, C. A. (2012). *How stable are value-added estimates across years, subjects and student groups?* Princeton, NJ: Carnegie Foundation for the Advancement of Teaching.
- Loeb, S., Kalogrides, D., & Bêteille, T. (2012). Effective schools: Teacher hiring, assignment, development, and retention. *Education Finance and Policy*, *7*(3), 269–304.
- McClellan, C., Donoghue, J., & Park, Y. S. (2013, April). *Commonality and uniqueness in teaching practice observation*. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.
- Mashburn, A. J., Downer, J. T., Rivers, S. E., Brackett, M. A., & Martinez, A. (2014). Improving the power of an efficacy study of a social and emotional learning program: Application of generalizability theory to the measurement of classroom-level outcomes. *Prevention Science*, *15*(2), 146–155.
- Mewborn, D. (2001). Teachers content knowledge, teacher education, and their effects on the preparation of elementary teachers in the United States. *Mathematics Teacher Education and Development*, *3*, 28–36.
- Neumerski, C. M. (2013). Rethinking instructional leadership, a review: What do we know about principal, teacher, and coach instructional leadership, and where should we go from here? *Educational Administration Quarterly*, *49*(2), 310–347.
- Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, *130*, 105–119.
- Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, *66*(4), 543–578.
- Park, Y. S., Chen, J., & Holtzman, S. (2014). Evaluating efforts to minimize rater bias in scoring classroom observations. In K. Kerr, R. Pianta, & T. Kane (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 383–414). San Francisco, CA: Jossey-Bass.
- Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal*, *45*(2), 365–397.
- Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109–119.
- Pianta, R. C., Hamre, B. K., & Mintz, S. (2012). *Upper elementary and secondary CLASS technical manual*. Retrieved from [cdn2.hubspot.net/hubfs/336169/](http://cdn2.hubspot.net/hubfs/336169/)
- Polikoff, M. S. (2014). The stability of observational and student survey measures of teaching effectiveness. *The American Journal of Education*, *121*, 183–212.
- Raudenbush, S. W., & Jean, M. (2014). To what extent do student perceptions of classroom quality predict teacher value added? In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project* (pp. 170–202). San Francisco, CA: Jossey-Bass.
- Remillard, J. T. (2005). Examining key concepts in research on teachers' use of mathematics curricula. *Review of Educational Research*, *75*(2), 211–246.
- Reutzel, D. R., Child, A., Jones, C. D., & Clark, S. K. (2014). Explicit instruction in core reading programs. *The Elementary School Journal*, *114*(3), 406–428.
- Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis*, *36*(3), 259–280.
- Stodolsky, S. (1988). *The subject matters: Classroom activity in math and social studies*. Chicago, IL: University of Chicago Press.
- Teh, B. R., Resch, A., Walsh, E., Isenberg, E., & Hock, H. (2013). *Is the stability of value-added underestimated?* Paper presented at the annual meeting of the Association of Education Finance and Policy, New Orleans, LA.
- Van Driel, J. H., & Berry, A. (2012). Teacher professional development focusing on pedagogical content knowledge. *Educational Researcher*, *41*(1), 26–28.
- Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod student perception survey. *American Educational Research Journal*, *53*(6), 1834–1866. doi:10.3102/0002831216671864

- Watson, J. G., Kraemer, S. B., & Thorn, C. A. (2009). *The other 69 percent*. Washington, DC: Center for Educator Compensation Reform at the U.S. Department of Education.
- Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*(1), 89–122.
- White, M., & Rowan, B. (2013). *User guide to Measures of Effective Teaching longitudinal database*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research, The University of Michigan.
- Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2015). Getting classroom observations right. *Education Digest, 80*(7), 20–28.
- Wood, T., Cobb, P., & Yackel, E. (1990). The contextual nature of teaching: Mathematics and reading instruction in one second-grade classroom. *The Elementary School Journal, 90*(5), 497–513.

### Authors

JULIE COHEN is an assistant professor of curriculum and instruction at the Curry School of Education at the University of Virginia. Her research focuses on the conceptualization and measurement of teaching quality as well as the development of effective instructional practices in pre-service teacher education and professional development.

ERIK RUZEK is a research assistant professor in the Center for the Advanced Study of Teaching and Learning at the Curry School of Education at the University of Virginia. He studies the impacts of classroom environments on student motivation, engagement, and academic achievement.

LIA SANDILOS is an assistant professor of psychological studies in education at Temple University. Her research focuses on teacher effectiveness and student-teacher interactions.