Educational Assessment, 13:267–300, 2008
Copyright © Taylor & Francis Group, LLC
ISSN: 1062-7197 print/1532-6977 online
DOI: 10.1080/10627190802602541

Routledge
Taylor & Francis Group

Toward Measuring Instructional Interactions "At-Scale"

Lindsay Clare Matsumura
University of Pittsburgh

Helen E. Garnier
UCLA/LessonLab Research Institute

Sharon Cadman Slater
LRDC/University of Pittsburgh

Melissa D. Boston
Duquesne University

This study explores two approaches to directly measuring the quality of instruction: teachers' assignments with student work and focused lesson observations. The technical quality and potential feasibility of these approaches for measuring instruction in large numbers of classrooms are compared within two different content areas (reading comprehension and mathematics). Generalizability and decision studies determined the optimal number of observations and assignments needed to obtain a reliable measure of a teacher's practice, and the association of these direct measures of instructional quality with student achievement was estimated. For both content areas, four assignments assessed by two raters yielded a reliable estimate of quality and as few as two observations yielded a reliable estimate of quality when teachers complied with the requirements of the research. The quality of observed instruction and teachers' assignments differentially predicted gains in students' achievement on the Stanford Achievement Test within each content area. The implications for measuring instruction "at-scale" in different content areas are discussed.

Correspondence should be addressed to Lindsay Clare Matsumura, University of Pittsburgh, 5808 Wesley Posvar Hall, Pittsburgh, PA 15260. E-mail: lclare@pitt.edu

School districts across the country are struggling to improve the quality of instruction. Districts invest large amounts of resources in professional development programs, curricula, and assessment systems, which are all intended to provide teachers with the resources they need to "teach well." The only information readily available to districts and schools regarding how well these efforts are working, however, is student outcomes on standardized achievement tests. The ways in which reform efforts influence (or fail to influence) what, and how, teachers teach remain unmeasured and unknown.

Why are measures of instruction not used on a more routine basis in schools and districts? Certainly politics play a role: teachers and teacher unions can be mistrustful of the motives of evaluators and concerned about subjectivity and bias, particularly when the evaluators are school or district personnel. Politics also play a role in how instructional quality is defined. Another reason, however, for the lack of attention to the specifics of instructional practice is that research is still at a beginning stage in terms of developing measures of good teaching within content areas that are feasible to use in large numbers of classrooms (Baker, 2007; Ball & Rowan, 2004).

Measuring instruction "at-scale" is important for several reasons. First, as described earlier, districts and schools need a way to systematically monitor the progress of reform initiatives on classroom practice. This information is critical to informing school leadership about the learning needs of their teachers and helping schools and district make more informed decisions about how to target professional development resources.

Measuring instruction also is important for directing attention to the quality of the learning environments teachers create for students. The quality of teaching is the most important factor—within the control of schools—influencing student learning (Darling-Hammond, 2000). What good teaching "looks like," however, is notably absent in the information reported to the public about school quality. Per the specifications of the *No Child Left Behind* act (2001), school quality generally is assessed by student performance on standardized achievement tests only. A growing body of research indicates, however, that this is inadequate for understanding the quality of students' learning opportunities. More alarmingly, some research indicates that excessive focus on achievement test scores can even be deleterious to student learning (Allensworth, Correa, & Ponisciak, 2008). This is because teachers who are under pressure to increase test scores often will narrow what they teach to the content represented in a test, and shape classroom activities and discussion to more closely mirror a test's question formats (Hamilton, 2003; Koretz & Barron, 1999; Koretz & Hamilton, 2006). Under these circumstances, students no longer have an opportunity to learn the span of knowledge and skills necessary to master grade-level content. Students may be rated "proficient" on their state's achievement test without a corresponding increase in the knowledge and skills necessary for success in the upper levels of schooling and beyond (Linn, 2003; Peterson & Hess, 2005; Resnick & Matsumura, 2007). Focusing attention on excellent instruction as an additional indicator of school quality could serve as a counter-balance to the pressure teachers feel to teach directly to the requirements of their state's achievement test. As such, it could mitigate some of the harmful effects of high-stakes student testing on instruction and help create more productive public reporting systems.

Measuring instruction could also improve communication about the characteristics of excellent instruction. Standards for instruction adopted by most states provide little guidance, mostly because the terms used to describe reformed practice are subject to multiple interpretations (Ball & Rowan, 2004). Teachers, principals, district leaders, and educational researchers can, and frequently do, disagree about what it means to "hold a discussion" or apply "critical thinking skills." Moreover, little opportunity exists for district leaders, principals, and teachers to develop consensus about what good teaching should look like. Measuring instructional quality on a routine basis in schools and districts could help create shared goals for teaching and a common vision of practice. As such, they could help create the conditions in schools and districts that are essential for instructional improvement (Spillane & Louis, 2002).

In this article, we describe two promising approaches to measuring the quality of teaching: teachers' assignments with student work and focused lesson observations. As noted by Ball and Rowan (2004), the measures used in education research have tended to assess instruction either broadly at a surface level (e.g., through teacher self-reports on surveys), or deeply on a small scale (e.g., through in-depth observation studies of a few classrooms). The measures described in this study were intended to bridge both perspectives—to provide independent, meaningful information about teaching quality in large numbers of classrooms. In the following section we describe some of the challenges associated with measuring instructional interactions at-scale and results from our research investigating the quality of our measures (their relation to student achievement) in different content areas. The potential feasibility of our measures for use in large numbers of classrooms is also explored.

The Challenge of Directly Assessing Instruction At-Scale

Teaching is a socially complex dynamic of interactions between teachers, students, and content (see Figure 1). Involved are teachers' knowledge of the subject-matter content (Hill, Ball, & Cohen, 2005), the affective climate of the classroom (Matsumura, Slater, & Crosson, in press), and the cultural match (or mismatch) between teachers and students (Delpit, 1988). *What* content is taught to students and *how* that content is delivered to students (i.e., the social

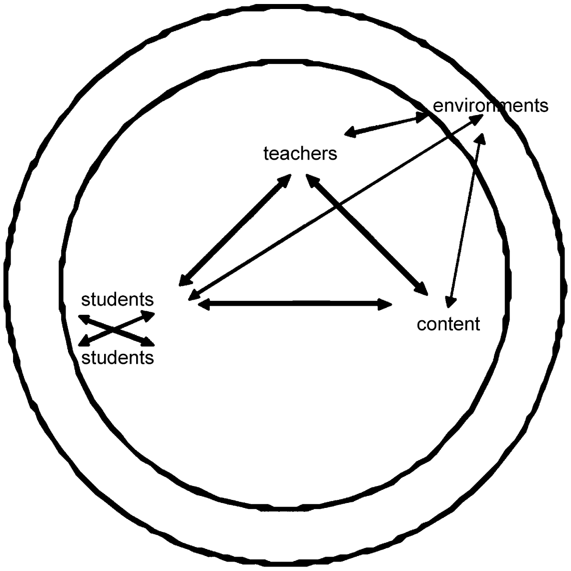

FIGURE 1 The instructional dynamic from Ball and Forzani, 2008.

interactions in the classroom around the content) is the core of instruction, however, and we argue, should be the focus of measurement for assessing instructional quality.

The challenge for measuring instruction is determining where in this complex dynamic to focus, and how much of any specific behavior to sample. The issue of how much to sample is important for considering the feasibility of measures intended for use in large-scale designs. While assessing instructional quality in a meaningful way is unlikely to be a cheap and easy endeavor, developing measures that are as minimally burdensome and inexpensive *as possible* is critical to investigating instruction in large numbers of classrooms.

Research conducted by the University of Michigan's Study for Instructional Improvement (SII) project indicates that instructional logs are an effective method for measuring the content of instruction in large numbers of classrooms (Rowan, Camburn, & Correnti, 2004). The results of this research indicate that teachers would reliably discriminate themselves—in terms of the content they teach—with as few as 20 log entries. Instructional logs, like surveys, are a reasonable approach for collecting information on instructional practice in large numbers of classrooms. Instructional logs are relatively inexpensive to administer, and pose a minimal burden on teachers as each log entry is short and takes little time to complete. Moreover, instructional logs represent a significant improvement over annual surveys because they circumvent the problem of inaccuracies in teachers' recollections that can arise when teachers are asked to retrospectively describe their teaching over a significant period of time. A limitation of instructional logs, however, is that they provide little (if any) insight on the interactions between teachers and students. For this reason, logs may be better suited for making broad distinctions between teachers in the *amount* of content they teach rather than for explaining differences in *how* content was taught.

To understand differences between teachers in how they teach and to provide a more complete picture of instructional quality other types of assessments are needed. Specifically, measures are needed that focus attention on the interactions between teachers and students—the other side of the instructional "triangle" pictured in Figure 1.

Lesson observations and the collection of teachers' assignments with student work are two approaches to directly measuring how content is enacted in classroom practice. Observations and assignments with student work capture information about how curricula are presented to students and the ways in which teachers—in the presenting of mathematical tasks or reading of a text—maintain or degrade the potential cognitive demand of the content. Different teachers can read the same rich text with their students (as specified in a district's scope and sequence plan, for example), but conduct discussions and engage students in assignment tasks that provide very different opportunities for students to deepen their comprehension and develop their academic skills (e.g., students' ability to use evidence from a text appropriately to support assertions in discussion or in their writing).

On the other hand, such performance-based assessments of teaching are expensive and vulnerable to many technical problems (see, for example, Dunbar, Koretz, & Hoover, 1991; Shavelson, Baxter, & Gao, 1993). Finding large numbers of qualified people who can consistently agree on the features of a "good" performance can be challenging. Variation in an individual's performance across tasks, activities, or occasions also can necessitate a large number of data points to yield a reliable estimate of performance and greatly increasing the expense and burden of data collection. Shavelson et al. (1993), for example, concluded that a reliable school-level estimate of science achievement required as many as 15 tasks from elementary school students. One can imagine that attempting to collect this number of tasks from teachers to make inferences about their science instruction, or observing teachers 20 times to make an inference about the content they emphasize in their practice (see Rowan et al., 2004), would not be possible. The expense to researchers or a district and the burden on teachers would be prohibitive.

In sum, more complete pictures of instructional quality rely on measures of instruction that yield information about what content is taught and how that content is enacted in practice. Instructional logs and other survey measures show promise for measuring the content of instruction at scale. The challenge is to develop assessments of instructional practice (enacted content) that provide meaningful information about instructional behaviors and do not entail large numbers of data collection points or excessive burden on teachers.

Goals of the Study

In the study described here, we explore and compare the quality of classroom observations and teacher assignment ratings for assessing instructional quality. These approaches are part of the Instructional Quality Assessment (IQA) designed to measure instruction across levels of schooling and content areas (Junker et al., 2006; Matsumura et al., 2006).

Our first goal for this study was to explore the potential feasibility of our measures for use in large numbers of classrooms by investigating the number of assignments and observations needed to yield a reliable estimate of a teacher's practice. In developing the IQA measures, we sought to minimize the problem of task-sampling variability by focusing on very specific activities and tasks (described in detail in the methods section). We expected that by narrowing the focus of measurement we would reduce the amount of data needed to obtain a reliable estimate of a teacher's practice.

The second goal of this study was to explore the relationship of the IQA ratings of classroom observations and teachers' assignments to each other and to student achievement. Since they measure different instructional behaviors, we investigated which type of measure is more strongly associated with student achievement, and whether to include both in the same design. The purpose of this goal was to contribute information and alternatives for districts and researchers by understanding the relationship of each type of measure with student achievement; in other words, to provide information about each measure so that districts and researchers could make an informed decision regarding the relative merit of each type of measure.

Finally, our research aimed to explore potential variation in the reliability and predictive validity of the IQA ratings in different content areas. A growing body of research indicates that instruction is not a generic practice, but is mediated by the subject-matter context (Grossman, Stodolsky, & Knapp, 2004; Stodolsky & Grossman, 1995; Stodolsky, 1988). Drawing on this research, we explored whether the subject matter context requires different approaches to instructional measurement.

The specific research questions addressed in this study were as follows:

1. What are the optimal numbers of observations and assignments needed to yield a reliable estimate of a teacher's practice within the different content areas of reading comprehension and mathematics?

2. What is the association between the IQA measures of classroom observations and teacher assignments? Does this relationship vary by the content areas of reading comprehension and mathematics?
3. Which approach to measuring instruction in the different content areas is a stronger predictor of student achievement: classroom observations, teacher assignments, or both?

METHODS

Participants

Grade 6 ($n = 22$) and Grade 7 teachers ($n = 12$) from five middle schools in an urban school district on the east coast participated in the study ($N = 34$, 71% female); 21 teachers taught English language arts and 13 taught mathematics. The teachers are mostly white ($n = 25$); the remaining teachers are Latino ($n = 5$) and African American ($n = 4$). From the pool of 34 teachers, 73% ($n = 25$) submitted assignments ($n = 16$ English language arts; $n = 9$ mathematics). Their students ($N = 492$, 51% female) were primarily from low-income families: 69% of the students qualified for free lunch and 9% were eligible for reduced-price lunch. Nearly half the students are Hispanic (48%). The remaining students are 20% African American, 22% white, 11% Asian, and 2% Native American. Almost none of the students were classified as Limited English Proficient (1%).

Student achievement was low overall, relative to student achievement in the state as a whole (New England Common Assessments Program Reporting, 2006). In our sample of English language arts classrooms, 62% of the students were categorized as Basic on total reading performance, 15% were categorized as Below Basic, 19% were categorized as Proficient, and 3% were categorized as Advanced. Of the students in the mathematics classes, 43.2% were categorized as Below Basic on total mathematics performance, 27% were categorized as Basic, 11% were categorized as Proficient, and 16% were categorized as Advanced.

Procedures

A member of the research team contacted the principals of each of the nine middle schools in the school district. Members of the research team visited each of the five schools that agreed to participate in order to discuss the study with interested teachers, to schedule the observations, and to distribute the assignment collection materials. Teachers ($N = 34$) were observed over a two-week time period (end of March to early April). Observations were conducted by members of the research team and a graduate student recruited and trained for the data collection ($N = 4$). Each teacher was observed on two consecutive days for the same class period by the same (single) rater. Teachers agreed in advance, as a condition for participation in the study, to hold a discussion about a text (for reading comprehension) or to engage students in a problem-solving activity and related discussion (in mathematics) on both days they were observed. Because of scheduling conflicts, four of the English language arts teachers were observed once (64 total observations).

Interrater agreement was assessed in non-sample classrooms prior to the study with each possible rater pair observing two consecutive lessons in each content area. The overall exact scale-point agreement between raters was 86% in total reading comprehension and 82% in total mathematics. Exact scale-point rater agreement for the individual rubrics ranged from moderate (around 70%) to excellent (100%). The exception to this was the rubric measuring the clarity of a teacher's expectations for student work. This rubric had poor interrater agreement in both reading comprehension (50%) and mathematics (43%; see Table 1).

The protocols for collecting and rating the quality of assignments follow a methodology established in past research conducted at the Center for Research on Evaluation, Standards, and Student Testing (Aschbacher, 1999; Clare, 2000; Matsumura, Garnier, Pascal, & Valdés, 2002). Of the 34 teachers who were observed, 25 teachers participated in the assignment collection ($n = 16$ English language arts; $n = 9$ mathematics). Teachers were asked to provide four assignments they considered to be challenging for their students (99 assignments[1]) in areas of instruction that are represented in most standards documents. Specifically, English language arts teachers were asked for response-to-literature assignments (e.g., an evaluation of a text, a character analysis, a comparison to multiple texts, etc.), and mathematics teachers were asked for assignments that engaged students in problem-solving activities. Teachers were asked to submit assignments they considered to be challenging for their students (rather than typical). The purpose of asking for challenging assignments from teachers was to establish a common basis for comparing teachers. We assumed that teachers would be more likely to provide challenging assignments and call them "typical" than to provide typical assignments and characterize them as challenging. Moreover, we were interested to know what teachers considered high-level, challenging work for students.

For each assignment, teachers completed a two-page cover sheet describing the context for the assignment, directions they provided to students, and their criteria for determining the quality of students' work. Teachers also included four representative samples of student work (two they considered to be of medium quality and two of high quality), the directions they gave to students, and the

[1] One teacher submitted only three assignments.

TABLE 1
Quality of Observed Instruction in Reading Comprehension
(n = 21 Teachers, 38 Observations)

Observation Rating	*Mean*	*SD*	*Range*	*0*	*1*	*2*	*3*	*4*
Classroom Talk								
Student participation in the discussion	2.00	1.47	0–4	26.3	7.9	23.7	23.7	18.4
Teacher links student contributions to each other	2.0	1.47	0–4	28.9	18.4	28.9	13.2	10.5
Students link to each other's contributions	1.11	1.16	0–4	28.9	52.6	7.9	0	10.5
Teacher presses for accurate knowledge and for students to explain thinking	1.55	1.29	0–4	28.9	18.4	28.9	15.8	7.9
Students provide accurate knowledge and explain their thinking	1.42	1.24	0–4	28.9	26.3	26.3	10.5	7.9
Analyzing and interpreting a text through discussion	1.41	1.26	0–4	37.8	5.4	40.5	10.8	5.4
Teacher Expectations								
Clarity and detail of the expectations for student learning	1.88	1.16	0–4	6.3	40.6	25.0	15.6	12.5
Rigor of the expectations for student learning	1.35	1.11	0–4	19.4	45.2	25.8	0	9.7
Student access to expectations	2.03	1.64	0–4	18.8	34.4	9.4	0	37.5
Cognitive Demand of Task								
Rigor of the text	2.28	0.92	0–3	3.4	20.7	20.7	55.2	
Analyzing and interpreting a text through lesson activities	1.45	0.91	0–3	15.2	36.4	36.4	12.1	0

rubrics they used to assess students' work. Teachers received the assignment materials in early January. After these materials were collected in early April, teachers received $100 gift certificates for participating in the assignment collection activity.

Three raters independently rated each of the assignments. Exact scale-point agreement, averaged across pairs of raters, was moderate: 71% in reading comprehension and 76% in mathematics (see Table 2). Exact scale-point agreement for the individual rubrics ranged from 61% to 93% in reading comprehension and from 63% to 85% in mathematics.

TABLE 2
Quality of Response to Literature Assignments (n = 16 Teachers, 64 Assignments)

Assignment Rating	*Mean*	*SD*	*Range*	*0*	*1*	*2*	*3*	*4*
Teacher Expectations (assessment criteria)								
Clarity and detail of the expectations for student learning	2.29	1.02	0–4	3.2	15.9	46.0	19.0	15.9
Rigor of the expectations for student learning	1.90	1.29	0–4	17.5	20.6	30.2	17.5	14.3
Student access to expectations	2.42	1.22	0–4	3.8	26.4	17.0	30.2	22.6
Cognitive Demand of Task								
Rigor of the text	2.53	.85	0–3	2.8	13.9	11.1	72.2	
Analyzing and interpreting the text	1.80	.91	0–4	4.7	37.5	32.8	23.4	1.6

Measures

Two lines of research supported development of the IQA rubrics. The first line of research focused on general features of "good" instruction, notably the reports published by the National Research Council (Bransford, Brown, & Cocking, 1999, 2000[2]). The second line of research focused on excellent practice within subject areas (e.g., Beck, McKeown, Hamilton, & Kucan, 1997; Doyle, 1988; Goldenberg, 1992/1993; Cobb, Boufi, McClain, & Whitenack, 1997; O'Connor & Michaels, 1996; Snow, 2002; Stein, Smith, Henningsen, & Silver, 2000). Based on this review, three broad and overlapping constructs were identified that characterize the quality of instruction common across subject-areas: level of cognitive demand of tasks and activities, classroom talk, and expectations communicated to students for the quality of their work.

The cognitive demand of tasks and the class discussion differ within each content area. In reading comprehension, the quality of the text (its potential for supporting high-level engagement with a text; Beck et al., 1997), the intellectual demand of the task/discussion (Newmann, Lopez, & Bryk, 1998; Newmann, Bryk, & Nagaoka, 2001; Snow, 2002), and the guidance students receive to provide/write extended responses and use appropriate evidence from a text to support their position is considered (Newmann et al., 1998). The rubrics in mathematics are based on the Mathematical Task Framework developed by Mary Kay Stein, Margaret Smith and their colleagues (Stein, Smith, Henningsen, &

[2] This research has been summarized for practitioners as the Principles of Learning (Resnick & Hall, 2001).

Silver, 2000). This framework considers the potential of a task to support higher-level, conceptual thinking including students' opportunity to engage with a range of representations in their responses, and the implementation of the task in practice (as enacted).

The rubrics for the IQA ratings of the quality of classroom talk and teachers' expectations are similar across content areas. The classroom talk rubrics focus on the percent of students participating in a discussion, the degree to which a teacher presses students to explain their thinking and engage with ideas and concepts, and a teacher's use of specific "talk moves" that help make reasoning public and accessible to all students (Goldenberg, 1992/1993; O'Connor & Michaels, 1996). The ratings of teachers' expectations focus on the amount and quality of the information teachers provide to students with regard to what "good" student work should look like, and how these expectations are communicated to students (Black & Wiliam, 1998; Doyle, 1983).

It is notable that both observations and assignments assess the interactions between teachers and students, but in somewhat different ways. Both types of measures focus on how tasks are presented to students (in class discussions or in the directions given to students), understood by students (as evidenced in the discussion or in their written work), and the criteria that a teacher accepts for completed (acceptable) work (either expressed to students in a discussion or in the grading criteria a teacher uses to assess students' work; Doyle, 1983; 1988). Observations focus on students' opportunity to participate in rich classroom discussions, however, while assignments focus on students' opportunity to develop their written communication skills.

Tables 3 to 6 list the individual dimensions that comprise the IQA, and the range of teacher performance as assessed on these dimensions for this study.

Student achievement was measured on the Stanford Test of Achievement 10th edition (SAT-10). The following subscores were used to assess student achievement in English language arts: Total Reading, Reading Comprehension, and Vocabulary. The following subscores were used to assess student achievement in mathematics: Total Mathematics, Procedures and Problem Solving (see Table 7).

Analyses

Generalizability studies were conducted to determine if collecting four assignments and observing teachers twice yielded a reliable estimate of an individual teacher's practice. Decision studies were conducted to investigate alternatives for future research designs, such as projects with different budget constraints. Data were analyzed using mGenova (Brennan, 2001a, 2001b). Correlation and multiple regression analyses were conducted to explore the relation of the

TABLE 3
Quality of Observed Instruction in Mathematics (n = 13 Teachers, 26 Observations)

Observation Rating	*Mean*	*SD*	*Range*	*0*	*1*	*2*	*3*	*4*
Classroom Talk								
Student participation in the discussion	2.57	1.04	0–4	8.7	0	30.4	47.8	13.0
Teacher links student contributions to each other	1.54	1.10	0–4	19.2	26.9	42.3	3.8	7.7
Students link to each other's contributions	1.08	.85	0–4	19.2	61.5	15.4	0	3.8
Teacher presses for accurate knowledge and for students to explain their thinking	1.81	1.27	0–4	19.2	23.1	23.1	26.9	7.7
Students provides accurate knowledge and explain their thinking	1.73	1.25	0–4	19.2	26.9	23.1	23.1	7.7
Rigor of discussion following the task	1.65	1.38	0–4	23.1	30.8	19.2	11.5	15.4
Teacher Expectations								
Clarity and detail of the expectations for student learning	1.81	1.27	0–4	11.5	42.3	11.5	23.1	11.5
Rigor of the expectations for student learning	2.08	1.16	0–4	7.7	23.1	38.5	15.4	15.4
Student access to expectations	2.65	1.57	0–4	11.5	23.1	3.8	11.5	50.0
Cognitive Demand of Task								
Potential of the task	2.46	.91	0–4	3.8	0	57.7	23.1	15.4
Implementation of the task	2.28	.74	0–4	4.0	0	64.0	28.0	4.0

observed instruction and assignment quality ratings with each other and with student achievement.

RESULTS

Number of Observations and Assignments Needed to Obtain a Reliable Estimate of Instructional Quality

Observations of reading comprehension and mathematics lessons. The ratings of each dimension of instructional quality assessed by the observation protocols were averaged to create an overall score of instructional quality in each content area. This summary score was used in the generalizability and decision analyses reported here. Results indicated that as few as two observations yielded a reliable estimate of quality *when teachers complied*

TABLE 4
Quality of Classroom Assignments in Mathematics (n = 9 Teachers, 35 Assignments[a])

Assignment Rating	*Mean*	*SD*	*Range*	*0*	*1*	*2*	*3*	*4*
Teacher Expectations (assessment criteria)								
Clarity and detail of the expectations for student learning	2.69	.86	1–4	0	6.3	37.5	37.5	18.8
Rigor of the expectations for student learning	3.03	.78	1–4	0	5.9	11.8	55.9	26.5
Student access to expectations	2.61	.84	1–4	0	6.5	41.9	35.5	16.1
Cognitive Demand of Task								
Potential of the task	3.06	.78	1–4	0	2.9	17.6	50.0	29.4
Implementation of the task	2.79	.77	1–4	0	2.9	32.4	47.1	17.6
Rigor of student work following task	3.03	.86	1–4	0	3.1	25.0	37.5	34.4

[a] One of the nine teachers submitted 3 assignments.

with the requirements of the data collection ($\hat{\phi}$ = .80 and .86 for reading comprehension and mathematics, respectively[3]; see Table 8). As described earlier, teachers agreed in advance to hold class discussions on each of the two days we visited. Four teachers (two in each content area) did not comply with the data collection requirements on one of these two days. One reading comprehension teacher engaged students in a writer's workshop the entire class period and another teacher had students work independently throughout the class period (e.g., listening to books on tape and following along with the text, etc.). Similarly, one math teacher tested students while another required student presentations for the entire class. These four teachers received zero scores (meaning target behavior not observed) on most IQA rubrics. When including their scores in analyses, the number of observations needed to obtain a reliable estimate of quality increased considerably. The variance component associated with Observation was zero or slightly negative in this analysis. This was likely because the observations for any given teacher were conducted by the same rater and occurred on adjacent days. Therefore, there was little variation in observation from one day to the next.

[3] For the analysis of observation data, a random-effects Teacher × Observation design was used. Raters never overlapped in their observations so rater was not included as a facet in this design. For the analysis of assignment data, a random-effects Teacher × Rater × Assignment analysis of variance was performed, and variance components were estimated. Variance component estimates and dependability coefficients are presented for each analysis. Dependability coefficients for absolute decisions, reported throughout the paper (i.e., phi-coefficients), describe absolute level of performance rather than generalizability coefficients for relative decisions intended for rank ordering and norm-referenced comparisons. To help interpret the variance component estimates, percent of total variability accounted for by each variance component is presented.

TABLE 5
Level of Agreement between Raters for the Classroom Observation Ratings (Reading Comprehension: 3 Observers, 4 Observations; Mathematics: 4 Raters, 7 Observations)

	Reading Comprehension		*Mathematics*	
Observation Rating	*% Agreement*	*ICC*	*% Agreement*	*ICC*
Overall	86.4	0.96	81.8	0.98
Classroom Talk				
Students participate in the discussion	75.0	0.80	85.7	0.99
Teacher links student contributions to each other	75.0	0.84	100.0	1.0
Students link to each other's contributions	100.0	1.0	100.0	1.0
Teacher presses for accurate knowledge and for students to explain their thinking	100.0	1.0	85.7	0.99
Students provide accurate knowledge and explain their thinking	75.0	0.57	71.4	0.99
Analyzing and interpret a text in the discussion	100.0	1.0		
Teacher Expectations				
Clarity and detail of the expectations for student learning	50.0	0.89	42.9	0.11
Rigor of the expectations for student learning	75.0	0.80	71.4	0.98
Student access to expectations	100.0	1.0	100.0	1.0
Cognitive Demand of Task				
Rigor of the text	100.0	1.0		
Analyzing and interpreting a text in lesson activities	100.0	1.0		
Potential of the task			71.4	0.75
Implementation of the task			100.0	1.0
Rigor of discussion following the task			71.4	0.99

Reading comprehension and mathematics assignments. As with the observations, the individual ratings of each dimension of assignment quality were averaged to create an overall score for instructional quality; this score was used in the analyses reported here. Results indicated that collecting four assignments from teachers yielded a generalizable estimate of quality in both content areas. Specifically, results for reading comprehension indicated a dependability coefficient of .82 for four assignments per teacher rated by three raters. Decision studies estimated that reducing the number of raters to two did not substantially change the dependability coefficient ($\hat{\phi} = .81$). Collecting three reading assignments per teacher was likely to be sufficient, with dependability estimated just below .80 ($\hat{\phi} = .78$; see Table 9). Results for mathematics assignments were similar, yielding a dependability coefficient of .80 for four

TABLE 6
Level of Agreement between Raters for the Assignment Ratings (Reading Comprehension: 3 Raters, 64 Assignments; Mathematics: 3 Raters, 35 Assignments)

	Reading Comprehension		*Mathematics*	
Assignment Rating	*% Agreement*	*ICC*	*% Agreement*	*ICC*
Overall	71.3	0.93	76.3	0.88
Teacher Expectations				
Clarity and detail of the expectations for student learning	64.1	0.79	82.9	0.77
Rigor of the expectations for student learning	60.9	0.80	75.2	0.63
Student access to expectations	74.5	0.81	84.8	0.90
Cognitive Demand of Task				
Rigor of the text	93.0	0.93		
Analyzing and interpreting the text	72.9	0.78		
Potential of the task			72.9	0.51
Implementation of the task			79.0	0.72
Rigor of students' work following task			62.9	0.67

TABLE 7
Descriptive Statistics of Student Achievement Scores in Reading and Mathematics

	Prior Achievment			*End-of-Year Achievement*		
Sub Score	*Mean*	*SD*	*Range*	*Mean*	*SD*	*Range*
Reading						
Total reading	645.58	32.78	576–757	649.58	31.04	571–799
Reading comprehension	643.93	34.21	567–753	649.58	32.44	568–805
Vocabulary	650.31	38.80	543–790	650.78	37.66	560–775
Mathematics						
Total mathematics	652.08	48.37	560–803	657.55	40.96	588–806
Problem solving	650.13	47.54	553–805	657.46	47.54	583–814
Procedures	655.89	57.71	548–797	658.93	47.96	550–819

assignments per teacher rated by three raters. Decision studies estimated that reducing the number of raters to two only minimally reduced the dependability coefficient ($\hat{\phi} = .77$), but collecting only three mathematics assignments per teacher may not be sufficient ($\hat{\phi} = .74$; see Table 9). Notice in Table 9 that variance components associated with rater (r, tr, and ra) are always negligible

TABLE 8
Decision Study for Observation Data [t × o Design]:
Two Noncompliant Teachers Removed from Dataset
(Reading Comprehension: n = 15 Teachers; Mathematics: $n_t = 11$)

	Reading Comprehension				*Mathematics*			
Source of Variation	*G Study*	*Alternative D Studies*			*G Study*	*Alternative D Studies*		
$n_o =$	2	3	4	5	2	3	4	5
Teacher (t)	41.2 (67%)	41.2 (86%)	41.2 (89%)	41.2 (91%)	63.2 (75%)	63.2 (90%)	63.2 (92%)	63.2 (94%)
Observation (o)[a]	0	0	0	0	0	0	0	0
Residual (to, e)	19.9 (33%)	6.66 (14%)	5.0 (11%)	4.0 (9%)	21.4 (25%)	7.2 (10%)	5.4 (8%)	4.3 (6%)
Absolute error *SD*	3.16	2.58	2.24	2.00	3.27	2.67	2.31	2.07
Dependability Coefficient	.80	.86	.89	.91	.86	.90	.92	.94

[a]Negative variance components set to zero.

or zero. This is a common finding in generalizability studies that is explained by the general rule that rater-sampling variability is less of an issue than task-sampling variability. Raters can be trained to consistently judge performance, and they were in this study (see Table 6). Our results were consistent with Shavelson, Baxter, and Gao (1993) who presented results from a number of G-studies in math, science, and the military. They found that person × task was consistently the major source of measurement error and that the variance components for rater, person × rater, and task × rater were always negligible or zero. Schoonen (2005) also reported that person and task contribute more to score variance than do raters.

Relationship of the IQA Ratings of Observed Instruction and Assignment Quality

The correlation between overall ratings of observed instruction and assignment quality was only moderate ($r = .20$, $p < .01$) but differed within the content areas. In reading comprehension, the ratings were uncorrelated ($r = .03$); in mathematics, the ratings were highly correlated ($r = .68$, $p < .01$).

To further investigate the association between the two reading comprehension ratings, additional correlations were computed between observed instruction and assignment quality within the three teaching domains used to create the overall ratings: classroom talk, teacher expectations, and cognitive demand. The quality of the expectations communicated to students in observed lessons was significantly but weakly correlated with assignment quality ($r = .17$ and $r = .05$, respectively). No correlation was found between the assignment ratings and the

TABLE 9
D-Study for Assignment Data [t × r × a Design] (Reading Comprehension: $n_t = 16$; Mathematics: $n_t = 9$)

	Reading Comprehension					*Mathematics*				
Source of Variation	*G Study*		*Alternative D Studies*			*G Study*		*Alternative D Studies*		
$n_r =$	3	2	2	2	3	3	2	2	2	3
$n_a =$	4	4	3	2	3	4	4	3	2	3
Teacher (t)	.385	.385	.385	.385	.385	.095	.095	.095	.095	.095
	(51%)	(81%)	(77%)	(70%)	(79%)	(36%)	(76%)	(71%)	(62%)	(74%)
Rater (r)[a]	0	0	0	0	0	.001	.0004	.0004	.0004	.0003
						(0.4%)	(0.3%)	(0.3%)	(0.3%)	(0.2%)
Assignment (a)	0	0	0	0	0	.002	.0005	.0006	.0009	.0006
						(0.8%)	(0.4%)	(0.5%)	(0.6%)	(0.5%)
tr	.021	.011	.011	.011	.007	0	0	0	0	0
	(3%)	(2%)	(2%)	(2%)	(2%)					
ta	0.273	.068	.091	.136	.091	.064	.016	.021	.032	.021
	(36%)	(14%)	(18%)	(25%)	(19%)	(24%)	(13%)	(16%)	(21%)	(17%)
ra	.0002	.0002								
	(<.01%)	(.04%)	0	0	0	0	0	0	0	0
Residual (tra, e)	.081	.010	.014	.020	.009	.102	.013	.017	.025	.011
	(11%)	(2%)	(3%)	(4%)	(2%)	(39%)	(11%)	(13%)	(17%)	(9%)
Absolute Error *SD*	.28	.30	.34	.41	.33	.16	.17	.20	.24	.18
Dependability Coefficient	.82	.81	.77	.70	.78	.80	.77	.71	.61	.74

[a] Negative variance components set to zero.

quality of classroom talk or cognitive demand of observed lesson activities. In mathematics, the ratings were moderately to highly correlated within all teaching domains (ranging from $r = .40$ to $r = .61$, $p < .01$).

Relationship of the IQA Ratings of Observed Instruction and Assignment Quality to Student Achievement

The different types of instructional quality measures varied in their relationships to student achievement. The ratings of assignment quality were moderately and significantly correlated with all the reading and mathematics achievement subscale scores. Correlations ranged from $r = .36$ to $r = .39$ ($p < .01$) for the reading comprehension subscale scores (Total Reading, Reading Comprehension, and Vocabulary) and from $r = .27$ to $r = .29$ ($p < .01$) for the mathematics achievement subscale scores (Total Math, Problem Solving, and Procedures). Observed instruction quality was significantly correlated with the mathematics Procedures score only ($r = .23$, $p < .01$).

Relationship of the IQA Ratings of Observed Instruction and Assignment Quality to Change in Student Achievement

Because the sample size of this study was too small to effectively analyze multi-level models, linear regression was used to investigate the relationship between the IQA observation and assignment ratings and student achievement controlling for students' past achievement and demographic characteristics. Separate multiple regression analyses specified the following subscale scores of the SAT-10 as the dependent variables: Total Reading, Reading Comprehension, and Vocabulary in English language arts; Total Mathematics, Procedures, and Problem-Solving in mathematics. The independent variables included students' prior year's achievement and demographic characteristics in addition to the IQA assignment quality and observed instruction ratings.

Reading comprehension. When both the observation and teacher assignment ratings were included in the same model, and after adjusting for students' background and prior achievement, the ratings of assignment quality positively and significantly predicted all the reading comprehension outcome scores: Total Reading ($\beta = .11$, $p = .01$), Reading Comprehension ($\beta = .11$, $p = .02$) and Vocabulary ($\beta = .19$, $p = .000$). The ratings of observed instruction positively and significantly predicted Reading Comprehension ($\beta = .10$, $p = .03$; see Table 10).

We conducted further regression analyses to investigate the potential of either the assignment ratings or the observation ratings in the models as a stand-alone measure of instructional quality. When used as the single indicator of

TABLE 10
Regression Results for Predicting SAT-10 Reading Scores

			Regression Estimates			
Dependent Variable	*Type of Predictors*	*Independent Variables*	*B*	*SE*	*β*	R^2
Total Reading Score	Prior Reading Achievement	Total Reading Score—Previous	.80	.05	.72***	.65***
	Student Characteristics	Gender	2.06	2.56	.03	
		Free/Reduced-Price Lunch	−5.23	3.14	−.06	
		Ethnicity	−5.45	3.57	−.06	
	IQA Ratings	IQA Assignment Score	4.97	1.84	.11**	
		IQA Observation Score	2.07	1.86	.04	
Reading Comprehension Score	Prior Reading Achievement	Reading Comprehension Score—Previous	.66	.05	.60***	.54***
	Student Characteristics	Gender	5.05	3.07	.07	
		Free/Reduced-Price Lunch	−11.01	4.03	−.12	
		Ethnicity	−8.45	4.26	−.09	
	IQA Ratings	IQA Assignment Score	5.20	2.20	.11*	
		IQA Observation Score	4.92	2.22	.10*	
Vocabulary Score	Prior Reading Achievement	Vocabulary Score—Previous	.70	.05	.64***	.53***
	Student Characteristics	Gender	−.33	3.46	−.00	
		Free/Reduced-Price Lunch	−1.03	7.76	−.01	
		Ethnicity	−4.46	4.91	−.04	
	IQA Ratings	IQA Assignment Score	9.70	2.45	.18***	
		IQA Observation Score	−1.70	2.52	−.03	

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

instructional quality, the quality of the assignments teachers gave to students significantly predicted change in student outcomes on all of the subscale scores (Total Reading, Reading Comprehension, and Vocabulary; Table 11). In contrast, the IQA observation rating score on its own did not significantly predict of any reading outcome (see Table 12).

Mathematics. Unlike reading comprehension, the IQA ratings of observed instruction and teacher assignments, when included in the same model, did not significantly predict change in student achievement on any of the SAT-10 mathematics subscale scores (see Table 13). One explanation for this pattern in mathematics may be a problem with multicollinearity with the strong correlation between mathematics observed instruction and assignment ratings ($r = .68$).

As with reading, we conducted additional regression analyses using either the IQA assignment quality ratings or the IQA observation ratings as the single measure of instructional quality. Results indicated that the IQA observation ratings alone significantly predicted change in student achievement on the Total Mathematics subscale (standardized $\beta = .16$, $p = .00$) and the Procedures subscale (standardized $\beta = .32$, $p = .000$; see Table 14). The IQA assignment ratings alone predicted change in student achievement on the Procedures subscore only (see Table 15).

SUMMARY AND DISCUSSION

This study compares two approaches that directly measure instruction in different content areas: teachers' assignments with student work and focused lesson observations. Our findings suggest that these approaches may be practical for large-scale applications.

Reliability of Ratings

Results verified an overall acceptable level of reliability and internal consistency for the rubrics assessing the quality of teachers' assignments with student work and observed instruction in both reading comprehension and mathematics. The exception to this finding was the rubric assessing the clarity and detail of the expectations that teachers hold for students. In both subject areas, raters disagreed on whether teachers provided sufficient information to do high-quality work. Teachers can be exceedingly clear to students regarding what they want them to do but not provide them with a great deal of detail, and our raters disagreed about what how much detail was "sufficient" or reasonable to expect from teachers. In addition, teachers can be overspecific regarding the amount of information given to students about what they need to do, or what high-quality

TABLE 11
Regression Results for Predicting SAT-10 Reading Scores (Assignment Ratings Only)

			Regression Estimates			
Dependent Variable	*Type of Predictors*	*Independent Variables*	*B*	*SE*	β	R^2
Total Reading Score	Prior Reading Achievement	Total Reading Score—Previous	.81	.05	.72***	.65***
	Student Characteristics	Gender	2.41	2.54	.04	
		Free/Reduced-Price Lunch	−5.67	3.42	−.07	
		Ethnicity	−5.40	3.58	−.06	
	IQA Ratings	IQA Assignment Score	4.80	1.84	.11**	
Reading Comprehension Score	Prior Reading Achievement	Reading Comprehension Score—Previous	.67	.05	.61***	.53***
	Student Characteristics	Gender	5.87	3.07	.08	
		Free/Reduced-Price Lunch	−12.18	4.02	−.14**	
		Ethnicity	−8.38	4.30	−.09	
	IQA Ratings	IQA Assignment Score	4.83	2.21	.10*	
Vocabulary Score	Prior Reading Achievement	Vocabulary Score—Previous	.69	.05	.64***	.53***
	Student Characteristics	Gender	−.60	3.43	−.01	
		Free/Reduced-Price Lunch	−.69	4.72	−.01	
		Ethnicity	−4.50	4.90	−.04	
	IQA Ratings	IQA Assignment Score	9.82	2.44	.18***	

***$p < .001$; **$p < .01$; *$p < .05$.

TABLE 12
Regression Results for Predicting SAT-10 Reading Scores (Observation Ratings Only)

			Regression Estimates			
Dependent Variable	*Type of Predictors*	*Independent Variables*	*B*	*SE*	*β*	R^2
Total Reading Score	Prior Reading Achievement	Total Reading Score—Previous	.85	.04	.76***	.63***
	Student Characteristics	Gender	2.02	2.26	.03	
		Free/Reduced-Price Lunch	−5.28	2.98	−.06	
		Ethnicity	−3.20	3.24	−.04	
	IQA Ratings	IQA Observation Score	.65	1.58	.01	
Reading Comprehension Score	Prior Reading Achievement	Reading Comprehension Score—Previous	.71	.05	.65***	.50***
	Student Characteristics	Gender	3.97	2.75	.06	
		Free/Reduced-Price Lunch	−8.90	3.52	−.10*	
		Ethnicity	−6.53	3.91	−.07	
	IQA Ratings	IQA Observation Score	2.64	1.89	.06	
Vocabulary Score	Prior Reading Achievement	Vocabulary Score—Previous	.73	.05	.67***	.48***
	Student Characteristics	Gender	1.70	3.13	.02	
		Free/Reduced-Price Lunch	−5.79	4.20	−.06	
		Ethnicity	−1.35	4.54	−.01	
	IQA ratings	IQA Observation Score	−3.11	2.18	−.06	

$^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$.

TABLE 13
Regression Results for Predicting SAT-10 Mathematics Scores

Dependent Variable	*Set of Predictors*	*Independent Variables*	*Regression Estimates*			
			B	*SE*	β	R^2
Total Mathematics Score	Prior Mathematics Achievement	Total Mathematics Score—Previous	.66	.04	.81***	.81***
	Student Characteristics	Gender	7.59	3.36	.08*	
		Free/Reduced-Price Lunch	−7.32	5.00	−.07	
		Ethnicity	−7.01	4.27	−.07	
	IQA Ratings	IQA Assignment Score	8.69	5.99	.07	
		IQA Observation Score	−1.22	4.67	−.01	
Problem Solving Score	Prior Mathematics Achievement	Problem Solving Score—Previous	.81	.06	.80***	.76***
	Student Characteristics	Gender	−12.22	4.35	−.12**	
		Free/Reduced-Price Lunch	−9.02	6.66	−.07	
		Ethnicity	−.02	5.61	−.00	
	IQA Ratings	IQA Assignment Score	7.08	7.83	.05	
		IQA Observation Score	−1.16	6.02	−.01	
Procedures Score	Prior Mathematics Achievement	Procedures Score—Previous	.45	.05	.62***	.52***
	Student Characteristics	Gender	2.88	5.49	.03	
		Free/Reduced-Price Lunch	−2.28	7.98	−.02	
		Ethnicity	−10.09	7.00	−.10	
	IQA Ratings	IQA Assignment Score	15.32	9.75	.12	
		IQA Observation Score	.83	7.68	.01	

***$p < .001$; **$p < .01$; *$p < .05$.

TABLE 14
Regression Results for Predicting SAT-10 Mathematics Scores (Observation Ratings Only)

			Regression Estimates			
Dependent Variable	*Set of Predictors*	*Independent Variables*	*B*	*SE*	*β*	R^2
Total Mathematics Score	Prior Mathematics Achievement	Total Mathematics Score—Previous	.68	.03	.83***	.77***
	Student Characteristics	Gender	7.35	2.84	.09**	
		Free/Reduced-Price Lunch	−6.26	4.00	−.07	
		Ethnicity	−6.39	3.68	−.07	
	IQA Ratings	IQA Observation Score	9.79	2.12	16***	
Problem Solving Score	Prior Mathematics Achievement	Problem Solving Score—Previous	.83	.04	.82***	.74***
	Student Characteristics	Gender	−11.18	3.46	−.12***	
		Free/Reduced-Price Lunch	−5.00	5.01	−.04	
		Ethnicity	−.80	4.56	−.01	
	IQA Ratings	IQA Observation Score	−3.95	2.56	−.06	
Procedures Score	Prior Mathematics Achievement	Procedures Score—Previous	.47	.05	.58***	.42***
	Student Characteristics	Gender	3.19	5.23	.03	
		Free/Reduced-Price Lunch	−3.09	7.24	−.03	
		Ethnicity	−9.18	6.77	−.08	
	IQA Ratings	IQA Observation Score	22.70	3.94	.32***	

$^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$.

TABLE 15
Regression Results for Predicting SAT-10 Mathematics Scores (Assignment Ratings Only)

			Regression Estimates			
Dependent Variable	*Set of Predictors*	*Independent Variables*	*B*	*SE*	*β*	R^2
Total Mathematics Score	Prior Mathematics Achievement	Total Mathematics Score—Previous	.66	.04	.81***	.81***
	Student Characteristics	Gender	7.57	3.35	.08*	
		Free/Reduced-Price Lunch	−7.36	4.98	−.07	
		Ethnicity	−7.03	4.26	−.07	
	IQA Ratings	IQA Assignment Score	7.69	4.59	.07	
Problem Solving Score	Prior Mathematics Achievement	Problem Solving Score—Previous	.81	.06	.80***	.76***
	Student Characteristics	Gender	−12.23	4.33	−.12**	
		Free/Reduced-Price Lunch	−9.08	6.63	−.07	
		Ethnicity	−.04	5.59	−.00	
	IQA Ratings	IQA Assignment Score	6.10	5.95	.04	
Procedures Score	Prior Mathematics Achievement	Procedures Score—Previous	.45	.05	.62***	.52***
	Student Characteristics	Gender	2.90	5.47	.03	
		Free/Reduced-Price Lunch	−2.25	7.95	−.02	
		Ethnicity	−10.09	6.98	−.10	
	IQA Ratings	IQA Assignment Score	16.00	7.46	.13*	

***$p < .001$; **$p < .01$; *$p < .05$.

work looks like. In mathematics particularly, if a teacher provides a great deal of detailed direction to students (e.g., provides a solution path for solving a complex problem), the rigor of the task can and likely will be compromised. We are continuing to refine this rubric so that it is more closely aligned with instructional demands of the different content areas.

Assignments. In both reading comprehension and mathematics, relatively few assignments and observations were needed to yield a reliable estimate of a teacher's practice. Specifically, our results indicated that as few as four assignments yielded a reliable teacher-level estimate of instructional quality. This finding is commensurate with the results of other research efforts investigating the use of teachers' assignments as an indicator of instructional practice in mathematics at the elementary school level (e.g., Boston & Wolf, 2006). These results also are in line with other research conducted through the Center for Research on Evaluation, Standards, and Student Testing indicating that four assignments yields a reliable estimate of practice in English language arts at the secondary level (Clare, 2000; Clare & Aschbacher, 2001; Matsumura, Garnier, Pascal, & Valdés, 2002).

That a reliable estimate of the quality of a teacher's practice can be gained with only four assignments greatly adds to the feasibility of this approach for measuring instruction at-scale. While teachers vary in their reports of how long it takes to complete the materials, on average it appears to take about 45 minutes per assignment to complete the cover sheets and choose and Xerox the student work samples. Collecting four assignments, therefore, would take about three hours of a teacher's time. Further adding to the feasibility of collecting assignments in large numbers of classrooms is the fact that relatively few raters need to be hired to rate large numbers of assignments. This reduces the cost and training burden on researchers (or schools and districts) and increases the potential for rater agreement.

Observations. Our results also confirmed that when teachers comply with the data collection requirements, as few as two observations might yield a reliable estimate of quality at the teacher-level in either content area. Unlike classroom measures focusing on the content of teachers' instruction in which more data points are required because the content of instruction is fluid over the course of the year, the IQA measures of assignment quality and classroom discussions were more stable over time and, therefore, required fewer data points (e.g., Rowan et al., 2004). One reason for this finding could be that the same rater observed any given classroom on both occasions. Another likely explanation is our decision to narrow the scope of measurement for the observations. As described earlier, we focused on a specific classroom activity (discussion in both content areas). We also observed on consecutive days, thus minimizing potential variation in an individual teacher's instructional practice that could result from variation in the curricula topic being covered. Observing on consecutive days also minimized the potential for variation in an individual teacher's practice as a function of the time of year. It is possible, for example, that instruction would look different at the beginning of the year (when a teacher is establishing class rules and norms for participation) than at the end of the year, or near the time when students participate in standardized testing.

Narrowing the scope of measurement, as we did, appears to reduce the number of data points needed to obtain a reliable estimate of practice. This strategy adds to a measure's potential feasibility for use at scale because it reduces the cost of data collection, burden on teachers, and the amount of content that needs to be addressed in a rater-training program. Rater training for the IQA observation protocol, for example, focuses solely on assessing instruction in whole-group settings rather than across multiple instructional activities (e.g., small group discussions, one-on-one conferring, etc.).

However, narrowing the scope of measurement reduces the generalizability of the results (the inferences one can make about a teacher's practice). Although it seems logical that a teacher with the ability to lead rich whole-class discussions about a text would do so as a general rule in other areas of their literacy instruction, it is not possible to make such a conclusion from these data. If the goal of measurement is to obtain an estimate of what a teacher's practice looks like across *different* instructional settings and content, then more data points likely would be needed. If the goal of measurement is to obtain a cross-sectional snapshot of teachers' instructional quality, then the results from this study provide evidence that these snapshots and students' achievement are significantly related.

In terms of feasibility for measuring instruction in large numbers of classrooms, reducing the number of observations needed to gain a reliable estimate of teaching quality greatly increases the potential of this approach for use in large numbers of classrooms. In this regard, observations have an advantage over assignments in that they pose minimal (if any) burden on teachers. Teachers do not need to prepare special lessons to be observed. On the other hand, observing large numbers of teachers would pose a tremendous burden on districts, school, or researchers because they would need to hire and train large numbers of raters. In our experience, it takes about four days of training (per content area) and at least two additional in-classroom sessions (novice and expert raters observing the same lesson and comparing scores afterward) to be considered a "reliable" rater. It is notable that some people, even after participating in the training, can never become reliable raters. If a district wanted to observe teachers within a fairly narrow window of time (as is required in most research designs), then many raters would need to be hired and trained, and not all of these people likely would complete the training successfully. Additionally, we found that it is necessary for raters to have significant knowledge of the subject matter they are rating—not just in terms of understanding the content but also understanding what is considered grade-level work in that content area (Matsumura et al., 2006). Finding enough qualified raters to observe large numbers of classrooms could pose a significant challenge to districts.

Ratings in Different Content Areas

While the reliability of the observation and assignment ratings were similar for mathematics and reading comprehension, the relation of these ratings to each other and to student achievement differed by content area. In mathematics, the quality of teachers' assignments was associated with the quality of observed classroom discussions, tasks, and expectations expressed to students for the quality of their work. This was not the case in English language arts. The quality of teachers' assignment tasks were not associated with the quality of teachers' classroom discussions and lesson activities and were only weakly associated with the expectations teachers expressed to students for the quality of their work.

Our results are different from Clare and Aschbacher (2001) who found that the ratings of teachers' language arts assignments in elementary and middle schools yielded similar estimates of quality for some aspects of observed instruction. One explanation for the difference in our findings is that in the current study we asked for examples of challenging tasks from middle school teachers. Clare and Aschbacher, in contrast, mostly looked at assignments considered by teachers to be "typical" of their practice, and combined elementary and middle school teachers in their analyses. In English language arts, more challenging tasks generally entail a longer period of time to complete than a few class periods—especially at the secondary level. Even when teachers provided students with more rigorous, extended opportunities to write about a text, we were only able to observe a part of the project in the limited number of visits we made to the classroom. Looking across observation field notes, teachers who were observed leading higher cognitive demand class discussions about a text may or may not have provided students with the opportunity to write meaningfully about what they were reading. It is notable, however, that we did not see any examples of teachers leading low-level discussions, and submitting cognitively challenging assignments and student work.

In mathematics, a greater consistency appeared between the quality of observed instruction and teachers' assignments likely because the mathematics teachers used a structured curriculum (Everyday Mathematics). The language arts teachers, in contrast, used a scope and sequence plan that set out the books intended for students to read at different points in the year, but did not prescribe the student activities teachers assigned to build comprehension or write in response to these texts. With more freedom to design and plan activities, the language arts teachers engaged in a broader range of classroom activities than did the mathematics teachers. Another reason for the similarity between observation and quality of assignment measures in mathematics also may be because students generally worked on a problem in its entirety during class time. In contrast to the multiple steps involved in producing an extended essay or research report, one is less likely to see just a portion of a multi-day project in a mathematics class.

Ratings and Student Achievement

Across both content areas, the ratings of assignment quality correlated with student outcomes on all the ratings subscales. The relation of the instructional quality ratings to student achievement outcomes within the different content areas changed, however, when we controlled for prior achievement and student characteristics. In English language arts, the assignment ratings were the strongest predictor of student achievement across all of the subscale scores, although both assignments and classroom observations predicted higher achievement on the more challenging reading content (reading comprehension). Including both measures in the same model did not make an appreciable difference in the percent of variance explained by the instructional quality measures. In sum, the assignment measure—focusing on the interactions between teachers, students, and content in the service of developing students' academic writing skills—may be the better measure of instructional quality in English language arts in terms of predicting gains in student achievement.

In mathematics, the relation of the instructional quality ratings and student achievement portrayed a different pattern. Although the assignment ratings correlated more strongly with mathematics achievement than with classroom observations, both observations and assignments predicted *gain* in mathematics procedures achievement in separate models, with observations being the stronger predictor. Classroom observations also better predicted gain in total mathematics scores. Even though the quality of teachers' assignment tasks and observed practice were highly correlated, our findings point to observed social interactions around the content presented in classrooms as the better strategy for capturing opportunities to learn mathematics.

Gains in the more challenging problem-solving content were predicted only by students' prior problem-solving achievement and characteristics (gender and SES). The amount of variance remaining to be accounted for by quality of instruction was quite small and may explain why the IQA measures were weak predictors of the Problem-Solving subscale. Interestingly, these results are consistent with other research that found that estimates of teacher effects on middle schools students' learning differed depending on whether the Procedures or Problem-Solving subscales of the SAT-9 were used as an outcome measure (Lockwood, McCaffrey, Hamilton, Stecher, Le, & Martinez, 2007).

Why did the IQA measures of instruction differentially predict student achievement in reading comprehension and mathematics? Possibly, when a structured curriculum such as mathematics is being used, observations may be a better method for describing instructional quality since it is the quality of the conversation around tasks (rather than the tasks themselves) that would likely exhibit greater variability. This explanation is supported by the fact that the quality of the class discussions in the observed mathematics lessons showed more variability than the quality of the tasks that teachers assigned to students. When teachers are responsible for creating their own tasks, as the reading comprehension teachers in this study were, assignments may be a more robust measure of instructional quality since a single assignment covers a significantly longer period of instructional time than a single observation. An assignment encapsulates an instructional cycle in which teachers communicate an objective (e.g., a skills or set of skills they want students to master), students practice or enact those skills in their work, and teachers then provide feedback to students on their efforts. An observation only provides insight into a single (in our study, 50-minute) classroom period. Collecting multiple assignments from teachers, as we did, clearly increased the amount of instructional time covered.

Limitations and Future Research

It is important to reiterate that the regression results should be interpreted with caution. Some other unmeasured aspect of teacher quality, for example, could influence the relationship between the IQA ratings and student achievement, and the IQA ratings could be correlated with this other construct. Moreover, without accounting for the clustering of students within classrooms and schools as do multilevel models, the linear regression analyses reported here may estimate effects that appear stronger than they actually are. Further research is needed in larger samples of classrooms that allow researchers to control for the nesting of students within classrooms. Research also is needed to disentangle content area and curricula specificity by examining in greater depth the relation of teachers' assignments and observations of instruction to student learning in different content areas.

Finally, while the IQA measures show promise for providing insight on teachers' instructional practice with a minimal number of data points, the amount of variation they predicted in student achievement was small overall. As described earlier, the inferences about student achievement may be limited. Additional research is needed concerning combining the IQA direct measures of instructional practice with other types of measures (such as instructional logs) that focus on the amount and type of content covered over the course of a year to provide a richer, more complete picture of students' opportunity to learn. Combining relatively inexpensive and less demanding measures that focus attention on different aspects of instructional practice may be the best approach for providing meaningful information about practice in large numbers of classrooms (i.e., measuring the instructional dynamic at scale). Which combination of measures to use in different content areas, and perhaps level of schooling, are questions to be addressed in future research.

In conclusion, measuring instruction at scale is critical for providing information to the public about school quality, supporting better decision-making in districts, and directing attention toward excellent instructional practice (students' opportunity to learn) as a desired outcome in and of itself for education reform. This issue is especially timely given the questionable quality of the achievement tests and standards for proficiency set in many states (e.g., Kingsbury, Olson, Cronin, Hauser, & Houser, 2003; Peterson & Hess, 2005). At the same time, it is important to recognize that measuring instruction in schools and districts could have negative consequences. This is likely to be especially the case if stakes were attached to the outcomes. Score inflation and narrowing of instruction to focus on the teaching behaviors assessed to the exclusion of other important dimensions of instructional practice are possibilities. Should measures of instruction be adopted for use by districts and schools as a routine course of action, the effect measuring instruction might have on teaching (for better or for worse) would need to be monitored. Careful attention also would need to be paid to how information from these assessments is used to increase the possibility for positive consequences; that is, for measures of instruction to support improvement in the quality of students' learning opportunities in schools.

REFERENCES

Allensworth, E., Correa, M., & Ponisciak, S. (May, 2008). *From high school to the future: ACT preparation—Too much, too late*. Chicago: Consortium on Chicago School Research.

Aschbacher, P. R. (1999). *Developing indicators of classroom practice to monitor and support school reform* (CSE Technical Report #513). University of California, Los Angeles. Center for the Study of Evaluation; Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.

Baker, E. L. (2007). The end(s) of testing. *Educational Researcher*, *36*, 309–317.

Ball, D. L., & Forzani, F. M. (2008). What makes education research "educational"? *Educational Researcher, 36*(9), 529–540.

Ball, D. L., & Rowan, B. (2004). Measuring instruction. *The Elementary School Journal*, *105*(1), 3–10.

Beck, I. L., McKeown, M. G., Hamilton, R. L., & Kucan, L. (1997). *Questioning the Author: An approach for enhancing student engagement with text.* Delaware: International Reading Association.

Black, P., & Wiliam, D. (1998). Inside the black box. Raising standards through classroom assessment. *Phi Delta Kappan 80*(2), 139–148.

Boston, M., & Wolf, M. (2006). *Assessing Academic Rigor in Mathematics Instruction: The Development of the Instructional Quality Assessment Toolkit* (CSE Technical Report #672). University of California, Los Angeles. Center for the Study of Evaluation; Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How people learn: Brain, mind, experience and school.* Committee on Developments in the Science of Learning, National Research Council Washington, DC: National Academy press.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience and school: Expanded edition* (Additional material from the Committee on Learning Research and Educational Practice, M. S. Donovan, J. D. Bransford, and J. W. Pellegrino, Eds.). Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, DC: National Academy press.

Brennan, R. L. (2001a). Manual for mGENOVA. Iowa City, IA: Iowa Testing Programs, University of Iowa.

Brennan, R. L. (2001b). mGENOVA (Version 2.1) [Computer software] [On-line]. Available: http://www.education.uiowa.edu/casma/computer_programs.htm

Clare, L. (2000). *Using teachers' assignments as an indicator of classroom practice* (CSE Technical Report #532). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Clare, L., & Aschbacher, P. (2001). Exploring the technical quality of using assignments and student work as indicators of classroom practice. *Educational Assessment, 7*(1), 39–59.

Cobb, P., Boufi, A., McClain, K., & Whitenack, J. (1997). Reflective discourse and collective reflection. *Journal for Research in Mathematics Education, 28*(3), 258–277.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Educational Policy Analysis Archives*, *8*(1). Retrieved February 28, 2001 from http://olam.ed.asu.edu/epaa/v8n1

Delpit, L. D. (1988). The silenced dialogue: Power and pedagogy in educating other people's children. *Harvard Educational Review, 58,* 280–298.

Doyle, W. (1983). Academic work. *Review of educational research, 53,* 159–199.

Doyle, W. (1988). Work in mathematics classes: The context of students' thinking during instruction. *Educational psychologist*, *23*, 197–180.

Dunbar, S., Koretz, D., & Hoover, H. D. (1991). Quality control in the development and use of performance assessment. *Applied Measurement in Education*, *4*(4), 289–303.

Goldenberg, C. (1992/1993). Instructional conversations: Promoting comprehension through discussion. The Reading Teacher, *46*, 316–326.

Grossman, P. L., Stodolsky, S. S., & Knapp, M. (2004). Making subject matter part of the equation: The intersection of policy and content. Occasional Paper (Document 0-04-1), Center for the Study of Teaching and Policy, University of Washington, Seattle.

Hamilton, L. (2003). Assessment as a policy tool. *Review of Research in Education*, *27*, 25–68 (focus on pages 25–32).

Hill, H., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Education Research Journal, 42*(2), 371–406.

Junker, B. W., Matsumura, L. C., Crosson, A., Wolf, M. K., Levison, A., Wiesberg, J., et al. (2006). *Overview of the Instructional Quality Assessment*. (CSE Technical Report #671). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kingsbury, G. G., Olson, A., Cronin, J., Hauser, C., & Houser, R. (2003). *The state of state standards: Research investigating proficiency levels in fourteen states*. Portland, OR: Northwest Evaluation Association.

Koretz, D., & Barron, S. (1999). *The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS).* RAND Research Brief.

Koretz, D., & Hamilton, L. (2006). Testing for accountability in K-12. In R. L. Brennen (Ed.), *Educational Measurement* (4th edition, pp. 531–542). Westport, CT: American Council on Education/ Praeger.

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher, 32*(7), 3–13.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, J. P. (2007). The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures. *Journal of Educational Measurement*, *44*(1), 47–67.

Matsumura, L. C., Garnier, H., Pascal, J., & Valdés, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and student achievement. *Educational Assessment, 8*(3), 207–229.

Matsumura, L. C., Slater, S. C., & Crosson, A. (2008). Classroom climate, rigorous instruction and curricula, and students' interactions in urban middle school classrooms. *Elementary School Journal, 108*(4), 293–312.

Matsumura, L. C., Slater, S. C., Junker, B., Peterson, M., Boston, M., Steele, M., et al. (2006). *Measuring reading comprehension and mathematics instruction in urban middle schools: A pilot study of the Instructional Quality Assessment*. (CSE Technical Report #681). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

New England Common Assessments Program Reporting. (2006). Retrieved February 26, 2006, from http://reporting.measuredprogress.org/NECAPpublicRI/select.aspx.

Newmann, F., Bryk, A. S., & Nagaoka, J. K. (2001). *Authentic Intellectual Work and Standardized Tests: Conflict or Coexistence*? Chicago: Consortium on Chicago School Research.

Newmann, F. M., Lopez, G., & Bryk, A. S. (October 1998). *The quality of intellectual work in Chicago schools: A baseline report,* 1–59. Retrieved December 2006, from Consortium on Chicago School Research database.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, §115 Stat. 1425 (2002).

O'Connor, M. C., & Michaels, S. (1996). Shifting participant frameworks: Orchestrating thinking practices in group discussions. In D. Ghicks (Ed.), *Discourse, Learning, and Schooling* (pp. 63–103). New York: Cambridge University Press.

Peterson, P. E., & Hess, F. M. (Summer, 2005). Johnny can read . . . in some states: Assessing the rigor of state assessment systems. *Education Next*, 52–53.

Resnick, L., & Hall, M.W. (2001). The Principles of Learning: study tools for educators. [CDROM]. Pittsburgh, PA: Institute for Learning, Learning Research and Development Center, University of Pittsburgh.

Resnick, L. B., & Matsumura, L. C. (2007). Academic proficiency: Bright hopes, blurry vision. *Voices in Urban Education*, *14*, 9–21.

Rowan, B., Camburn, E., & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum: A study of literacy teaching in third-grade classrooms. *Elementary School Journal, 105*, 75–102.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*(1), 1–30.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215–232.

Snow, C. E. (2002). *Reading for understanding: Toward a research and development program in reading comprehension.* Prepared for the Office of Research and Improvement. Santa Monica, CA: RAND Corporation.

Spillane, J., & Louis, K. S. (2002). School improvement processes and practices: Professional learning for building instructional capacity. *Yearbook of the National Society for the Study of Education*, *101*(1), 83–104.

Stein, M. K., Smith, M. S., Henningsen, M. A., & Silver, E. A. (2000). *Implementing standards-based mathematics instruction: A casebook for professional development.* New York: Teachers College Press.

Stodolsky, S. S. (1988) *The subject matters: Classroom activity in math and social studies*. Chicago: University of Chicago Press.

Stodolsky, S. S., & Grossman, P. A. (1995). The impact of subject matter on curricular activity: An analysis of five academic subjects. *American Educational Research Journal*, *32*(2), 227–249.