Teachers College Record
The Voice of Scholarship in Education

The Instructional Challenge in Improving Teaching Quality: Lessons From a Classroom Observation Protocol

by Drew Gitomer, Courtney Bell, Yi Qi, Daniel McCaffrey, Bridget K. Hamre & Robert C. Pianta — 2014

Background/Context: *Teacher evaluation is a major policy initiative intended to improve the quality of classroom instruction. This study documents a fundamental challenge to using teacher evaluation to improve teaching and learning.*

Purpose: *Using an observation instrument (CLASS-S), we evaluate evidence on different aspects of instructional practice in algebra classrooms to consider how much scores vary, how well observers are able to judge practice, and how well teachers are able to evaluate their own practice.*

Participants: *The study includes 82 Algebra I teachers in middle and high schools. Five observers completed almost all observations.*

Research Design: *Each classroom was observed 4–5 times over the school year. Each observation was coded and scored live and by video. All videos were coded by two independent observers, as were 36% of the live observations. Observers assigned scores to each of 10 dimensions. Observer scores were also compared with master coders for a subset of videos. Participating teachers also completed a self-report instrument (CLASS-T) to assess their own skills on dimensions of CLASS-S.*

Data Collection and Analysis: *For each lesson, data were aggregated into three domain scores, Emotional Support, Classroom Organization, and Instructional Support, and then averaged across lessons to create scores for each classroom.*

Findings/Results: *Classroom Observation scores fell in the high range of the protocol. Scores for Emotional Support were in the midlevel range, and the lowest scores were for Instructional Support. Scores for each domain were clustered in narrow ranges. Observers were more consistent over time and agreed more when judging Classroom Organization than the other two domains. Teacher ratings of their own strengths and weaknesses were positively related to observation scores for Classroom Organization and unrelated to observation scores for Instructional Support.*

Conclusions/Recommendations: *This study identifies a critical challenge for teacher evaluation policy if it is to improve teaching and learning. Aspects of teaching and learning in the observation protocol that appear most in need of improvement are those that are the hardest for observers to agree on, and teachers and external observers view most differently. Reliability is a marker of common understanding about important constructs and observation protocols are intended to provide a common language and structure to inform teaching practice. This study suggests the need to focus our efforts on the instructional and interactional aspects of classrooms through shared conversations and clear images of what teaching quality looks like.*

Almost 30 years ago, researchers began documenting the mediocrity of teaching practice. Goodlad (1984) described classrooms in which teachers controlled almost all of the discourse and students were not intellectually engaged, were asked few questions beyond factual recall, and did not explore ideas in any depth. Studies since then have provided substantial evidence that little has changed (Gonzales et al., 2008; Horizon Research, Inc., 2000; Rowan, Harrison, & Hayes, 2004). Researchers, policymakers, and citizens from a range of political perspectives have argued that the mediocre state of U.S. teaching is related to many critical outcomes—the economy, the health of our democracy, and our global status (Hanushek & Woessmann, 2011; National Commission on Excellence in Education, 1983). Thus, if we are to accomplish more ambitious learning for more students, the quality of interactions between teachers and students must change.

One increasingly popular approach to improving interactions is to use teacher evaluation to leverage change inside classrooms. Current teacher evaluation approaches are built on assumptions about both how best to improve the teacher workforce and how to improve an individual teacher's practices. Focusing on individual improvement, a teacher evaluation system should create information (e.g., scores on observation protocols, value-added scores, narratives about areas of strength and areas for growth) that can be used to identify specific teaching practices that need to be strengthened. Improvements in those teaching practices will then lead to improvements in student learning. In this paper, we document a fundamental challenge to using teacher evaluation to improve the mediocrity of teaching and learning.

IMPROVING TEACHING AND LEARNING THROUGH TEACHER EVALUATION SYSTEMS

Current policy initiatives are focused on differentiating teachers to identify them in terms of their relative effectiveness as referenced against measures of student achievement (U.S. Department of Education, 2009). This policy press, and particularly the emphasis on quantitative approaches to rating teachers, has grown out of a deep frustration with the status quo practices of teacher evaluation. Almost all teachers receive the same evaluation; an unsatisfactory evaluation is a rare event in most school districts. Weisberg, Sexton, Mulhern, and Keeling (2009) demonstrated the lack of differentiation among teachers by school administrators in formal evaluation reports, even when administrators acknowledged, in private conversations, the differential

performance of these same teachers. Weisberg et al. and others (e.g., Glazerman et al., 2011) have made the argument that school administrators, for a variety of reasons, do not have the will to make judgments that differentiate teachers, even when they can recognize that some teachers are substantially stronger or weaker than others. Other metrics used in the teacher preparation and compensation arenas (e.g., teacher licensure exams and advanced degrees) also fail to discriminate or relate to student performance and are thus suspect as indicators of effectiveness (Wayne & Youngs, 2003). The larger literature showing that particular teachers are more consistently associated with stronger achievement gains in their students than other teachers offers the promise of systems that recognize these differences.

In response, new teaching evaluation systems are being designed and implemented by states and districts across the country. These systems vary in many details. For example, some include measures of parent or student perceptions (e.g., Burniske & Meibaum, 2012; Ripley, 2012); others include measures of teachers' professional activities outside of the classroom (e.g., Rhode Island Department of Education, 2011). Despite these differences, at the heart of most systems are two classes of measures—one based on student achievement and the other based on classroom observation. In this paper, we investigate empirical issues associated with using observations to improve teaching quality in middle and high school Algebra I classrooms.

Observations are used as a means of obtaining a direct measure of teaching practice and can generally be described as follows: A protocol is structured around a set of *domains* that describe the core constructs of teaching valued by the protocol (e.g., emotional support, classroom organization, and instructional support). Each domain is then defined by a set of *dimensions* of teaching (e.g., engagement and productivity). Though these dimensions are not intended to overlap, the integrated nature of teaching and inherent connection among different aspects of teaching typically result in strong correlations among dimensions. Each dimension is assigned scores by a trained observer on a numerical scale that typically has descriptive anchors explaining most, if not all, of the scale points. Observations of the teacher are made for some specified number of lessons over the school year. For each observation, observers assign scores using the method mandated by the protocol system. In some systems, a full lesson is observed and scored in its entirety. In other cases, a full lesson is divided into shorter segments of time, and each segment is scored. Finally, in some instances, only portions of a lesson are scored. Dimension scores are subsequently aggregated in some way to determine an overall observation score, or set of domain scores, for the teacher for the year. Often, observation scores are then combined with other information (e.g., value-added estimates and student learning objective scores) to inform a summary impression of the teacher's effectiveness.

The validity of observation scores can be evaluated through a set of processes and arguments described by Bell et al. (2012). If the purpose of the observation is to improve teaching, then a number of assumptions would be evaluated to support the validity argument. First, the observation protocol must be able to provide information about different aspects of teaching practice. For example, the protocol must define classroom behaviors in such a way that scores and any written notes from the observation provide useful information about different aspects of teaching, but only to the extent that the validity of inferences made on the basis of the observation has support.

Second, the protocol should provide a common framework and language for considering and developing teaching. For a framework and language to have common meaning among teachers, principals, and administrators within a district, a protocol's descriptions of teaching, definitions of terms, and distinctions across levels of quality need to be understood by all participants in the same way. Do mathematics teachers and their principals have a common understanding of what it means to engage students in deep reasoning about a mathematical concept? Do they agree on whether such reasoning is evident when they look at the same segment of teaching? Varied research literatures, from teacher learning to education policy and education administration, suggest that educators lack common understandings of critical teaching practices as well as the behaviors that instantiate a specific type of practice (e.g., deep reasoning and conceptual understanding; Cohen, 1990; Grossman & McDonald, 2008; Sherin & Han, 2004). Therefore, discussions of teaching practice within a system will be limited to the extent that such understandings are idiosyncratic and not shared within a community of practice.

In the same way, the potential for using observation protocols to make valid judgments about teaching quality depends, in large part, on the extent to which different observers make the same judgments about teaching quality, given the same classroom evidence. For a system to be valid, judgments should not be determined by who makes the judgments or when they make such judgments. This means that observers should generally agree on what they are seeing and how it should be scored using the observation system. The degree of observer agreement is one indicator of the extent to which there is a common understanding of teaching within the community of practice. Finally, it should not matter if a classroom's observations occurred at the beginning, middle, or end of the school year unless the actual quality of teaching changes during the course of the year.

If observers cannot agree on scores because of a lack of common understanding about aspects of teaching, this should not automatically be taken as a limitation of the protocol. Instead, the lack of agreement may be a signal that the observers lack the shared understanding that is necessary to move the system forward in productive and coherent ways. Improved reliability over time may not only result in more accurate scores, but may indicate an increasingly shared understanding of the nature of teaching quality among observers.

To summarize, in order for observations to be valid and useful for improving teaching, at least two basic criteria must be met: The protocol must be able to distinguish among different aspects of teaching practice, and stakeholders must have a common understanding of teaching and learning such that who observes or when that individual observes does not unduly influence the scores assigned to the teaching. In this paper, we argue that the validity of scores is not uniform within the observation protocol under investigation. Observers are more apt to agree on scores for certain aspects of teaching, as represented by domains in the protocol, than for other aspects. Specifically, the instructional aspects of classroom practice are particularly difficult for observers to see in similar ways. Scores on these aspects of teaching are also more apt to be influenced by who observes the teacher and

when the teacher is observed. On the other hand, more traditional measures of teaching (i.e., classroom management) are most readily understood. We then argue that if the relative lack of agreement is a signal that there is not a shared understanding of teaching quality on these instructional dimensions of teaching, the ability to improve instructional practices may be particularly challenging.

RESEARCH QUESTIONS

This study involved observations and other measures of algebra classrooms using the *Classroom Assessment Scoring System for Secondary Classrooms* (CLASS-S) instrument (Pianta, Hamre, Haynes, Mintz, & La Paro, 2009) to investigate the following questions:

1. *To what degree does the classroom observation protocol produce scores that vary across different dimensions of teaching practice in algebra classrooms?* Previous studies of mathematics classrooms have suggested that across most classrooms nationally, aspects of teaching associated with intellectual rigor, cognitive challenge, and sense-making are executed relatively poorly (Weiss, Pasley, Smith, Banilower, & Heck, 2003). On the other hand, Weiss et al. found evidence that mathematics classrooms had rated more highly on dimensions of respect for students, student participation, and other markers of positive social practices. In this study, we asked whether a classroom observation protocol designed for large-scale application was sensitive to and helped provide insight into teaching practices in algebra classrooms.

2. *How well are observers able to judge different dimensions of teaching practice?* This study explored whether certain aspects of teaching are more or less difficult for observers to judge. Of interest is the extent to which observers make judgments consistent with expert observers. Also, to what degree do observers' scores change as they practice scoring and learn to consistently score different aspects of teaching practice?

3. *How well are teachers able to evaluate their own teaching practice across different dimensions?* We investigate the extent to which teachers estimate the quality of their own teaching on different dimensions of practice. We are interested in whether teachers' judgments about their own teaching vary across different aspects of practice, whether their judgments align with those of the observation protocol, and if that alignment depends on which aspect of practice is considered.

METHOD

STUDY DESIGN

This study is part of the *Toward an Understanding of Classroom Context* (TUCC) validation project of a relatively new classroom observation instrument, which was conducted in algebra classrooms (Bell et al., 2012). As part of the TUCC project, teachers and classrooms were studied with a range of instruments, including a classroom observation protocol measuring classroom interactions; a self-report measure about classroom interactions; several tests of teacher knowledge; value-added scores based on tests of student achievement in algebra; a survey of teacher attitudes about different aspects of teaching; and a survey of student and teacher views about intelligence (Blackwell, Trzesniewski, & Dweck, 2007). This study focuses only on the observation and self-report instruments concerning classroom interactions.

OBSERVATION INSTRUMENT: THE *CLASSROOM ASSESSMENT SCORING SYSTEM FOR SECONDARY CLASSROOMS* (CLASS-S)

CLASS-S is part of the *Classroom Assessment Scoring System*, a system of research-based observation protocols (preK, K–3, and secondary) developed at the University of Virginia that is designed to measure PK–12 classroom interactions (Hamre & Pianta, 2005; La Paro, Pianta, & Stuhlman, 2004; Pianta et al., 2005; Rimm-Kaufman, La Paro, Downer, & Pianta, 2005). CLASS-S measures the *Teaching Through Interactions* (TTI) model of classroom interaction (Hamre et al., 2013). The TTI model developed from extensive theoretical and empirical work in preschool and early elementary school classrooms and takes a developmental perspective on teaching and learning, paying particular attention to teacher and student interactions that support academic and emotional growth as well as the purposefulness and productivity of the classroom.

All observation protocols in the CLASS system, including CLASS-S, are built on the same theoretical model and organized around the same three domains of teacher–student interactions: Classroom Organization, Emotional Support, and Instructional Support (Allen et al., 2013). Each domain is assessed by three or four specific dimensions of teacher–student interactions (Figure 1). Specific features of each dimension are indicated by discrete behaviors (and their patterns), and these differ across the protocols of the CLASS system. For example, having a teacher who is sensitive (e.g., well-calibrated to cues from the student) is important in both high school and elementary school, but because teachers in these two kinds of schools would demonstrate sensitivity differently, the indicators of sensitivity used in the elementary and secondary CLASS protocols differ. In essence, the CLASS system assumes heterotypic continuity (Costello, Mustillo, Erkanli, Keeler, & Angold, 2003), a concept taken from developmental science in which processes remain stable over time although particular manifestations change.

Figure 1. CLASS-S: Domains and Dimensions

Domain	Dimension	Dimension Description
Emotional Support	Positive Climate	reflects the emotional connection and relationships among teachers and students, and the warmth, respect, and enjoyment communicated by verbal and nonverbal interactions

	Teacher Sensitivity	reflects the teacher's responsiveness to the academic and social/emotional needs and developmental levels of individual students and the entire class, and the way these factors impact students' classroom experiences
	Regard for Adolescent Perspectives	focuses on the extent to which the teacher is able to meet and capitalize on the social and developmental needs and goals of adolescents by providing opportunities for student autonomy and leadership; also considered are the extent to which student ideas and opinions are valued and content is made useful and relevant to adolescents
Classroom Organization	Negative Climate	reflects the overall level of negativity among teachers and students in the class; frequency, quality, and intensity of teacher and student negativity are important to observe
	Behavior Management	encompasses the teacher's use of effective methods to encourage desirable behavior and prevent and redirect misbehavior
	Productivity	considers how well the teacher manages time and routines so that instructional time is maximized; captures the degree to which instructional time is effectively managed and down time for students is minimized; it is not a code about student engagement or about the quality of instruction or activities
Instructional Support	Instructional Learning Formats	focuses on the ways in which the teacher maximizes student engagement in learning through clear presentation of material, active facilitation, and the provision of interesting and engaging lessons and materials
	Content Understanding	refers to both the depth of lesson content and the approaches used to help students comprehend the framework, key ideas, and procedures in an academic discipline; at a high level, refers to interactions among the teacher and students that lead to an integrated understanding of facts, skills, concepts, and principles
	Analysis and Problem Solving	assesses the degree to which the teacher facilitates students' use of higher level thinking skills, such as analysis, problem solving, reasoning, and creation through the application of knowledge and skills; opportunities for demonstrating metacognition (i.e., thinking about thinking), also included
	Quality of Feedback	assesses the degree to which feedback expands and extends learning and understanding and encourages student participation; in secondary classrooms, significant feedback may also be provided by peers; regardless of the source, focus here should be on the nature of the feedback provided and the extent to which it "pushes" learning

Evidence for the domain structure for CLASS-S is described in Hamre et al. (2013). Each dimension is scored on a scale of 1 to 7. Observers are trained with video anchors and elaborated descriptions of practice at the low (1, 2), middle (3, 4, 5), and high (6, 7) score bands. An example of the scale definition for one dimension appears in Appendix A.

In contrast to traditional observation protocols that focus on teacher actions, CLASS-S is representative of more recent evaluation protocols that focus on the actions and interactions of both teachers and students. Such protocols raise the question of whether scores not only reflect the teacher, but also the students in the class. For this reason, we refer to teaching quality rather than teacher quality (Bell et al., 2012) and report summary findings of observations in terms of classrooms instead of teachers.

Importantly, the domains also differ in the degree to which complex social interactions occur between teachers and students. Dimensions such as Behavior Management and Productivity (dimensions of Classroom Organization) largely capture very discrete actions by individual actors in the classroom. For example, did a student act out or did the teacher have to stop the class to correct behavior? Productivity can largely be quantified by time on task and time off task. The Emotional Support and Instructional Support domains and constituent dimensions are less directly observed, for they are intended to capture interactions between teachers and students and among students. The dimensions of Analysis and Problem Solving and Quality of Feedback, for example, not only include the nature of the teacher's questions, but also how the students respond and interact with the problems they are working on. Positive Climate requires judgment of the nature of the emotional tenor within a class.

In this study, lessons were divided into segments. Observers viewed a lesson for 15 minutes, recording evidence perceived as relevant to one or more dimensions. At the end of 15 minutes, observers assigned scores to each of the dimensions during a 7-minute scoring period. Thus, a segment lasted 22 minutes, meaning that a 45-minute lesson consisted of two segments, while longer lessons consisted of up to four segments. Each lesson was scored live and video-recorded for later scoring by observers who had not done the live scoring for that lesson.

Two types of analyses were conducted on these data—analyses of calibration data and analyses of the observational scoring. For analyses of the observational scoring, if two observers assigned scores, these scores were averaged to create a segment/dimension score. Otherwise, the single observer's score was used as the segment/dimension score. Segment/dimension scores were then averaged across segments to create a set of lesson/dimension scores. Finally, all lesson/dimension scores were averaged across lessons to create a classroom/dimension score.

At each level—segment/lesson/classroom—domain scores were also created, averaging across the respective dimensions described in Figure 1. Given the domain structure of CLASS-S, results focus on the domain level, although dimension scores are also reported. Internal reliabilities (Cronbach's alpha) of domain scores are reported in Table 1 at the segment, lesson, and classroom levels. Scores were highly related within each domain, and reliabilities increased as the data were aggregated. Domain scores were more reliable than dimension scores, and classroom scores were more reliable than single lesson scores.

Table 1. CLASS-S: Cronbach's Alpha by Domain, Live Scoring, All Levels

	Level		
Domain	**Segment**	**Lesson**	**Classroom**
Emotional Support	.83	.86	.90
Classroom Organization	.75	.78	.86
Instructional Support	.86	.89	.91

SELF-REPORT INSTRUMENT: *CLASSROOM ASSESSMENT SCORING SYSTEM FOR TEACHERS* (CLASS-T)

The CLASS-Teacher (CLASS-T) (excerpted in Appendix B), also developed at the University of Virginia, is a self-report questionnaire that asks teachers to provide an assessment of their skills on dimensions of CLASS-S. Teachers are asked to rate their skills in each dimension on a 5-point Likert scale ranging from 1 (*Area for Much Growth. This is an area with which I very often struggle.*) to 5 (*Area of Great Strength. This is an area in which I think I do very well.*). CLASS-T also asks teachers to identify the most and least important dimensions for teaching. Each participating teacher was asked to complete the instrument one time, and the administration was distributed so that one quarter of the teachers completed this survey during each quarter of the school year.

DESCRIPTION OF PARTICIPATING TEACHERS AND SCHOOLS

The study included 82 Algebra I teachers in 20 middle schools and 20 high schools in a large urban-fringe district. Participating teachers, all of whom volunteered to participate, were comparable to the entire pool of Algebra I teachers both in terms of teacher demographics and in the achievement of students they taught. For each teacher, a single section (e.g., Period 4) was studied over the course of the school year. Demographics of study teachers are presented in Table 2.

Table 2. Demographics of Study Teachers

Ethnicity	**%**		
African American or Black	50		
Asian or Asian American	26		
Caucasian or White	20		
Hispanic or Latino	1		
Other[a]	4		
Highest Education Level	**%**		
BA/BS degree	48		
MA/MS degree	42		
More than one degree	10		
EdD./PhD.	1		
Years of Full-time Teaching Experience	**K-12 (%)**	**Mathematics (%)**	**Algebra (%)**
0–1 year	10	10	21
2–4 years	12	12	21
5–10 years	33	38	28
11–20 years	23	22	18
> 20 years	17	13	7
No response	5	5	5
Certificate Type	**%**		
Alternate route status	8		
Standard professional certificates	33		
Advanced professional certificate	52		
Other/no response	6		

[a] Other ethnicities include Indian, German, Pakistani, and Jamaican.

OBSERVERS AND TRAINING

Six observers, all former secondary public school teachers, were originally part of the study. However, one observer left very early in the study, leaving the project with five observers who completed the vast majority of live observations and all of the video observations. The observers underwent extensive CLASS-S training that included a certification test, weekly calibration tests, and conference calls to discuss calibration results.

All observers had been secondary public school teachers for at least two years. Three taught in the state's public schools for at least some of their teaching experience, and one worked as a professional developer in the state's schools. Two of the five

previously taught in the study's school district. Four of the observers had experience teaching secondary mathematics, and the fifth observer taught English language arts.

Once scoring began, weekly calibration sessions were held. The calibration process was designed to ensure observers assigned scores that accurately reflected the quality of interactions seen in the video. To do this, videos that had been scored live as a part of the study were selected for calibration and scored by three "master coders" who were members of the CLASS-S development team at the University of Virginia. These individuals were selected by authors of the measure, were highly experienced with CLASS-S, typically led training and scoring efforts for CLASS-S, and were designated as master coders for the project. The CLASS-S authors reported very high agreement of the master coders with measure authors (over 90% within 1 scale point). For each calibration video, at least two master coders rated all of the segments of the lesson and then reconciled any discrepancies by discussing their codes and resolving any disagreements. These reconciled codes constituted master codes and are the project's best attempt at developing a surrogate for *true scores*. Master codes for segments were aggregated in the same way as in operational scoring to create master lesson scores.

In the next step of calibration, observers coded the same video lesson the master coders did. After submitting codes for review, observers were given the master codes and participated in a telephone discussion led by one of the project's principal investigators. Weekly calibration discussions focused on the discrepancies between the master codes and observer codes for each video and worked to clarify observer understandings of the CLASS-S scoring procedures and rubrics. These weekly calibration discussions continued throughout the project until all the lessons had been scored according to the scoring design. Although there was some small variability, observers participated in approximately 30 calibrations each, including approximately 60 segments and 30 lessons. Observational scores on the calibration videos collected before the telephone discussion serve as the basis of our assessment of rater agreement.

OBSERVATIONS

Each classroom was observed 4–5 times over the course of the 2009–2010 school year. Every observation was coded and scored *live* (i.e., during the observation). In addition, all observations were *video* recorded. All videos were coded and scored by two independent observers, while 36% of the live observations were double coded.

Scoring was designed so that observers were evenly distributed across classrooms and no observer scored the same lesson in both live and video modes. Having multiple observers across the lessons mitigated scoring errors attributable to differences in the severity with which particular observers assigned scores. Elaborated detail on the scoring design is presented in Casabianca, McCaffrey, Gitomer, Bell, and Hamre (2013).

Casabianca et al. (2013) reported the results of comparing live and video observations from this study. Though results are highly correlated, the live scores are somewhat more reliable. Therefore, for parsimony of presentation, this study focuses on the results from the live observation data.

The one exception is that we consider results from the video scoring in the analysis of changes in scoring consistency over time. In live scoring, not only do observers become more experienced over time and regular calibrations, but the teaching itself may change over the school year. Therefore, changes in live scores inherently confound changes in scoring behavior with changes in teaching. However, the study's video scoring avoided this confound by scoring videos in an order that was not systematically correlated with the day of the school year. By separating out any day-of-year effects, we were able to directly estimate changes in observers' rating behavior.

RESULTS

To what degree does the classroom observation protocol produce scores that vary across different dimensions of teaching practice in algebra classrooms? In this analysis, we consider scores in light of the descriptions of teaching articulated in the CLASS-S protocol. These descriptions give meaning to the nature of effective teaching on each of the respective dimensions, and these descriptions are independent of the distribution of scores for any sample or population of classrooms. Thus, when scores are said to differ between one dimension or domain and another, the reference is always to the protocol's respective description and scale only. No other kinds of comparisons of relative quality are implied.

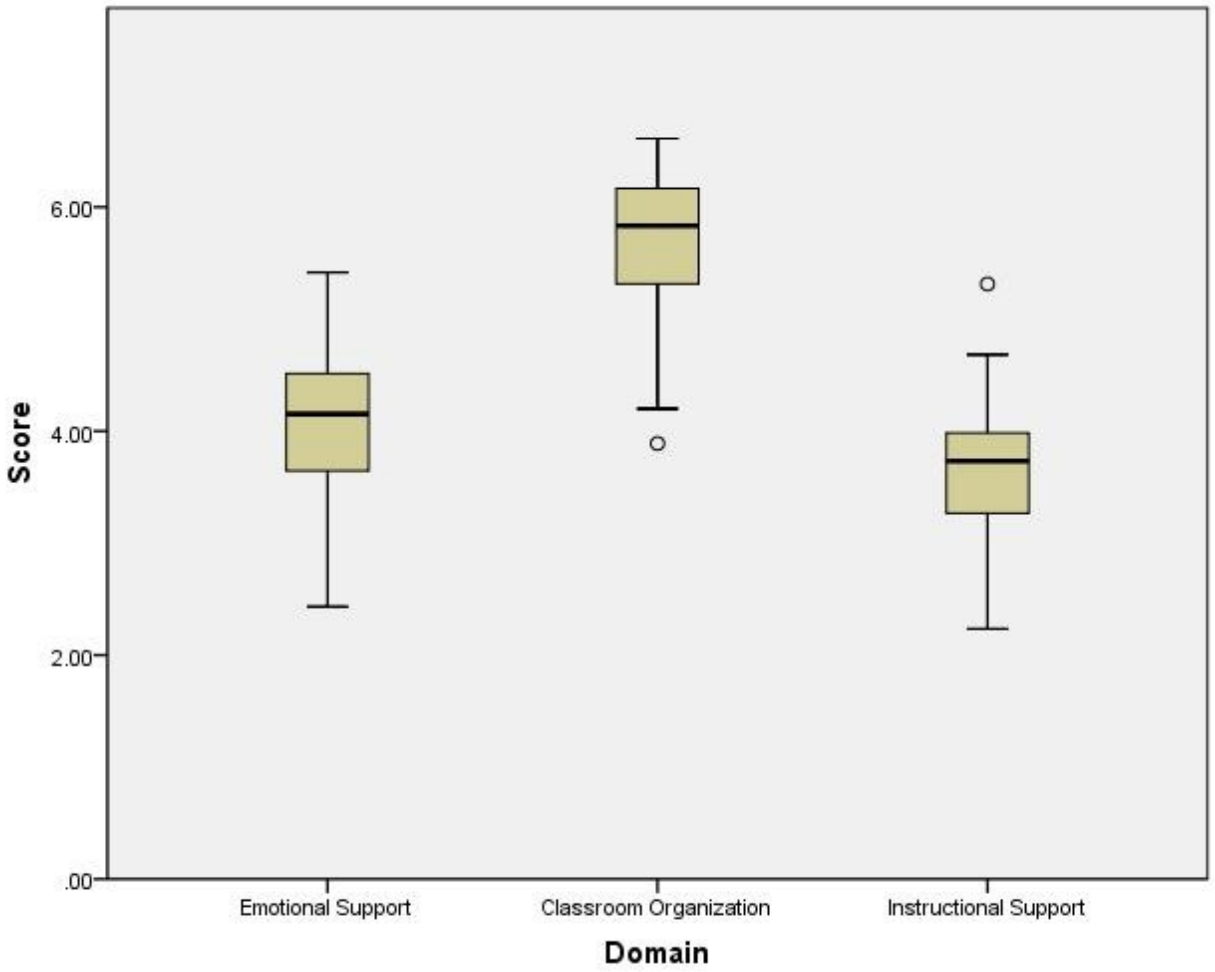
Mean lesson- and classroom-level dimension and domain scores and standard deviations are shown in Table 3. Not surprisingly, the variance is greater for lesson-level scores than for classroom-level scores. The Figure 2 boxplots display the distribution of domain scores at the classroom level. While mean scores vary across domains, the range of scores within domains is fairly narrow.

Table 3. CLASS-S: Descriptive Statistics by Domain and Dimension, Live Scoring, Lesson and Classroom Levels

	Lesson		Classroom	
	Mean	SD	Mean	SD
Domains				
Emotional Support				
Classroom Organization				
Instructional Support	4.00	0.95	4.05	0.65
	5.67	0.84	5.70	0.63
Dimensions	3.61	0.98	3.64	0.56

Positive Climate	4.32	1.14	4.39	0.87
Teacher Sensitivity	4.39	0.95	4.43	0.64
Regard for Adolescent Perspectives	3.29	1.13	3.33	0.61
Negative Climate (Inverted)	6.52	0.70	6.54	0.48
Behavior Management	5.17	1.22	5.21	0.95
Productivity	5.34	1.02	5.35	0.62
Instructional Learning Formats	4.00	0.89	4.03	0.56
Content Understanding	4.11	1.00	4.15	0.61
Analysis and Problem Solving	2.57	1.27	2.60	0.65
Quality of Feedback	3.75	1.34	3.79	0.70

Figure 2. Boxplots of CLASS-S domain scores at the classroom level. The circles represent observations outside the standard deviation range.



Mean scores for Classroom Organization are relatively high on the CLASS-S scale. Paired-sample t-tests were conducted to compare scores for Classroom Organization with scores for the other two domains. There were significant differences found in the t-tests: with Emotional Support, $t\ (81) = 34.21$, $p < .01$; and with Instructional Support, $t\ (81) = 38.88$, $p < .01$. Given that a score of 4 defines the middle of the scale, there are very few lessons that fall below the midpoint. Behavior Management in these classrooms is good, and students are productive in that they are engaged in the assigned instructional tasks with little downtime within and between activities. Very few instances of Negative Climate were observed. The vast majority of scores falls in the 5–7 range.

Mean scores for Emotional Support fell in the midlevel range. Support for the social and emotional needs of students as measured by CLASS-S is, on average, modest. The vast majority of dimension scores fall within the 3–5 range. The lowest scores, which came

in the Regard for Adolescent Perspectives dimension, reflect, in part, a lack of establishing a connection between algebra instruction and the lives of students.

Although scores on the Emotional Support domain were only modest, they were still higher on average than scores for Instructional Support. The paired-sample t-test between scores for Emotional Support and Instructional Support was significant, $t\ (81) = 10.50$, $p < .01$. Among the dimensions of Instructional Support, mean scores for the Instructional Learning Formats and Content Understanding dimensions are midlevel on the scale. The scores are lower on the Quality of Feedback dimension, which assesses if feedback will enhance student learning. The scores are lowest for the Analysis and Problem Solving dimension, suggesting that raters did not find evidence of the higher level reasoning skills described in the CLASS-S protocol. Scores on this dimension fall within the 2–4 range, the lower end of the CLASS-S scale.

In order to assess the degree to which experience might shape CLASS-S scores, we conducted a series of checks that examined the relationship of teaching experience and algebra teaching experience to the observation scores. We found no evidence of systematic relationships and, therefore, do not consider the effects of experience in the remainder of the paper.

How well are observers able to judge different dimensions of teaching practice? We evaluate observer agreement in a number of ways. The CLASS-S system treats agreement as having scores that are within one point of each other. Thus, if one observer assigns a score of 3 and another assigns a score of 4, the scores are judged to be in agreement. As described above, we assessed the accuracy of ratings by comparing observers' scores to the master codes. Comparing to master codes avoids the problem of finding high rates of agreement among observers because all observers are making similar errors. Agreement with the master codes means observational raters can produce scores of teaching as the CLASS-S instrument intends.

Table 4 displays agreement at the segment and lesson levels for CLASS-S domains and dimensions. For segment levels, we report exact agreement (the two raters give the same score) as well as adjacent agreement (the raters differ by at most one point) rates. We include the latter measure, given the calibration and certification procedures used in the CLASS-S system. Lesson-level agreement was calculated by rounding the scores averaged over segments to the nearest score point and then comparing the observers' and master coders' rounded scores. Agreement rates were then averaged across all observer-master coder pairs. The agreement criterion of being within one point for lesson-level scores averaged across segments is displayed in addition to the intraclass correlation (ICC), a correlation that summarizes the similarity of data that is organized into groups.

Table 4. Consistency of Observer and Master Codes

	% Agreement	Segment %1 Off Agreement	ICC	Lesson %1 Off Agreement	ICC
Domains					
Emotional Support	32.7	75.4	0.35	77.4	0.33
Classroom Organization	48.1	87.2	0.39	90.5	0.43
Instructional Support	29.3	73.5	0.33	75.5	0.29
Dimensions					
Positive Climate	30.2	70.3	0.25	73.4	0.23
Teacher Sensitivity	31.0	76.3	0.35	79.0	0.32
Regard for Adolescent Perspectives	36.9	79.7	0.44	79.8	0.44
Negative Climate (Inverted)	67.3	96.3	0.38	95.1	0.40
Behavior Management	37.9	92.1	0.49	96.5	0.51
Productivity	39.0	82.0	0.38	86.7	0.43
Instructional Learning Formats	29.7	78.3	0.33	83.9	0.36
Content Understanding	27.3	70.9	0.32	70.6	0.28
Analysis and Problem Solving	32.3	80.7	0.35	83.2	0.34
Quality of Feedback	28.0	69.0	0.31	72.7	0.27

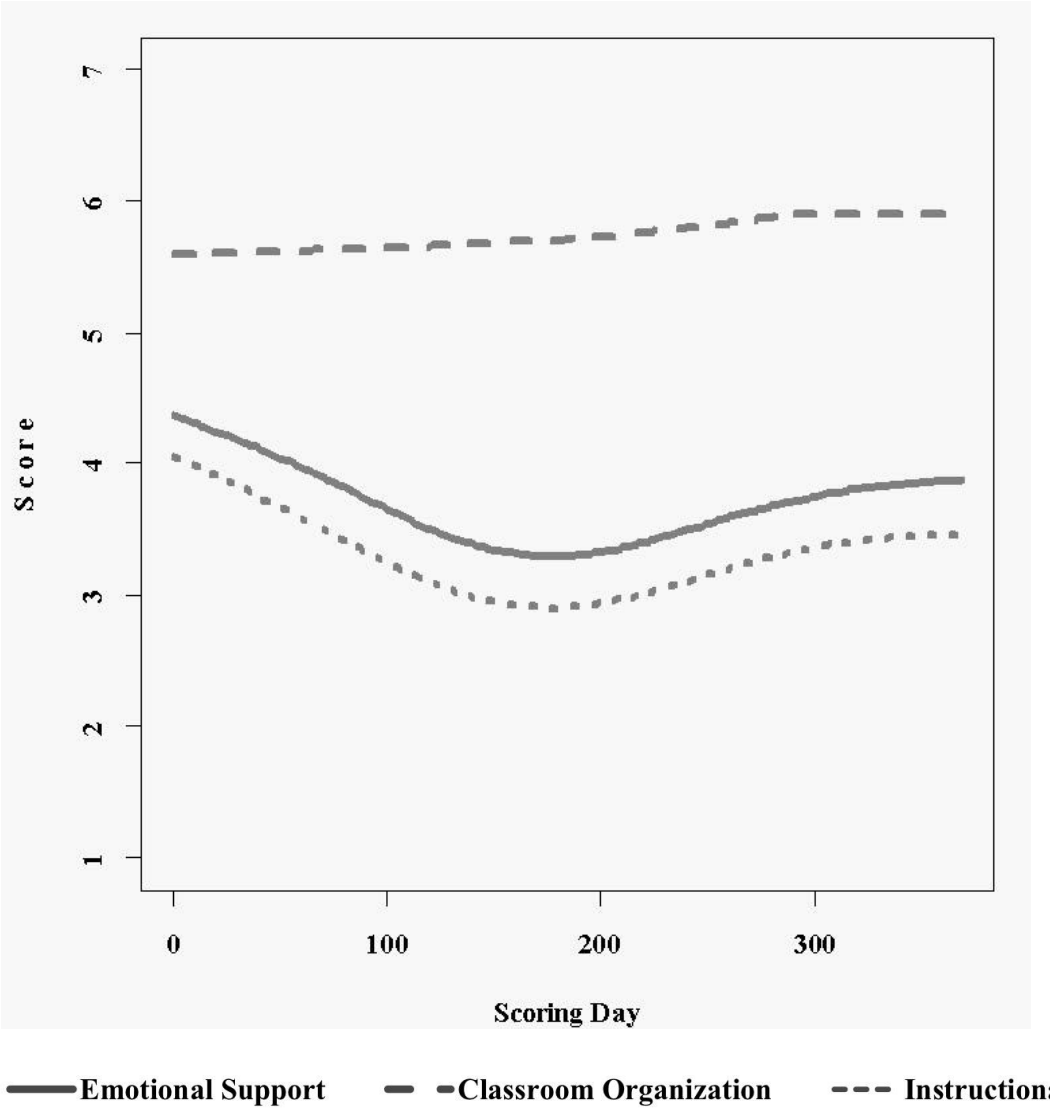
Agreement rates, using either agreement metric, are consistently higher for the Classroom Organization domain. Agreement rates are also slightly higher at the lesson level than at the segment level. Emotional Support and Instructional Support adjacent agreement rates are not significantly different from each other. The ICCs for all three domains are quite modest, however, suggesting that observers and master coders are not consistently rank ordering segments and lessons in similar ways. Given the high rates of adjacent agreement, the modest ICCs may reflect the relatively narrow distribution of scores for each domain. That is, while master coders and observers seem to be assigning scores within one point of each other, most especially for Classroom Organization, there is not highly consistent agreement in the ordering of individuals within a narrow range of scores.

The second way we examine rater judgment is to consider the stability of ratings over time. To what extent do observers' judgments change over time, presumably as they become more skilled through more experience as well as through weekly calibrations? To explore this question, Casabianca et al. (in press), as part of the TUCC project, examined changes in observer scores over the course of the school year.

Looking simply at the time trend of when in the school year the observation took place, Emotional Support and Instructional Support scores from both live and video observations decreased across days. Scores for Classroom Organization were fairly constant over time. Live scores showed a steeper decline than video scores. In fact, the decreasing trend for video scores by lesson date did not attain significance.

However, these trends conflate two possible sources of difference in scores: changes in observers' use of the rating scale over time and changes in the teaching and classroom interactions over time as the school year progresses. To separate trends due to observer changes rather than changes in teaching, Casabianca et al. (2013) plotted scores from videos versus the scoring date rather than the video capture date (Figure 3). Classroom Organization scores did not vary with the observers' experience in scoring. However, scores for the Emotional Support and Instructional Support domains declined during the first part of the scoring with a slight trend upward in scores on these domains at the end of the study (around 110 days, as shown in Figure 3). In models for video scores that included trends for both the day the lesson occurred and the day it was scored, trends in Classroom Organization were not significant for the day the lesson occurred nor the day it was scored. For Emotional Support and Instructional Support, the trends in the day the lesson was scored were significant, but trends in the day the lesson occurred were not. Hence, the visual trends in Figure 3 result from changes in how the raters were scoring the observations across the school year, not from real changes in teaching.

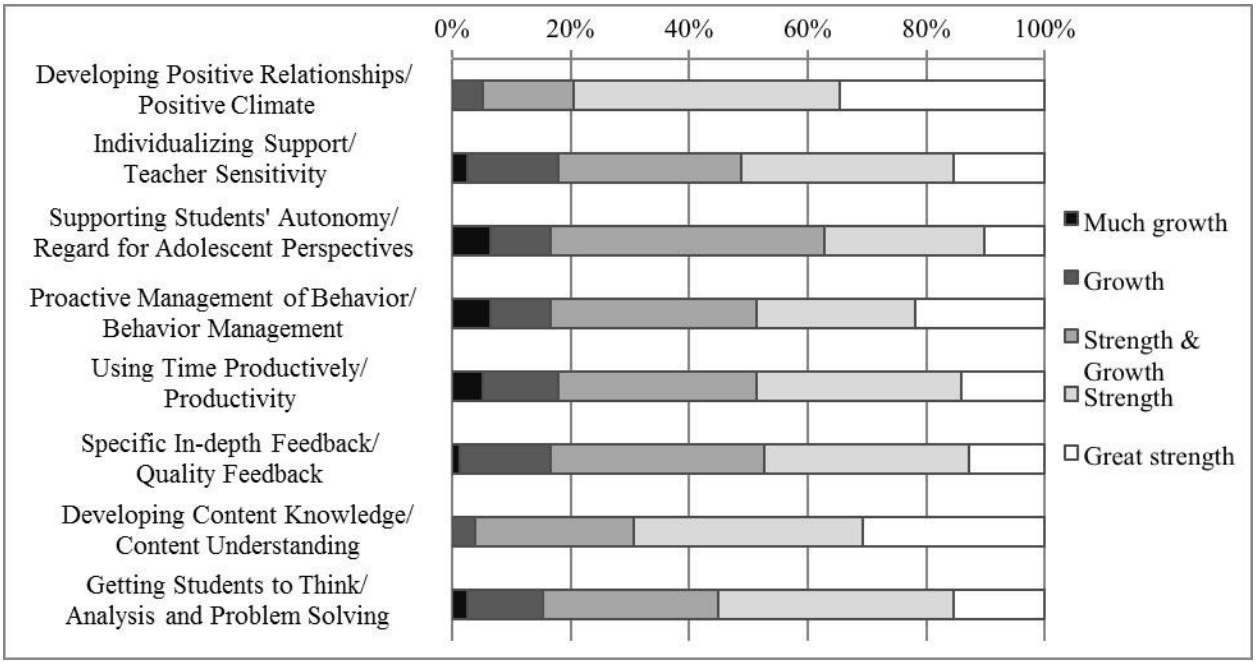
Figure 3. Time trends for video observations by date scored (Casabianca et al., in press)



Thus, there is evidence of changes in how teaching was scored on the Emotional Support and Instructional Support domains across the school year, with the highest levels in the first quarter of the year and the lowest levels in the second semester. Observers' use of the score scale is distinct from changes in teaching, with observers tending to score teaching lower as the school year progresses up to about the 180th day of scoring; the scores tended to move up again, although they remain below the initial values.

How well are teachers able to evaluate their own teaching practice across different dimensions? Teacher self-report data on the CLASS-T instrument are presented in Figure 4. The CLASS-T and their CLASS-S analogs are given. Proportions of responses in each of the five categories from *much growth* to *great strength* are given.

Figure 4. Distribution of ratings on each of the CLASS-T dimensions (with corresponding CLASS-S dimension)

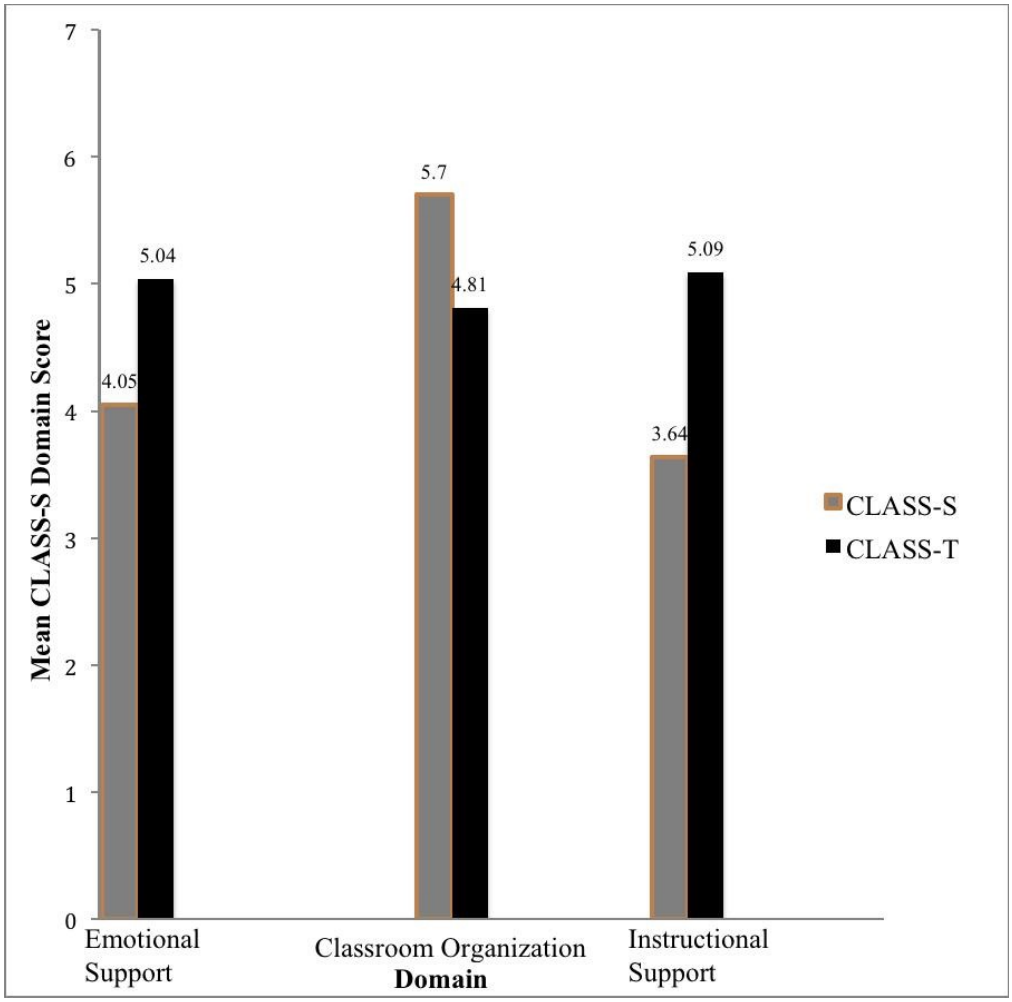


Interestingly, the patterns for self-reports are quite different from the CLASS-S observations. First, a one-way analysis of variance showed that there were differences in ratings among the eight CLASS-T dimensions, $F(7,616) = 6.96$, $p < .01$. A series of Bonferroni tests was used to compare each of the ratings. Teachers tended to rate themselves higher on Positive Climate and Content Understanding and somewhat lower and more similarly on the remaining dimensions. Positive Climate self-ratings were greater ($p < .01$) than those for all dimensions except Content Understanding. Content Understanding ratings were significantly higher ($p < .05$) than all of the other dimensions except Analysis and Problem Solving, Content Understanding, and Behavior Management (marginal $p < .06$). There were no significant differences across domains in teachers' average self-rating, in contrast with the CLASS-S findings where differences across domains were large.

In order to determine the degree to which teachers and external observers evaluated teaching practice similarly, we compared the scores they gave on the CLASS-T self-reports with the CLASS-S observation scores. To facilitate the comparison, we converted CLASS-T scores from a 5-point scale to a 7-point scale. Of course, because these measures are on different scales, the magnitude of scores on CLASS-S and CLASS-T is not directly comparable. The interactions between domains and measures are of interest. Therefore, we tested if the differences across domains identified by observers were similarly identified by teachers. We then compared individual teachers' self-reports to the scores raters gave, based on the live coding.

A repeated measures analysis of variance showed that there was a significant test by domain interaction, $F(1,77) = 15.32$, $p. < .01$. While mean scores on CLASS-T domains did not differ from each other, they varied substantially for CLASS-S. Given the use of different scales on the two instruments, it is the relative comparisons of domain scores by instrument rather than absolute scores that is of most interest. This interaction can be seen graphically in Figure 5. In relative terms, teachers seem to underestimate their abilities in Classroom Organization and overestimate their abilities on Emotional Support and Instructional Support.

Figure 5. Mean domain classroom-level scores for CLASS-S (observation) and CLASS-T adjusted scores (self-report)



Correlations provide another way of understanding the relationship between these two measures. Table 5 presents correlations for live scoring on CLASS-S with CLASS-T. The strongest correlation is for Classroom Organization—teachers whose classrooms are observed to be strong in Classroom Organization also tend to rate themselves as having strength in that domain. There is a significant, although lower, correlation for Emotional Support. What is of greatest interest is that there is no relationship between observations and self-reports in the Instructional Support domain. Teachers whose classrooms have strong observation scores rate themselves just as highly as teachers who have weak observation scores. Patterns are highly similar for both live and video-based scores.

Table 5. Pearson Correlations of Live CLASS-S Scores and CLASS-T Scores

CLASS-S	CLASS-T		
	Emotional Support	Classroom Organization	Instructional Support
Emotional Support	.28*	.27	-.08
Classroom Organization	.26*	.50**	-.02
Instructional Support	.26*	.23*	-.06

Notes. * $p < 0.05$. ** $p < 0.01$

DISCUSSION

In order for teacher evaluation to improve interactions in classrooms, at a minimum, protocols must be able to distinguish between different aspects of practice, and observers must be able to agree across time on how specific instantiations of teaching and learning should be scored by the protocol. This study identifies a critical challenge for teacher evaluation policy if it is to improve teaching and learning at scale. The challenge is that the aspects of teaching and learning that appear most in need of improvement on CLASS-S are those that are the hardest for observers to agree on and those that teachers external reviewers view most differently. Although this challenge is consistent with the patterns of observation scores and observer agreement statistics in a number of recent studies considering similar data (Bill and Melinda Gates Foundation, 2012), it is not insurmountable.

Consistent with studies of CLASS across grade levels, subject areas, and contexts (Hamre et al., 2013), we find evidence that the CLASS-S protocol can distinguish among different dimensions of classroom interactions. As others have reported on various observation protocols (Bill and Melinda Gates Foundation, 2012; Sartain, Stoelinga, & Brown, 2011), we find that scores are

generally lower for dimensions associated with the instructional aspects of classroom interactions, which are also ones that involve complex interactions among individuals in the classroom. The algebra classrooms that were part of this study were generally well managed, and there were relatively few instances of substantial misbehavior or Negative Climate. Further, students generally were on-task and doing what was asked of them.

On the other hand, scores for the Emotional Support domain were lower, largely because of low scores on the Regard for Adolescent Perspectives dimension. Given CLASS-S criteria, there was modest evidence of allowing for student autonomy or leadership or for making content relevant to the adolescents' own experiences. Students also had little opportunity to express ideas and opinions.

Instructional Support domain scores were lowest. According to the CLASS-S criteria, there was very little evidence that classrooms supported student development and used higher order thinking skills, including analysis, problem solving, reasoning, and creation through the application of knowledge and skills.

For all three domains and across dimensions, the striking result was the relative homogeneity of practice. Scores, by and large, tended to be distributed close to the mean score for a dimension, though means did vary substantially by dimension. Scores were highest for dimensions that could be more easily characterized by simple counts of discrete actions or attention to the timing of interactions in the classroom. We also found that scores were relatively consistent across the school year, once changes in scoring practices by observers were taken into account.

We found that well-trained and regularly calibrated observers had a more difficult time making judgments for the Emotional Support and Instructional Support domains than they did for the Classroom Organization domain. Agreement with master coders was highest for Classroom Organization. We also found that observers became more stable in their scoring more quickly for the Classroom Organization domain. There was a much more substantial learning curve for the other two domains, in which scores change substantially over time. Allen et al. (2013) observed similar patterns of rater agreement for CLASS-S. Though their study reported overall higher levels of agreement, the trends across domains were the same, suggesting that these patterns of agreement are robust across multiple studies.

Finally, we found that teachers' perceptions of the level of their practice were more aligned with those of external observers on the Classroom Organization domain. There was no relationship between the observed quality of Instructional Support in teachers' classrooms and their self-perception of their performance. There was a modest but significant relationship between perceptions and observed scores for the Emotional Support domain.

To summarize, the classroom interactions most in need of attention were the same ones that observers experienced the most difficulty scoring accurately and that teachers had the most difficulty evaluating in the same way external observers evaluated them. If this finding is true in other settings, we must consider how to work on this instructional challenge. What do we already know that can make it more likely that the implementation of teacher evaluation policy will be successful?

As a starting point, it is worth noting that recent experimental and observational work suggests that using observation protocols to improve teachers' practice has resulted in improvements in teaching. For example, Taylor and Tyler's (2011) recent findings from Cincinnati's teacher evaluation system suggest that teachers improve their teaching practices in anticipation of the year in which they are evaluated by external observers and continue to improve after that evaluation year.

Though researchers are only able to speculate on what mechanism causes this improvement, experimental data suggests that teachers can improve their observation scores when they are taught the observation protocol and work with a coach around short videos of their own teaching practice (Allen, Pianta, Gregory, Mikami, & Lun, 2011). In the *My Teaching Partner* program studied by Allen et al. (2011), teachers are taught the CLASS domains and dimensions. Over the course of the year, teachers go through multiple sharing and feedback cycles in which they share video clips from their classroom with a remote coach. After that video is scored, the teacher and coach, interacting almost exclusively online, discuss ways to improve specific aspects of the teacher's practice. Observations indicate that teachers make significant changes in their practice following a year of coaching, and these changes lead to improvements in student learning in the subsequent year (Allen et al., 2011).

If we look beyond the nascent literature that uses observation protocols as a primary way to improve practice, there is evidence that teachers can improve instructional aspects of their teaching, which, in turn, leads to improvements in student achievement (Yoon, Duncan, Lee, Scarloss, & Shapley, 2007). Though there is disagreement about what professional development must contain, effective professional development is generally viewed as content specific, intensive, and coherent, and it requires active learning and collective participation (Desimone, 2011). Many researchers have documented the challenges associated with professional development (e.g., Opfer & Pedder, 2011), and we should expect those lessons will apply to the implementation of teacher evaluation policies. Specifically, teaching will only change when teachers learn new things and have the support they need to successfully implement their new knowledge and skills. Thus, if past is prelude, teacher evaluation policy will need to pay careful attention to providing effective professional development that is aligned with teachers' specific learning needs.

In addition to suggesting the need to link specific areas of weakness to appropriate learning activities, this research suggests we need to think more carefully about the validity and interpretation of observation data in formal teacher evaluation systems. Except for one person, the observers in this study were former mathematics teachers; the other observer had experience using mathematics in other contexts. Observers were also highly trained and calibrated on a weekly basis, with ongoing feedback. Further, they conducted ratings on an almost full-time basis over the course of the project. Finally, these observers did not have relationships with the teachers in the study and, therefore, did not have to make judgments in the context of being a supervisor or

coworker. All of this suggests that the individuals who conducted observations in this study were operating in a supportive condition for providing high-quality ratings, a condition likely to be more supportive than existing contexts. For example, principals will likely have the greatest responsibility for teacher evaluation. Yet, the organizational demands on their time (Horng, Klasik, & Loeb, 2009) leave little space for the kind of evaluation processes that new systems demand. This need not mean that scores assigned by observers who are supervisors or coworkers have to be inferior to the ones assigned here. However, it does mean that teacher evaluation policies will have to pay careful attention to the training and calibration observers undergo.

Reliability of scores is not simply a psychometric hurdle. It is a marker of common understanding about important constructs, and what constitutes successful accomplishment with respect to those constructs. If such understanding does not exist, then it is difficult to imagine how pre-service and in-service teachers can effectively be supported. Observation protocols such as CLASS-S are intended to provide a common language and structure to inform teaching practice. However, if teachers, principals, administrators, and teacher educators hold idiosyncratic views of teaching practice, then any observation tool is likely to perpetuate the kinds of focus and discussions that already exist. In general, teaching does not have such shared understandings (e.g., Grossman & McDonald, 2008). Commonly held definitions of teaching quality are rare (Goe, Bell, & Little, 2008). Further, discussions grounded in actual artifacts of practice (e.g., clips of classroom interactions, conversations about student work) have historically been rare or limited to small subsets of professionals such as the elementary faculty in a single teacher education program or a grade-level team within a middle school. In order for teacher evaluation policy to be successful, it will have to be implemented in such a way that a common language and understanding of teaching is fostered. If shared understandings are then tied to the scoring criteria of the observation protocol, the second part of the instructional challenge can be addressed. Observers will be more likely to score reliably, and teachers will have views of their own instruction that are more consistent with those of external observers.

Of course, reliability can also be improved by constraining observation systems to those aspects of teaching that are most readily judged in a reliable manner. This research suggests that doing so might skew definitions of teaching in ways that would privilege those things that are easily observed and/or readily counted. However, doing this would also lead to observation systems that ignore critical features of teaching. While good classroom organization is a necessary condition for effective teaching, it is not sufficient to result in improved practice (Doyle, 1986). Effective teaching is characterized by the kinds of instructional interactions that are measured by dimensions in the Emotional Support and Instructional Support domains.

These findings also indicate that the historic tendency to give high and undifferentiated evaluations to teachers may not simply be a matter of institutional and administrative will, as suggested by recent reports and policy statements (e.g., Glazerman et al., 2010; Weisberg et al., 2009). If observers feel unsure of judgments about instruction and instead focus on aspects of classroom organization, then they are likely to legitimately evaluate teaching as being much more successful than they would if accurate judgments of the more instructional dimensions of teaching were also taken into account.

We are at a critical junction in the history of our profession. Never before has there been such strong federal intervention into many areas of education (Sykes & Dibner, 2009). Current teacher evaluation policies press us to be specific about the strengths and weaknesses in our classrooms. The implementation of the Common Core State Standards (National Governors Association [NGA] Center for Best Practices & Council of Chief State School Officers [CCSSO], 2010) provides a fresh opportunity to engage in a collective conversation about what cognitively complex student work looks like and how classroom interactions can make that work more successful for more students. The general weakness of teaching practice noted by Goodlad (1984) and confirmed by others can be identified and described using classroom observation methods. This study suggests the need to focus our efforts on the instructional and interactional aspects of classroom instruction through shared conversations, supported by clear images about what teaching quality looks like and how best to improve it through professional learning.

Acknowledgments

This research was generously supported by grants from the W. T Grant (9622) and Spencer Foundations (200900181). We are grateful for the reviews of Jim Carlson, Skip Livingston, and Danielle Guzman-Orth in addition to the anonymous reviewers for *Teachers College Record*. Finally, we thank the teachers, administrators, and staff of our partner district. Without their willingness to open their district to us, we would not be able to learn.

References

Allen, J. P., Gregory, A., Mikami, A. Y., Lun, J., Hamre, B. K., & Pianta, R. C. (2013). Observations of effective teaching in secondary school classrooms: Predicting student achievement with the CLASS-S. *Journal of School Psychology, 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*, 1034–1037. doi:10.1126/science.1207998

Bell, C. A., Gitomer, D. H., McCaffrey, D., Hamre, B., Pianta, R., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*, 62–87. doi:10.1080/10627197.2012.715014

Bill and Melinda Gates Foundation. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Washington, DC: Author. Retrieved from http://www.metproject.org/downloads /MET_Gathering_Feedback_Research_Paper.pdf

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an

adolescent transition: A longitudinal study and an intervention. *Child Development, 78*, 246–263. doi:10.1111/j.1467-8624.2007.00995.x

Burniske, J., & Meibaum, D. (2012). *The use of student perceptual data as a measure of teaching effectiveness* (Texas Comprehensive Center Briefing Paper). Austin, TX: Texas Comprehensive Center at SEDL. Retrieved from http://txcc.sedl.org/resources/briefs/number_8/bp_teacher_eval.pdf

Casabianca, J., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., & Hamre, B. K. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*(5), 757-783.

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis, 12*, 311–330. doi:10.3102/01623737012003311

Costello, E. J., Mustillo, S., Erkanli, A., Keeler, G., & Angold, A. (2003). Prevalence and development of psychiatric disorders in childhood and adolescence. *Archives of General Psychiatry, 60*, 837–844.

Desimone, L. (2011). R&D: A primer on effective professional development. *Phi Delta Kappan, 92*(6), 68–71.

Doyle, W. (1986). Classroom organization and management. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (4th ed.). New York, NY: Macmillan.

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D., & Whitehurst, G. (2011). *Passing muster: Evaluating teacher evaluation systems*. Washington, DC: Brown Center on Education Policy at Brookings. Retrieved from http://www.brookings.edu/~/media/research/files/reports/2011/4/26%20evaluating%20teachers/0426_evaluating_teachers

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added*. Washington, DC: Brown Center on Education Policy at Brookings. Retrieved from http://www.brookings.edu/~/media/research/files/reports/2010/11/17%20evaluating%20teachers/1117_evaluating_teachers.pdf

Goe, L., Bell, C. A., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://www.gtlcenter.org/sites/default/files/docs/EvaluatingTeachEffectiveness.pdf

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., & Brenwald, S. (2008). *Highlights from TIMSS 2007: Mathematics and science achievement of U.S. fourth- and eighth-grade students in an international context* (NCES 2009-001 Revised). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://nces.ed.gov/pubs2009/2009001.pdf

Goodlad, J. (1984). *A place called school: Prospects for the future*. New York, NY: McGraw-Hill.

Grossman, P., & McDonald, M. (2008). Back to the future: Directions for research in teaching and teacher education. *American Educational Research Journal, 45,* 184–205. doi:10.3102/0002831207312906

Hamre, B. K. (2008). *My areas of strength and growth*. Unpublished manuscript, University of Virginia, Charlottesville, VA.

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949–967. Retrieved from http://niwl.fhi360.org/events/uploads/UserFiles/File/Hamre_Pianta_paper.pdf

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S., . . . Hakigami, 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.

Hanushek, E. A., & Woessmann, L. (2011). How much do educational outcomes matter in OECD countries? *Economic Policy, 26*, 427–491. doi:10.1111/j.1468-0327.2011.00265.x

Horizon Research, Inc. (2000). *Inside the classroom observation and analytic protocol*. Chapel Hill, NC: Author. Retrieved from http://www.horizon-research.com/instruments/clas/cop.pdf

Horng, E. L., Klasik, D., & Loeb, S. (2009). *Principal time-use and school effectiveness* (CALDER Working Paper 34). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research. Retrieved from http://larrycuban.files.wordpress.com/2012/08/1001441-school-effectiveness.pdf

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The Classroom Assessment Scoring System: Findings from the prekindergarten year. *The Elementary School Journal, 104*, 409–426. Retrieved from http://www.jstor.org/stable/pdfplus/3202821.pdf?acceptTC=true

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC:

U.S. Department of Education. Retrieved from http://www.scribd.com/doc/49151492/A-Nation-at-Risk

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common core state standards for mathematics*. Washington, DC: Authors. Retrieved from http://www.corestandards.org/assets /CCSSI_Math%20Standards.pdf

Opfer, V. D., & Pedder, D. (2011). Conceptualizing teacher professional learning. *Review of Educational Research, 81,* 376–407. doi:10.3102/0034654311413609

Pianta, R. C., Hamre, B. K., Haynes, N. J., Mintz, S. L., & La Paro, K. M. (2009). *Classroom Assessment Scoring System (CLASS), secondary manual*. Charlottesville, VA: University of Virginia Center for Advanced Study of Teaching and Learning.

Pianta, R. C., Howes, C., Burchinal, M., Byrant, D., Clifford, R., Early, D., & Barbarin, O. (2005). Features of prekindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions? *Applied Developmental Science, 9*, 144–159. doi:10.1207/s1532480xads0903_2

Rhode Island Department of Education. (2011). *The Rhode Island model: Guide to evaluating building administrators and teachers*. Providence, RI: Author.

Rimm-Kaufman, S. E., La Paro, K. M., Downer, J. T., & Pianta, R. C. (2005). The contribution of classroom setting and quality of instruction to children's behavior in kindergarten classrooms. *The Elementary School Journal, 105,* 377–394. doi:10.1086/429948

Ripley, A. (2012, October). Why kids should grade teachers. *The Atlantic*. Retrieved from http://www.theatlantic.com/magazine /archive/2012/10/why-kids-should-grade-teachers/309088/

Rowan, B., Harrison, D. M., & Hayes, A. (2004). Using instructional logs to study mathematics curriculum and teaching in the early grades. *The Elementary School Journal, 105,* 103–127.

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation*. Chicago, IL: Consortium on Chicago School Research, University of Chicago Urban Education Institute. Retrieved from http://ccsr.uchicago.edu/sites/default/files/publications /Teacher%20Eval%20Report%20FINAL.pdf

Sherin, M. G., & Han, S. (2004). Teacher learning in the context of a video club. *Teaching and Teacher Education, 20,*163–183.

Sykes, G., & Dibner, K. (2009). *Fifty years of federal teacher policy: An appraisal*. Washington, DC: Center on Education Policy.

Taylor, E. S., & Tyler, J. H. (2011). *The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers* (NBER Working Paper 16877). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w16877.pdf?new_window=1

U.S. Department of Education. (2009). *Race to the Top program: Executive summary*. Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73,* 89–122.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New York, NY: The New Teacher Project. Retrieved from http://widgeteffect.org/downloads /TheWidgetEffect.pdf

Weiss, I. R., Pasley, J. D., Smith, P. S., Banilower, E. R., & Heck, D. J. (2003). *Looking inside the classroom: A study of K–12 mathematics and science education in the United States*. Chapel Hill, NC: Horizon Research, Inc. Retrieved from http://www.horizon-research.com/insidetheclassroom/reports/looking/complete.pdf

Yoon, K. S., Duncan, T., Lee, S. W-Y., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues & Answers Report, REL 2007–No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/southwest/pdf/REL_2007033.pdf

Appendix A. Excerpt From *Classroom Assessment Scoring System for Secondary Classrooms* (CLASS-S)

From *Classroom Assessment Scoring System (CLASS), secondary manual* (p. 21), by R. C. Pianta, B. K. Hamre, N. J. Haynes, S. L. Mintz, & K. M. La Paro, 2009, Charlottesville, VA: University of Virginia Center for Advanced Study of Teaching and Learning.

Positive Climate

Positive Climate reflects the emotional connection and relationships among teachers and students, and the warmth, respect, and enjoyment communicated by verbal and nonverbal interactions.

	Low (1, 2)	Mid (3, 4, 5)	High (6, 7)
Relationships • Physical proximity • Peer interactions • Shared positive affect • Interest in each others' lives • Social conversation	There are few, if any, indications that the teacher and students enjoy warm, supportive, and respectful relationships with one another.	There are some indications that the teacher and students enjoy warm, supportive, and respectful relationships with one another.	There are many indications that the teacher and students enjoy warm, supportive, and respectful relationships with one another.
Positive affect • Smiling • Laughter • Enthusiasm	There are few, if any, displays of positive affect among the teacher and students.	There are some displays of positive affect among the teacher and students; at other times these interactions are not evident.	There are frequent displays of positive affect among the teacher and students.
Positive communications • Positive comments • Positive expectations	There are rarely, if ever, positive communications among the teacher and students.	There are sometimes positive communications among the teacher and students; at other times these interactions are not evident.	There are frequent positive communications among the teacher and students.
Respect • Respectful language • Use of each others' names • Warm, calm voice • Listening to each other • Cooperation	The teacher and students rarely, if ever, demonstrate respect for one another.	The teacher and students sometimes demonstrate respect for one another.	The teacher and students consistently demonstrate respect for one another.

Appendix B. Excerpt From *Classroom Assessment Scoring System*—Teacher (Class-T)

From *My areas of strength and growth* (pp. 2–3), by B. K. Hamre, 2008, Charlottesville, VA: University of Virginia. Copyright 2008 by Bridget Hamre. Reprinted with permission.

1. **Using time productively**. Productive classrooms are like "well-oiled machines"—students in these classrooms know what they should be doing and always have something to do. In productive classrooms, teachers maximize instructional time throughout each class period. Teachers prepare for instructional activities in advance so that all materials are ready and accessible. In the face of inevitable distractions, such as someone entering the room or school announcements, teachers keep the students' focus on the activity at hand with quick redirections. Teachers minimize time spent on managerial tasks such as recording attendance or checking homework and put students in charge of some managerial tasks. Teachers transition smoothly from one activity to another.

- **Area for Much Growth. This is an area in which I very often struggle.**
- **Area for Growth. This is an area in which I most often struggle but occasionally feel I do well.**
- **Area of Strength and Growth. Sometimes, I do very well in this area. Other times, it is more of a struggle.**
- **Area of Strength. This is an area in which I think I do well most of the time but occasionally struggle.**
- **Area of Great Strength. This is an area in which I think I consistently do very well.**

2. **Getting students to think deeply and critically.** Teachers who help students think deeply and critically ask students to *reason* and *problem solve*. They ask many how and why questions. Teachers provide many opportunities for students to analyze and synthesize information. Teachers ask students to identify problems and generate multiple solutions to those problems. Teachers also help students learn to think by regularly modeling and encouraging students to "think out loud."

- **Area for Much Growth. This is an area in which I very often struggle.**
- **Area for Growth. This is an area in which I most often struggle but occasionally feel I do well.**
- **Area of Strength and Growth. Sometimes, I do very well in this area. Other times, it is more of a struggle.**
- **Area of Strength. This is an area in which I think I do well most of the time but occasionally struggle.**
- **Area of Great Strength. This is an area in which I think I consistently do very well.**

Cite This Article as: *Teachers College Record* Volume 116 Number 6, 2014, p. 1-32