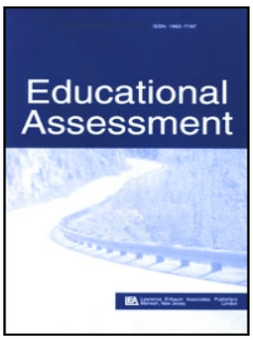

Educational Assessment

ISSN: 1062-7197 (Print) 1532-6977 (Online) Journal homepage: http://www.tandfonline.com/loi/heda20

# Attending to General and Mathematics-Specific Dimensions of Teaching: Exploring Factors Across Two Observation Instruments

David Blazar, David Braslow, Charalambos Y. Charalambous & Heather C. Hill

**To cite this article:** David Blazar, David Braslow, Charalambos Y. Charalambous & Heather C. Hill (2017) Attending to General and Mathematics-Specific Dimensions of Teaching: Exploring Factors Across Two Observation Instruments, Educational Assessment, 22:2, 71-94, DOI: 10.1080/10627197.2017.1309274

**To link to this article:** http://dx.doi.org/10.1080/10627197.2017.1309274

Accepted author version posted online: 22 Mar 2017.
Published online: 22 Mar 2017.

EDUCATIONAL ASSESSMENT
2017, VOL. 22, NO. 2, 71–94
http://dx.doi.org/10.1080/10627197.2017.1309274

Routledge
Taylor & Francis Group

# Attending to General and Mathematics-Specific Dimensions of Teaching: Exploring Factors Across Two Observation Instruments

David Blazar[a], David Braslow[a], Charalambos Y. Charalambous[b], and Heather C. Hill[a]

[a]Harvard Graduate School of Education; [b]University of Cyprus

**ABSTRACT**

New systems that seek to evaluate teachers with regard to their classroom quality often rely on observation instruments that capture general instructional pedagogies. However, decades of research suggest that content-specific dimensions of instruction also are important to differentiate teachers and improve student outcomes. We explore the degree of overlap between a general and a content-specific instrument when capturing upper elementary teachers' mathematics instruction. To do so, we conducted exploratory and confirmatory factor analyses on data from more than 2,000 videotaped lessons scored using both the Classroom Assessment Scoring System, a general instrument, and the Mathematical Quality of Instruction, a content-specific instrument. Findings indicate that there is some overlap between instruments but that preferred factor structures include both general and content-specific practices.



## Introduction

As schools and districts revamp their approaches to teacher evaluation, many are using observations of teaching practice as one metric of effectiveness (Center on Great Teachers and Leaders, 2013). Although evaluating teachers using on-the-job performance measures is not a new endeavor (Darling-Hammond, Wise, & Pease, 1983; Shavelson & Dempsey-Atwood, 1976), current approaches are meant to improve upon "cursory evaluations" (Hill & Grossman, 2013, p. 371) that did little to differentiate teachers with regard to performance standards (Weisberg, Sexton, Mulhern, & Keeling, 2009). Specifically, new systems rely on research-based observation rubrics and trained observers to ensure that these efforts meet stated goals.

Researchers who study these approaches to observation and evaluation highlight their potential to improve the quality of the teacher workforce by providing individualized feedback about teachers' instruction and matching them to appropriate development programs (Danielson & McGreal, 2000; Darling-Hammond, 2013; Hill & Grossman, 2013; Odden, 2004; Papay, 2012). At the same time, many of these same researchers raise concerns about logistical and practical constraints that schools face in implementing teacher observation and feedback. Hill and Grossman (2013) articulated three such challenges: (a) the use of general versus content-specific observation instruments, (b) the limited expertise of personnel and school leaders who serve as the primary observers, and (c) the capacity of schools to collect sufficient data on each teacher in a way that leads to robust and valid inferences about their effectiveness in the classroom.

In this article we explore the first of these challenges, one that forces districts and schools to consider their underlying assumptions about teaching and the nature of instructional guidance. For some, use of a general instrument stems from a theoretical perspective on teaching in which generic teaching skills transcend subject matter and discipline. Developers of one such instrument, the Framework for

**CONTACT** David Blazar dblazar@umd.edu University of Maryland, 2311 Benjamin Bldg., 3942 Campus Dr., College Park, MD 20742.

David Blazar is now affiliated with the University of Maryland.

Teaching (FFT), imply that subject specificity is not a concern, stating, "Teaching, in whatever context, requires the same basic tasks, namely, knowing one's subject, knowing one's students, having clear outcomes, establishing a culture for learning, engaging students in learning, etc." (Danielson Group, 2013b). This, in combination with practical challenges associated with the adoption of instruments and training of raters, has led many states and districts to rely on general instruments that are agnostic to content and grade-level differences across classrooms (Center on Great Teachers and Leaders, 2013).

However, the process of improving teaching requires targeted feedback tied directly to teachers' own strengths and weaknesses (Darling-Hammond, Wei, Andree, Richardson, & Orphanos, 2009; Garet, Porter, Desimone, Birman, & Yoon, 2001; Hill, 2007; Little, 2001; Wayne, Yoon, Zhu, Cronen, & Garet, 2008), many of which will be content specific (e.g., weak content or pedagogical content knowledge, difficulty responding to students in a way that gets at the root of their misunderstanding, lack of implementation of best practices in the content area). Improving content-specific teaching practices is particularly important given the relationship between these skills and students' academic performance (see, e.g., Blazar, 2015, for these relationships as they pertain to mathematics, and Grossman, Loeb, Cohen, & Wyckoff, 2013, as they pertain to English language arts), as well as students' social and emotional development including their self-efficacy in math (Blazar & Kraft, 2017). Thus, observations likely require content-specific protocols and observers adept at differentiating teachers and instruction in this way.

The ways in which these perspectives and approaches to the development of observation instruments play out in practice is not trivial. Two general instruments, the Classroom Assessment Scoring System (CLASS) and the FFT, ask broadly about teachers' "content understanding" (Pianta, Hamre, & Mintz, 2010) or "knowledge of content and pedagogy" (Danielson Group, 2013a). Comparatively, the Mathematical Quality of Instruction (MQI) articulates more than 10 subject-specific competencies (e.g., linking between representations, providing mathematical explanations, exploring patterns and generalizations) and provides examples on how observers should score each. Hill and Grossman (2013) emphasized that, even when competencies listed on general and content-specific instruments appear similar, they may serve different functions when applied in context. Take, for example, the practice of providing feedback to students. From a content-independent perspective, this feedback needs to be timely and descriptive, should elicit any clarification of students' thinking, and should outline the next steps that students need to undertake in order to improve their work (Stiggins & Chappuis, 2012). From a mathematics-specific perspective, this feedback likely requires additional features, including making public the most common student misconceptions and offering conceptual rather than purely procedural remediation for incorrect responses. When misconceptions are not present, a teacher offering content-specific feedback may choose to ask whether the students' strategy for solving a particular problem was the most efficient, or whether alternative strategies exist (Ball, 1988; Lampert, 2001; National Council of Teachers of Mathematics, 2014). In turn, the instructional guidance given to a teacher could differ quite dramatically based on use of a general versus content-specific instrument.

In this article we explore the degree of overlap between a general and a content-specific instrument and ask, *To what extent do a generic and a content-specific instrument overlap in capturing instructional quality in upper-elementary mathematics classes?* To answer this question, we used data from fourth- and fifth-grade teachers from five school districts, where teachers each had scores from two observation instruments, the CLASS, a general instrument, and the MQI, a content-specific instrument. Through correlational as well as exploratory and confirmatory factor analyses, we examined the relationship between instructional quality scores captured by these two instruments. We use the results of these analyses to discuss the practical tradeoffs when implementing evaluation and instructional feedback systems in our conclusion.

## Background

Many who study teaching view it as a complex craft made up of multiple dimensions and competencies (e.g., Cohen, 2011; Lampert, 2001; Leinhardt, 1993). In particular, older (Brophy, 1986) and more recent (Grossman & McDonald, 2008; Hamre et al., 2013) work calls on researchers,

practitioners, and policymakers to consider both general and content-specific elements of instruction. General classroom pedagogy often includes eliciting student thinking through effective questioning, giving timely and relevant feedback to students, and maintaining a positive classroom climate (e.g., Pianta, Belsky, Vandergrift, Houts, & Morrison, 2008; Pianta & Hamre, 2009). Content-specific elements, particularly in mathematics, include ensuring the accuracy of the content taught, providing opportunities for students to think and reason about the content, and using evidence-based best practices (e.g., linking between representations or using multiple solution strategies in mathematics; Ball, Thames, & Phelps, 2008; Lampert, 2001; National Council of Teachers of Mathematics, 1989, 1991, 2000).

However, research studies rarely integrate these views of teaching in practice. Most recent studies of teaching quality largely draw on just one observation instrument, either general or content-specific (see, e.g., Grossman et al., 2013; Hafen et al., 2015; Hill et al., 2008; Kane, Taylor, Tyler, & Wooten, 2011; McCaffrey, Yuan, Savitsky, Lockwood, & Edelen, 2015; Pianta et al., 2008). This tendency might be attributed to practical considerations (e.g., lack of resources to employ more than one instrument) or deeper philosophical reasons as to what counts as quality teaching and how it can best be measured. The few studies that have employed both general and more content-specific instruments provide empirical evidence attesting to the importance of considering both types of instruments as a means of better capturing instructional quality.

Our review of the research literature identified three studies that examine both general and content-specific teaching practices concurrently. Two of these studies draw on data from the Measures of Effective Teaching (MET) project, which includes scores on multiple observation instruments from teachers across six urban school districts. Using a principal components analysis framework, Kane and Staiger (2012) found that items tended to cluster within instrument to form up to three principal components each: one that captured all competencies from a given instrument simultaneously, analogous to a single dimension for "good" teaching; a second that focused on classroom or time management; and a third that captured a specific competency highlighted by the individual instrument (e.g., teachers' ability to have students describe their thinking for the FFT, and classroom climate for the CLASS). Using the same data, McClellan, Donoghue, and Park (2013) examined overlap between general and content-specific observation instruments. Factor analyses indicated that instruments did not have the same common structure. In addition, factor structures of individual instruments were not sensitive to the presence of additional instruments, further suggesting independent constructs. Without much overlap between instruments, the authors identified as many as 12 unique factors. In the third study, Lockwood, Savitsky, and McCaffrey (2015) analyzed data from lessons scored on three observation instruments, two general (CLASS and FFT) and one content-specific (either the MQI for math lessons or the Protocol for Language Arts Teaching for English language arts lessons). Using Bayesian exploratory factor analysis to account for how dimensions were ordered, they found two distinct teaching constructs: one for teachers' instructional practice and the other for classroom management. Unlike in the MET study, items from different instruments did tend to cluster onto the same factor. Together, this work suggests that instruments that attend solely to general or content-specific aspects of instruction may miss other important elements of teaching.

At the same time, these findings point to a challenge often associated with looking for factors across instruments: the existence of instrument-specific variation. Due to differences in the design and implementation of each instrument—such as the number of score points, the construction of dimensions, or the pool of raters—scores will tend to cluster more strongly within instruments than across them (Crocker & Algina, 2008). Therefore, distinctions found between teaching constructs across instruments may be driven by measurement artifacts (D. T. Campbell & Fiske, 1959).

To our knowledge, no research has explored the measurement artifacts associated with use of multiple observation instruments in the context of teaching quality. However, a handful of studies examining factor structures from a single instrument have found that more complex modeling strategies that account for construct-irrelevant sources of variation often lead to different solutions than factor structures from simpler models. For example, using confirmatory factor analysis (CFA),

McCaffrey et al. (2015) found that hierarchical models that accounted for rater errors led to a more parsimonious structure of the CLASS instrument than models in which teacher-level scores were conflated with rater error. When using CFA to specify bifactor models, Hamre, Hatfield, Pianta, and Jamil (2014) also found a more parsimonious structure for the CLASS instrument than was demonstrated in prior work. In a study using multiple observation instruments, Lockwood et al. (2015) accounted for construct-irrelevant variance due to segments, lessons, and raters, finding a similar number of factors as the studies just mentioned; however, their work did not account for instrument-specific variance.

Additional studies outside of the teaching quality literature suggest that bifactor CFA offers a way to separate out the variation introduced by different methods or instruments, including those that seek to capture respondents' personality traits and academic achievement/aptitude (Chen, Hayes, Carver, Laurenceau, & Zhang, 2012; Gustafsson & Balke, 1993). We build on these approaches to examine overlap between teaching quality scores captured by the CLASS and MQI observation instruments.

## Research hypotheses

Our analyses were guided by hypotheses surrounding the intended structures of both the CLASS and MQI instruments. The CLASS includes 11 items split into three theoretically distinct domains (see Table 1 for a full list of items and descriptions): "Emotional Support" captures "teachers' abilities to support social and emotional functioning in the classroom"; "Classroom Organization" focuses on "classroom practices that contribute to students' self-regulatory abilities"; and "Instructional Support" takes a general view of the content, curriculum, and learning activities, and in particular "the ways in which teachers implement these to effectively support cognitive and academic development" (Pianta & Hamre, 2009, p. 113). A 12th item, Student Engagement, is separated into its own domain. Several factor analyses conducted by instrument developers support this structure (Bell, Gitomer, McCaffrey, Hamre, & Pianta, 2012; Hafen et al., 2015; Hamre et al., 2013). However, other studies highlight variation in the factor structure of the CLASS, particularly when specifying more complex models including bifactor or hierarchical CFA models (Hamre et al., 2014; Kane & Staiger, 2012; McCaffrey et al., 2015). Generally, this latter work points to two prominent factors related to teachers' positive management and routines, and their cognitive facilitation. In light of these mixed findings, some researchers (Sandilos, DiPerna, & Family Life Project Key Investigators, 2014) have called for continued examination of factor structures in the CLASS instrument.

The MQI includes 13 items that also have a theoretically driven design laid out by instrument developers (see Table 2 for a full list of items and descriptions): "Richness of the Mathematics" captures the "depth of the mathematics offered to students … focus[ing] either on the meaning of facts and procedures or on key mathematical practices"; "Working with Students and Mathematics" captures "whether teachers can understand and respond to students' mathematically substantive productions (utterances or written work) or mathematical errors"; "Student Participation in Meaning-Making and Reasoning" captures "evidence of students' involvement in cognitively activating classroom work"; and "Errors and Imprecisions" focuses on "teacher errors or imprecision of language and notation, uncorrected student errors, or the lack of clarity/precision in the teacher's presentation of the content" (MQI, 2014). To date, instrument developers have not published formal factor analyses.

When mapping the content of the CLASS to the MQI, we envisioned a spectrum of teaching practices, with some more focused on content than others. On one extreme is Classroom Organization from the CLASS, which is theoretically unrelated to the specific content delivered in the classroom (Brophy & Good, 1986; Muijs et al., 2014) and which we hypothesized would form a unique factor from any of the instructional components captured by either the CLASS or the MQI. In prior work, this factor was found to be distinct from other instructional components of the CLASS, even when using either simpler or more complex factor structures than the typical three-factor structure advanced by instrument developers (Hamre et al., 2013; Kane & Staiger, 2012; McCaffrey et al., 2015).

**Table 1.** Item Descriptions from the Classroom Assessment Scoring System (CLASS) Instrument.

| Items | Description |
|---|---|
| Classroom organization | |
| Negative climate | Negative climate reflects the overall level of negativity among teachers and students in the class. |
| Behavior management | Behavior management encompasses the teacher's use of effective methods to encourage desirable behavior and prevent and redirect misbehavior. |
| Productivity | Productivity considers how well the teacher manages time and routines so that instructional time is maximized. This dimensions captures to degree to which instructional time is effectively managed and down time is minimized for students. |
| Emotional support | |
| 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. |
| Teacher sensitivity | Teacher sensitivity reflects the teacher's timely responsiveness to the academic, social/emotional, behavioral, and developmental needs of individual students and the entire class. |
| Respect for student perspectives | Regard for student perspectives captures the degree to which the teacher's interactions with students and classroom activities place an emphasis on students' interests and ideas and encourage student responsibility and autonomy. Also considered is the extent to which the content is made useful and relevant to the students. |
| Instructional support | |
| Instructional learning formats | 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 | 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, this refers to interactions among the teacher and students that lead to an integrated understanding of facts, skills, concepts, and principles. |
| Analysis and problem solving | 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, that is, thinking about thinking, are also included. |
| Quality of feedback | Quality of feedback assesses the degree to which feedback expands and extends learning and understanding and encourages student participation. Significant feedback may also be provided by peers. Regardless of the source, the focus here should be on the nature of the feedback provided and the extent to which it "pushes" learning. |
| Instructional dialogue | Instructional dialogue captures the purposeful use of dialogue—structured, cumulative questioning and discussion which guide and prompt students—to facilitate students' understanding of content and language development. The extent to which these dialogues are distributed across all students in the class and across the class period is important to this rating. |
| Student engagement | This scale is intended to capture the degree to which all students in the class are focused and participating in the learning activity presented and facilitated by the teacher. The difference between passive engagement and active engagement is of note in this rating. |

*Note*. Descriptions of items from Pianta, Hamre, and Mintz (2010).



On the other extreme is Instructional Support, which includes some items that have theoretical overlap with a content-specific instrument. In particular, the Content Understanding item from CLASS's Instructional Support dimension might align with multiple items and dimensions from the MQI. At face value, other items from Instructional Support, including Instructional Learning Formats, Analysis and Problem Solving, and Quality of Feedback, appear also to align with dimensions of the MQI that capture the quality of teachers' interactions with students and the extent to which these interactions develop intellectual challenge (Pianta & Hamre, 2009), namely, "Richness of the Mathematics," "Working with Students and Mathematics," and "Student Participation in Meaning-Making and Reasoning." However, as described earlier by Hill and Grossman (2013), a content-specific instrument such as the MQI likely enables observers to capture subtle distinctions in teachers' interactions with students around the content. Thus, we hypothesized that the MQI would capture additional variability in teachers' mathematics instruction than could be captured by the CLASS instrument, leading to correlated but unique factors.

Our interpretation of the CLASS instrument suggests that the third dimension, Emotional Support, falls in the middle of the spectrum of general to content-specific teaching practices. We

**Table 2.** Item Descriptions from the Mathematical Quality of Instruction (MQI) Instrument.

| Items | Description |
|---|---|
| Richness of the mathematics | |
| Linking and connections | Linking and connections of mathematical representations, ideas, and procedures. |
| Explanations | Explanations that give meaning to ideas, procedures, steps, or solution methods. |
| Multiple methods | Multiple procedures or solution methods for a single problem. |
| Generalizations | Developing generalizations based on multiple examples. |
| Mathematical language | Mathematical language is dense and precise and is used fluently and consistently. |
| Working with students and mathematics | |
| Remediation | Remediation of student errors and difficulties addressed in a substantive manner. |
| Teacher uses student productions | Responding to student mathematical productions in instruction, such as appropriately identifying mathematical insight in specific student questions, comments, or work; building instruction on student ideas or methods. |
| Student participation in meaning-making and reasoning | |
| Student explanations | Student explanations that give meaning to ideas, procedures, steps, or solution methods. |
| SMQR | Student mathematical questioning and reasoning, such as posing mathematically motivated questions, offering mathematical claims or counterclaims. |
| ETCA | Task cognitive demand, such as drawing connections among different representations, concepts, or solution methods; identifying and explaining patterns. |
| Errors and imprecisions | |
| Major errors | Major mathematical errors, such as solving problems incorrectly, defining terms incorrectly, forgetting a key condition in a definition, equating two nonidentical mathematical terms. |
| Language imprecisions | Imprecision in language or notation, with regard to mathematical symbols and technical or general mathematical language. |
| Lack of clarity | Lack of clarity in teachers' launching of tasks or presentation of the content. |

*Note*. Descriptions of items from Mathematical Quality of Instruction (MQI; 2014). SMQR = Student Mathematical Questioning and Reasoning; ETCA = Enacted Task Cognitive Activation.



state this given that codes falling under this dimension, such as Teacher Sensitivity, require some content-specific awareness around teachers' ability to identify students with academic problems and support them accordingly, as well as more general support of students' emotional needs in the classroom (Pianta et al., 2010). Thus, similar to Instructional Support, we hypothesized that this dimension would be correlated with dimensions from the MQI that also capture teachers' interactions with students but would not cluster onto the same factor. This interpretation is consistent with some empirical work in which items from Instructional Support and Emotional Support form a single factor (Hamre et al., 2013; McCaffrey et al., 2015).

Finally, we hypothesized that items from the Errors and Imprecisions dimension from the MQI would not be correlated with or cluster onto the same factor as any items from the CLASS, with the possible exception of Content Understanding. We state this given the purely mathematical nature of these items from the MQI. They do not intend to capture the sorts of teacher supports for students that are the main focus of the CLASS (Pianta & Hamre, 2009).

## Methods

### *Data and participants*

Our sample consists of 390 fourth- and fifth-grade teachers from five school districts on the East Coast of the United States. Four of the districts were part of a large-scale project from the National Center for Teacher Effectiveness focused around the collection of observation scores and other teacher characteristics. Teachers from the fifth district participated in a separate randomized controlled trial of a mathematics professional development program that collected similar data on teachers as the first project. Both projects spanned the 2010–11 through the 2012–13 school years. In the first project, schools were recruited based on district referrals and size; the study required a minimum of two teachers in each school and sampled grade. Of eligible teachers in these schools, roughly 55% agreed to participate. Despite moderate participation rates, a comparison of teachers

who chose to participate in the project and those who did not suggests no differences in their value-added scores on state mathematics tests (Blazar, Litke, & Barmore, 2016). In the second study, we included only the treatment teachers for the first 2 years, as observation data were not collected for the control group teachers. We have video data on teachers in both groups in the third year.

Teachers' mathematics lessons ($N = 2{,}276$) were captured over a 3-year period, with a yearly average of three lessons per teacher for the first project and six lessons per teacher for the second project. Videos were recorded using a three-camera, unmanned unit; site coordinators turned the camera on prior to the lesson and off at its conclusion. Most lessons lasted between 45 and 60 min. Teachers were allowed to choose the dates for capture in advance and were directed to select typical lessons and exclude days on which students were taking a test. Although it is possible that these videotaped lessons were different from teachers' general instruction, teachers did not have any incentive to select lessons strategically as no rewards or sanctions were involved with data collection. In addition, analyses from the MET project indicate that teachers are ranked almost identically when they choose lessons to be observed compared to when lessons are chosen for them (Ho & Kane, 2013).

The research projects scored these lessons using both the CLASS and MQI instruments. Data for the CLASS and MQI were generated from separate scoring protocols based on decision rules set by instrument developers. For the CLASS, one rater watched each lesson and scored teachers' instruction on 12 items for each 15-min segment on a scale from 1 (*low*) to 7 (*high*). Raters were recommended to the research projects by instrument developers based on their work as raters in other studies. The project recruited additional raters who were undergraduates from local colleges. For the MQI, two raters watched each lesson and scored teachers' instruction on 13 items for each 7½-min segment on a scale from 1 (*low*) to 3 (*high*). Raters were recruited from a separate pool of applicants based on their background in mathematics. Project leaders posted notices on mathematics education listservs and sent e-mails to colleagues working in mathematics education departments (see Hill, Charalambous, Blazar, et al., 2012, for more information). For both instruments, raters had to complete an online training, pass a certification exam, and participate in ongoing calibration sessions. Raters were not provided any background information on teachers. One item from the CLASS (Negative Climate) and three from the MQI (Major Errors, Language Imprecisions, and Lack of Clarity) have a negative valence, which we maintained in this analysis.

For our primary analyses, we reduced these raw data to a teacher-level data set by averaging scores across raters (for the MQI), segments, and lessons (both instruments). Our primary reason for doing so was that both observation instruments have been used primarily to draw inferences about individual teachers (Hill, Charalambous, Blazar, et al., 2012; Hill, Charalambous, & Kraft, 2012; Kane & Staiger, 2012; McClellan et al., 2013; Pianta & Hamre, 2009). Later in this article, we describe sensitivity analyses that examined the robustness of findings to use of additional data sets that allowed us to account for sources of variation—that is, raters, segments, and lessons—that were masked by aggregating the data to this level.

We present descriptive statistics on these teacher-level scores in Table 3. For the CLASS instrument, many items have means around the middle of the 7-point scale and are roughly normally distributed. Some exceptions include Negative Climate, which has a strong left skew, and both Behavior Management and Productivity, which have moderate right skews. For the MQI instrument, means tend to sit below the middle of the 3-point scale. Some items including Explanations, Mathematical Language, Remediation, and Enacted Task Cognitive Activation are roughly normally distributed; others have a long right tail.

### Analysis strategy

We conducted three sets of analyses. We began by examining pairwise correlations of items across instruments. This allowed us to explore the degree of potential overlap in the dimensions of instruction captured by each instrument. Next, we conducted a set of exploratory factor analyses (EFA) to identify the number of factors we might expect to see, both within and across instruments. In running these analyses, we attempted to get parsimonious models that would explain as much of

**Table 3.** Descriptive Statistics of Teacher-Level Observation Scores.

| Items | M | SD | Maximum | Minimum |
|---|---|---|---|---|
| CLASS | | | | |
| Negative climate | 1.23 | 0.31 | 4.00 | 1.00 |
| Behavior management | 6.02 | 0.62 | 7.00 | 2.73 |
| Productivity | 6.30 | 0.45 | 7.00 | 3.93 |
| Positive climate | 4.56 | 0.69 | 6.33 | 2.50 |
| Teacher sensitivity | 4.56 | 0.51 | 6.00 | 2.77 |
| Respect for student perspectives | 3.41 | 0.58 | 4.93 | 1.68 |
| Instructional learning formats | 4.46 | 0.45 | 5.67 | 3.00 |
| Content understanding | 4.23 | 0.54 | 5.75 | 2.00 |
| Analysis and problem solving | 2.98 | 0.59 | 4.50 | 1.18 |
| Quality of feedback | 4.12 | 0.65 | 6.17 | 1.88 |
| Instructional dialogue | 3.81 | 0.62 | 5.53 | 1.93 |
| Student engagement | 5.19 | 0.50 | 7.00 | 3.50 |
| MQI | | | | |
| Linking and connections | 1.32 | 0.21 | 2.27 | 1.00 |
| Explanations | 1.27 | 0.15 | 1.88 | 1.00 |
| Multiple methods | 1.15 | 0.13 | 1.76 | 1.00 |
| Generalizations | 1.04 | 0.05 | 1.33 | 1.00 |
| Mathematical language | 1.47 | 0.20 | 2.25 | 1.00 |
| Remediation | 1.37 | 0.17 | 2.08 | 1.00 |
| Teacher uses student productions | 1.21 | 0.15 | 1.72 | 1.00 |
| Student explanations | 1.21 | 0.15 | 1.83 | 1.00 |
| SMQR | 1.21 | 0.14 | 1.78 | 1.00 |
| ETCA | 1.32 | 0.19 | 2.06 | 1.00 |
| Major errors | 1.08 | 0.09 | 1.86 | 1.00 |
| Language imprecisions | 1.17 | 0.12 | 1.69 | 1.00 |
| Lack of clarity | 1.12 | 0.12 | 2.14 | 1.00 |

*Note*. Classroom Assessment Scoring System (CLASS) items are on a 7-point scale, and Mathematical Quality of Instruction (MQI) items are on a 3-point scale. SMQR = Student Mathematical Questioning and Reasoning; ETCA = Enacted Task Cognitive Activation.



the variation in the assigned teaching quality ratings with as few factors as possible. We opted for oblique rotations (i.e., direct oblimin rotation), which allow the extracted factors to be correlated. We did so given theory (Hill, 2010; Hill, Kapitula, & Umland, 2011; Pianta & Hamre, 2009) and empirical findings (Hill et al., 2008; Pianta et al., 2008) suggesting that the different constructs within each instrument are intercorrelated.[1]

Although we conducted this EFA to look for cross-instrument factors, prior research suggests that we would not expect to see much overlap across instruments because of the substantial variation attributable to each instrument (McClellan et al., 2013). Therefore, we used confirmatory factor analyses (CFA) to account for construct-irrelevant variation caused by the use of two instruments.[2] In particular, we utilized bifactor models (Chen et al., 2012) to extract instrument-specific variation and then compared the fit of various factor structures that allowed items to cluster across instruments. In these models, all items were specified to load on one instrument factor (CLASS or MQI) and one instructional factor. To help support our interpretation of these models, we also compared results to nonbifactor CFA models.

[1]To ensure that the resulting factor solutions were not affected by the differences in the scales used across the two instruments (MQI uses a 3-point scale, whereas CLASS employs a 7-point scale), we ran the analyses twice, first with the original instrument scales and a second time collapsing the CLASS scores into a 3-point scale (1–2 = low, 3–5 = mid, 6–7 = high) that aligns with the developers' use of the instrument (see Pianta & Hamre, 2009). Because there were no notable differences in the factor solutions obtained from these analyses, in what follows we report on the results of the first round of analyses, in which we used the original scales for each instrument.

[2]Generally, CFA involves using null hypothesis significance testing to examine whether there is evidence to support a theorized model. In some instances, CFA is used beyond a strictly confirmatory approach (i.e., testing only the theorized models at hand); it also can be used for model generation purposes, namely to generate different models, which satisfy three conditions: (a) they make theoretical sense, (b) they are reasonably parsimonious, and (c) their correspondence to the data is "acceptably close" (R. B. Kline, 2011, p. 8). Like others who have used CFA in this model-building fashion (see, e.g., Fabrigar, Wegener, MacCallum, & Strahan, 1999), we looked at incremental improvements in fit to evaluate different instructional factor structures.

Three constraints—both methodological and substantive—guided our construction of the bifactor CFA models. First, we could not allow all items from the same instrument to cluster together to form one instructional factor, because the instrument factor would be identical to the instructional factor. Second, for the CLASS and MQI instrument factors to extract instrument-specific variation, it was necessary to have at least one instructional factor with items from both the CLASS and MQI. This cross-instrument instructional factor ensures that shared variance caused by characteristics of instruction is explained by the instructional factors, whereas the instrument factors explain other sources of shared variation uncorrelated with instruction (for more information on this approach as it pertains to the multitrait multimethod analysis, see Pohl & Steyer, 2010). Inclusion of a cross-instrument instructional factor also aligns with our main purpose for specifying bifactor models: to examine the degree of overlap of items from the CLASS and MQI instruments, after accounting for variance attributed to using different instruments. Last, the instrument factors were constrained to be uncorrelated with the instructional factors to ensure that the variation captured by the instrument factors did not include variation attributable to the instructional characteristics being modeled.

One concern when attempting to account for instrument-specific variation is that these factors may, in fact, capture some variation in teaching ability in addition to the noninstructional method or instrument variance they were designed to capture. As such, our bifactor CFA approach may introduce additional instructional factors that are more than pure nuisance. At the same time, introducing these factors allows for comparisons of fit among models with cross-instrument factor structures of theoretical interest by accounting for the instrument-specific variance—both instructional and noninstructional—that is uncorrelated with the instructional factors of interest. We interpret models next in light of this challenge.

Some researchers have raised concern about additional sources of construct-irrelevant variation that can influence observed factor structures (McCaffrey et al., 2015). There are five such sources of variation in our data set: raters, segments, lessons, teachers, and instruments. Although Savitsky and McCaffrey (2014) were able to use Bayesian methods to simultaneously model variation caused by raters, segments, observations, and teachers, we were unable to do so in our analysis for two reasons. First, the scoring design differed across the two instruments, with different pools and numbers of raters used for each instrument, hindering our ability to model rater effects along with other sources of variation. Second, the bifactor CFA models described next already are computationally challenging when accounting for two sources of variation (instrument and teacher). Models including additional sources of random variation often failed to converge.

Despite our inability to simultaneously model all possible sources of variation, we attempted to understand the extent to which this limitation might affect our results with three supplementary analyses. First, we fit models in which teacher-level scores were adjusted for rater severity. Here, a differential was calculated for each rating based on how far it was from the average of the segment ratings for that teacher. Severity scores were calculated for each rater by averaging these differentials. We also fit two nonbifactor multilevel CFA models: one with segments nested within teachers, and another with lessons nested within teachers.[3] (CFA models simultaneously accounting for all three facets did not converge.) These additional analyses (available upon request) produced factor structures identical to that from our teacher-level analysis, so we present results only from this latter analysis.

[3]We note two important differences between instruments at the segment level. First, whereas the MQI has two raters score instruction, the CLASS has only one. Therefore, for the MQI, we averaged scores across raters within a given segment to match the structure of the CLASS. Second, whereas the MQI has raters provide scores for each 7½-min segment, the CLASS instrument has raters do so every 15 min. Therefore, to match scores at the segment level, we assigned CLASS scores for each 15-min segment to the two corresponding 7½-min segments for the MQI.

## Results

Our correlation matrix shows that some items on the CLASS and MQI are moderately correlated at the teacher level (see Table 4). For example, both Analysis and Problem Solving and Instructional Dialogue from CLASS are correlated with multiple items from the MQI (Mathematical Language, Teacher Uses Student Productions, Student Explanations, Student Mathematical Questioning and Reasoning, and Enacted Task Cognitive Activation) above 0.30. Three items from the MQI—Mathematical Language, Teacher Uses Student Productions, and Student Mathematical Questioning and Reasoning—are correlated with multiple items from the CLASS at similar magnitudes. The largest observed cross-instrument correlation of 0.41 is between Analysis and Problem Solving and Teacher Uses Student Productions. Even though we ran 156 separate tests, the 104 statistically significant correlations are much higher than the 5% we would expect to see by chance alone. These findings suggest that items from the two instruments may capture somewhat similar facets of instruction. Therefore, factor structures might include factors with loadings across instruments.

At the same time, there do appear to be distinct elements of instruction captured by each instrument. The three items under the Errors and Imprecisions dimension from the MQI, embedded deeply in a content-specific view of teaching, are not related to items from the CLASS. Five items from the CLASS—the three items under the Classroom Organization dimension, Negative Climate, Behavior Management, and Productivity, as well as Student Engagement and Positive Climate—correlate with items from the MQI no higher than 0.30. These findings suggest that these items might form distinct factors.

Next, we present results from the EFA in order to examine how these items cluster together to form instructional factors. At this stage in our analysis, we did not expect items from different instruments to load onto the same factor, as this analysis did not account for instrument-specific variation. Rather, results from the EFA inform our CFA models and help place an upper bound on the total number of factors we should model, both within and across instruments. First we note that the Kaiser-Mayer-Olkin value in all factor analyses exceeded guidelines for acceptable threshold of meritorious values (0.80), thus suggesting that the data lent themselves to forming groups of variables, namely, factors (Kaiser, 1974). Initial results point to six factors with eigenvalues above 1.0, a conventionally used guideline for selecting factors (P. Kline, 1994); scree-plot analysis also supports these six as unique factors (Hayton, Allen, & Scarpello, 2004). However, even after rotation, no item loads onto the sixth factor at or above 0.40, which is often taken as the minimum acceptable factor loading (Field, 2013; P. Kline, 1994). Two considerations guide our decision regarding which of the more parsimonious models best fit our data: the percent of variance explained by each additional factor, and the extent to which the factors have loadings that support substantive interpretations. We discard the five-factor solution given that it explains only 3% more of the variance in our data and therefore contributes only minimally to explaining the variance in the assigned teaching quality ratings (Field, 2013; Tabachnick & Fidell, 2001). In addition, items that load onto the fifth factor almost all cross-load onto other factors; generally these loadings are weak. We also exclude one- and two-factor solutions, as neither explains more than 50% of variation in our data, a guideline often used for accepting a factor structure (P. Kline, 1994).

In Tables 5 and 6, we present eigenvalues, percent of variance explained, and factor loadings for a parsimonious list of factors generated from the remaining three- and four-factor solutions. In both tables, cells are highlighted to identify instructional factors and potential cross-loadings (i.e., loadings on two factors of similar magnitude). In the three-factor solution—in which we name factors numerically (i.e., Factor 1, Factor 2, Factor 3)—23 of the 25 items load clearly onto only one factor. Of the remaining two items, Analysis and Problem Solving from the CLASS loads strongly onto the first factor and has a notable loading on the second factor; Mathematical Language from the MQI has loadings on both the first and second factors of similar magnitudes, though both fall below the acceptable guiding threshold of 0.40. Mathematical Language and Generalizations also have

**Table 4.** Item Correlations.

| Items | Negative Climate | Behavior Management | Productivity | Positive Climate | Teacher Sensitivity | Respect for Student Perspectives | Instructional Learning Formats | Content Understanding | Analysis and Problem Solving | Quality of Feedback | Instructional Dialogue | Student Engagement |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Linking and connections | −.104* | .103* | .155** | .099† | .170*** | .163** | .112* | .195*** | .308*** | .184*** | .247*** | .078 |
| Explanations | −.104* | .141** | .215*** | .132** | .293*** | .201*** | .190*** | .270*** | .338*** | .237*** | .264*** | .112* |
| Multiple methods | −.078 | .022 | .059 | .035 | .136** | .203*** | .056 | .062 | .313*** | .113* | .208*** | −.011 |
| Generalizations | −.103* | .114* | .199*** | .104* | .163** | .129* | .147** | .267*** | .165** | .241*** | .185*** | .089† |
| Mathematical language | −.223*** | .299*** | .299*** | .257*** | .281*** | .198*** | .215*** | .351*** | .295*** | .248*** | .308*** | .208*** |
| Remediation | −.007 | .067 | .169*** | .049 | .208*** | .140** | .104* | .176*** | .220*** | .218*** | .192*** | .051 |
| Teacher uses student productions | −.146** | .182*** | .203*** | .168*** | .314*** | .340*** | .287*** | .232*** | .410*** | .303*** | .393*** | .228*** |
| Student explanations | −.068 | .107* | .128* | .094† | .273*** | .257*** | .223*** | .175*** | .342*** | .221*** | .322*** | .148** |
| SMQR | −.090† | .134** | .159** | .150** | .249*** | .262*** | .186*** | .199*** | .340*** | .208*** | .305*** | .143** |
| ETCA | −.127* | .146** | .216*** | .147** | .305*** | .313*** | .268*** | .219*** | .324*** | .296*** | .321*** | .174*** |
| Major errors | .001 | .058 | .056 | .005 | −.049 | .037 | −.036 | .069 | .003 | .052 | .002 | −.023 |
| Language imprecisions | −.072 | .102* | .079 | .045 | .058 | .016 | −.029 | .071 | .087~ | .024 | −.002 | .030 |
| Lack of clarity | .027 | .027 | .049 | −.014 | −.013 | .018 | −.024 | .038 | −.012 | .015 | .000 | −.010 |

*Note*. Classroom Assessment Scoring System items are listed along the columns, and Mathematical Quality of Instruction items are listed along the rows. Cells with correlations above 0.3 are shaded. SMQR = Student Mathematical Questioning and Reasoning; ETCA = Enacted Task Cognitive Activation.
†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

**Table 5.** Exploratory Factor Analyses Loadings for a Three-Factor Solution.

| | Factor 1 | Factor 2 | Factor 3 | Communalities |
|---|---|---|---|---|
| Eigenvalues | 8.49 | 4.02 | 1.94 | |
| Cumulative % of variance explained | 32.32 | 46.67 | 52.95 | |
| CLASS | | | | |
| Negative climate | −0.578 | −0.110 | −0.003 | 0.343 |
| Behavior management | 0.597 | 0.141 | 0.045 | 0.360 |
| Productivity | 0.691 | 0.218 | 0.059 | 0.478 |
| Positive climate | 0.806 | 0.165 | 0.030 | 0.662 |
| Teacher sensitivity | 0.852 | 0.330 | −0.016 | 0.730 |
| Respect for student perspectives | 0.761 | 0.343 | 0.062 | 0.592 |
| Instructional learning formats | 0.687 | 0.253 | −0.035 | 0.475 |
| Content understanding | 0.832 | 0.289 | 0.082 | 0.696 |
| Analysis and problem solving | 0.711 | 0.459 | 0.052 | 0.570 |
| Quality of feedback | 0.812 | 0.329 | 0.059 | 0.667 |
| Instructional dialogue | 0.841 | 0.410 | 0.031 | 0.729 |
| Student engagement | 0.717 | 0.166 | −0.001 | 0.522 |
| MQI | | | | |
| Linking and connections | 0.199 | 0.556 | −0.190 | 0.314 |
| Explanations | 0.261 | 0.809 | −0.236 | 0.657 |
| Multiple methods | 0.119 | 0.549 | −0.151 | 0.307 |
| Generalizations | 0.209 | 0.394 | −0.098 | 0.162 |
| Mathematical language | 0.352 | 0.363 | −0.138 | 0.199 |
| Remediation | 0.167 | 0.609 | −0.306 | 0.400 |
| Teacher uses student productions | 0.332 | 0.889 | −0.184 | 0.792 |
| Student explanations | 0.236 | 0.808 | −0.123 | 0.658 |
| SMQR | 0.254 | 0.701 | −0.013 | 0.515 |
| ETCA | 0.296 | 0.839 | −0.236 | 0.707 |
| Major errors | 0.011 | −0.195 | 0.835 | 0.698 |
| Language imprecisions | 0.058 | −0.172 | 0.509 | 0.267 |
| Lack of clarity | −0.005 | −0.174 | 0.858 | 0.739 |

*Note*. Extraction method is Principal Axis Factoring. Rotation method is Oblimin with Kaiser Normalization. Cells are shaded to identify instructional factors and potential cross-loadings (i.e., loadings on two factors of similar magnitude). CLASS = Classroom Assessment Scoring System; MQI = Mathematical Quality of Instruction; SMQR = Student Mathematical Questioning and Reasoning; ETCA = Enacted Task Cognitive Activation.



communalities that are considerably low. Together, these three factors explain roughly 53% of the variance in our data, and all have acceptable reliability indices (0.91 for Factor 1, 0.87 for Factor 2, and 0.76 for Factor 3).

With one exception, these factors do not align with structures laid out by instrument developers. Factor 3 is consistent with the Errors and Imprecisions dimension from the MQI, and thus we continue to use this name. Of the other two factors, one includes all items from the CLASS instrument (Factor 1); the other includes the rest of the items from the MQI (Factor 2). That said, correlations among these factors support substantive interpretations (see Table 7). Factors 1 and 2 are moderately correlated ($r$ = .33), suggesting that teachers who engage in a range of organizational, instructional, and emotional supports for students may also provide rich mathematical activities to students and positively engage them in this content. The negative correlation between Factors 2 and 3 ($r$ = –.23) also makes sense given the negative valence of items under the Errors and Imprecisions dimension (Factor 3) and the positive valence of the other items from the MQI (Factor 2). The correlation between Factors 1 and 3 is negligible ($r$ = .03), supporting our hypothesis that purely math-specific practices and the extent to which teachers make errors in their instruction (Factor 3) would not overlap with or be related to teachers' general interactions with students in the classroom (Factor 1).

When we add a fourth factor, items from the CLASS split into two dimensions (see Table 6). One of these, Factor 4, aligns substantively with the Classroom Organization dimension described by instrument developers; we continue to use this name to describe this cluster of items. The other, Factor 1, includes the rest of the items from the CLASS. MQI items load substantively onto the same

**Table 6.** Exploratory Factor Analyses Loadings for a Four-Factor Solution.

| | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Communalities |
|---|---|---|---|---|---|
| Eigenvalues | 8.49 | 4.02 | 1.94 | 1.48 | |
| Cumulative % of variance explained | 32.56 | 47.04 | 53.33 | 58.06 | |
| CLASS | | | | | |
| Negative climate | −0.459 | −0.122 | −0.005 | −0.687 | 0.489 |
| Behavior management | 0.428 | 0.163 | 0.067 | 0.930 | 0.876 |
| Productivity | 0.572 | 0.232 | 0.065 | 0.772 | 0.646 |
| Positive climate | 0.803 | 0.151 | 0.005 | 0.504 | 0.679 |
| Teacher sensitivity | 0.815 | 0.325 | −0.034 | 0.611 | 0.719 |
| Respect for student perspectives | 0.850 | 0.320 | 0.031 | 0.302 | 0.747 |
| Instructional learning formats | 0.656 | 0.249 | −0.050 | 0.492 | 0.468 |
| Content understanding | 0.819 | 0.279 | 0.060 | 0.544 | 0.693 |
| Analysis and problem solving | 0.784 | 0.443 | 0.025 | 0.292 | 0.664 |
| Quality of feedback | 0.851 | 0.311 | 0.030 | 0.426 | 0.725 |
| Instructional dialogue | 0.896 | 0.392 | 0.000 | 0.416 | 0.811 |
| Student engagement | 0.650 | 0.167 | −0.011 | 0.606 | 0.528 |
| MQI | | | | | |
| Linking and connections | 0.212 | 0.557 | −0.194 | 0.101 | 0.314 |
| Explanations | 0.267 | 0.816 | −0.238 | 0.158 | 0.671 |
| Multiple methods | 0.162 | 0.546 | −0.157 | −0.021 | 0.309 |
| Generalizations | 0.198 | 0.398 | −0.099 | 0.160 | 0.169 |
| Mathematical language | 0.309 | 0.370 | −0.140 | 0.325 | 0.221 |
| Remediation | 0.181 | 0.611 | −0.308 | 0.075 | 0.401 |
| Teacher uses student productions | 0.359 | 0.889 | −0.191 | 0.155 | 0.792 |
| Student explanations | 0.273 | 0.806 | −0.129 | 0.070 | 0.656 |
| SMQR | 0.277 | 0.701 | −0.018 | 0.114 | 0.516 |
| ETCA | 0.316 | 0.841 | −0.241 | 0.148 | 0.710 |
| Major errors | 0.018 | −0.199 | 0.835 | 0.005 | 0.697 |
| Language imprecisions | 0.042 | −0.171 | 0.513 | 0.084 | 0.273 |
| Lack of clarity | 0.006 | −0.177 | 0.860 | −0.013 | 0.742 |

*Note*. Extraction method is Principal Axis Factoring. Rotation method is Oblimin with Kaiser Normalization. Cells are shaded to identify instructional factors and potential cross-loadings (i.e., loadings on two factors of similar magnitude). CLASS = Classroom Assessment Scoring System; MQI = Mathematical Quality of Instruction; SMQR = Student Mathematical Questioning and Reasoning; ETCA = Enacted Task Cognitive Activation.

**Table 7.** Correlations Among the Three Factors Emerging from the Exploratory Factor Analysis.

| | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| Factor 1 | 1.00 | | |
| Factor 2 | 0.33 | 1.00 | |
| Factor 3 | 0.03 | −0.23 | 1.00 |

two factors described above (Factors 2 and 3). Although we explain 5% more variation compared to a three-factor solution, for a total of 58%, the low communality values for Mathematical Language and Generalizations persist. In addition, a number of items from the CLASS cross load onto both Factors 1 and 4. As expected given these cross loadings, the two factors with items from the CLASS are correlated more strongly ($r$ = .51) than other combinations of factors (see Table 8). Of interest, we find that the nine items that have their primary loading on Factor 1 in this four-factor solution have stronger internal consistency reliability (0.96) than all 12 items from the CLASS (0.91), which formed Factor 1 in the three-factor solution. The three items that appear to cluster together to form Factor 4 (i.e., Classroom Organization) also have reasonable reliability (0.82), lending support to this solution as a plausible factor structure.

The cross loadings from the EFA suggest that some shared variation may exist across instruments, even though the suggested factors are largely within instrument. To explore this further, we proceed to CFA to test whether extracting instrument-specific variation leads us to one of the two solutions just described, or to another solution. We focus on a parsimonious set of models based both on findings from the EFA and hypotheses just described. We present the structure of these theory-

**Table 8.** Correlations Among the Four Factors Emerging from the Exploratory Factor Analysis.

| | Factor 1 | Factor 2 | Factor 3 | Factor 4 |
|---|---|---|---|---|
| Factor 1 | 1.00 | | | |
| Factor 2 | 0.35 | 1.00 | | |
| Factor 3 | 0.02 | −0.24 | 1.00 | |
| Factor 4 | 0.51 | 0.15 | 0.01 | 1.00 |

driven models in Tables 9 and 10, which document nonbifactor and bifactor models, respectively. The nonbifactor models provide a basis for comparison with the more complex, bifactor models. Models 1 through 4 are nonbifactor models. Models 3 and 4 correspond to the three- and four-factor solutions from the EFA analyses, with items restricted to load only on their primary factors from the EFA. As in the EFA, all items were specified to cluster within instrument. We also ran models with just one instructional factor (Model 1) and two factors comprising items from each instrument (Model 2) in order to examine whether the suggested models that emerged from the EFA had better fit as compared to that of more parsimonious models. This is a common practice when running CFA (R. B. Kline, 2011).

Models 5 through 8 are bifactor models, each with two instrument factors (CLASS and MQI) that attempt to extract instrument-specific variation, as well as varying numbers of instructional factors, including cross-instrument factors. We specified models with no more than three instructional factors because the EFA suggested no more than four instructional factors, and we hypothesized that extraction of instrument-specific variation would allow items to cluster across instruments to form fewer instructional factors. Model 5 includes two instrument factors and just one instructional factor. Based on theory and results from the EFA, we did not expect findings to point to just one instructional factor; however, similar to earlier in this article, we specified this model anyway as a comparison to more complex models. Models 6

**Table 9.** Confirmatory Factor Analysis Model Organization for Nonbifactor Models.

| Items | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| CLASS | | | | |
| Negative climate | Factor 1 | Factor 1 | Factor 1 | Factor 1 |
| Behavior management | Factor 1 | Factor 1 | Factor 1 | Factor 1 |
| Productivity | Factor 1 | Factor 1 | Factor 1 | Factor 1 |
| Positive climate | Factor 1 | Factor 1 | Factor 1 | Factor 2 |
| Teacher sensitivity | Factor 1 | Factor 1 | Factor 1 | Factor 2 |
| Respect for student perspectives | Factor 1 | Factor 1 | Factor 1 | Factor 2 |
| Instructional learning formats | Factor 1 | Factor 1 | Factor 1 | Factor 2 |
| Content understanding | Factor 1 | Factor 1 | Factor 1 | Factor 2 |
| Analysis and problem solving | Factor 1 | Factor 1 | Factor 1 | Factor 2 |
| Quality of feedback | Factor 1 | Factor 1 | Factor 1 | Factor 2 |
| Instructional dialogue | Factor 1 | Factor 1 | Factor 1 | Factor 2 |
| Student engagement | Factor 1 | Factor 1 | Factor 1 | Factor 2 |
| MQI | | | | |
| Linking and connections | Factor 1 | Factor 2 | Factor 2 | Factor 3 |
| Explanations | Factor 1 | Factor 2 | Factor 2 | Factor 3 |
| Multiple methods | Factor 1 | Factor 2 | Factor 2 | Factor 3 |
| Generalizations | Factor 1 | Factor 2 | Factor 2 | Factor 3 |
| Mathematical language | Factor 1 | Factor 2 | Factor 2 | Factor 3 |
| Remediation | Factor 1 | Factor 2 | Factor 2 | Factor 3 |
| Teacher uses student productions | Factor 1 | Factor 2 | Factor 2 | Factor 3 |
| Student explanations | Factor 1 | Factor 2 | Factor 2 | Factor 3 |
| SMQR | Factor 1 | Factor 2 | Factor 2 | Factor 3 |
| ETCA | Factor 1 | Factor 2 | Factor 2 | Factor 3 |
| Major errors | Factor 1 | Factor 2 | Factor 3 | Factor 4 |
| Language imprecisions | Factor 1 | Factor 2 | Factor 3 | Factor 4 |
| Lack of clarity | Factor 1 | Factor 2 | Factor 3 | Factor 4 |
| No. of factors | 1 | 2 | 3 | 4 |

*Note.* CLASS = Classroom Assessment Scoring System; MQI = Mathematical Quality of Instruction; SMQR = Student Mathematical Questioning and Reasoning; ETCA = Enacted Task Cognitive Activation.

**Table 10.** Confirmatory Factor Analysis Model Organization for Bifactor Models.

| Items | Model 5 | Model 6 | Model 7 | Model 8 |
|---|---|---|---|---|
| CLASS | | | | |
| Negative climate | Factor 1 | Factor 1 | Factor 1 | Factor 1 |
| Behavior management | Factor 1 | Factor 1 | Factor 1 | Factor 1 |
| Productivity | Factor 1 | Factor 1 | Factor 1 | Factor 1 |
| Positive climate | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| Teacher sensitivity | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| Respect for student perspectives | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| Instructional learning formats | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| Content understanding | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| Analysis and problem solving | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| Quality of feedback | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| Instructional dialogue | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| Student engagement | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| MQI | | | | |
| Linking and connections | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| Explanations | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| Multiple methods | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| Generalizations | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| Mathematical language | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| Remediation | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| Teacher uses student productions | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| Student explanations | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| SMQR | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| ETCA | Factor 1 | Factor 1 | Factor 2 | Factor 2 |
| Major errors | Factor 1 | Factor 2 | Factor 2 | Factor 3 |
| Language imprecisions | Factor 1 | Factor 2 | Factor 2 | Factor 3 |
| Lack of clarity | Factor 1 | Factor 2 | Factor 2 | Factor 3 |
| No. of factors | 3 | 4 | 4 | 5 |

*Note.* All models also include two instrument factors with all items cross loading onto their respective instrument factors. CLASS = Classroom Assessment Scoring System; MQI = Mathematical Quality of Instruction; SMQR = Student Mathematical Questioning and Reasoning; ETCA = Enacted Task Cognitive Activation.



and 7 each have two instructional factors, one that overlaps substantially with the single instructional factor in Model 5 and another from the CLASS or from the MQI that the correlational analyses and EFA results suggested were important (i.e., Errors and Imprecisions from the MQI in Model 6 and Classroom Organization from the CLASS in Model 7). Finally, Model 8 includes both of these latter instructional factors, Errors and Imprecisions and Classroom Organization; a third instructional factor includes the rest of the items from both the CLASS and MQI instruments. Bifactor models such as Model 8 allow us to test the hypothesis that these sorts of instructional components from the CLASS and MQI might be correlated but would not cluster onto the same factor. We hypothesized that the MQI would capture additional variability in teachers' use of student ideas in mathematics instruction than could be captured by a general instrument such as the CLASS. If our hypothesis holds, Model 8 should not be a better fit to the data than a model structure such as Model 4, where related CLASS and MQI items load onto separate factors.

In Table 11, we present standard fit statistics (chi-square) and fit indices (root mean square error of approximation, comparative fit index, standardized root mean square residual, Akaike information criterion [AIC], and Bayesian information criterion [BIC]) for each of these models using robust maximum likelihood estimation to account for the non-normality of some items (Sivo, Fan, Witta, & Willse, 2006). Although the fit of nested models normally can be compared using a chi-square-difference test, it is not appropriate here given that our base model has a significant chi-square at the nominal .05 alpha level (Yuan & Bentler, 2004). Further, these models do not meet expected guidelines of fit indices: less than 0.06 for root mean square error of approximation, greater than 0.95 for comparative fit index, and less than 0.08 for standardized root mean square residual (Hu & Bentler, 1999). We instead rely on AIC and BIC indices when identifying models with the best fit, looking for those models with the smallest AIC and BIC indices.

**Table 11.** Model Fit Indices for Confirmatory Factor Analysis Models.

| | Single Factor—No Cross Loadings | | | | Bifactor—Items Load Onto Their Respective Instruments, Plus Other Factors | | | |
|---|---|---|---|---|---|---|---|---|
| Fit Indices | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 | Model 8 |
| No. of factors | 1 | 2 | 3 | 4 | 3 | 4 | 4 | 5 |
| No. of parameters | 75 | 76 | 78 | 81 | 101 | 102 | 102 | 104 |
| Akaike (AIC) | −1639.91 | −3128.08 | −3499.79 | −3795.80 | −3570.68 | −4000.36 | −3886.38 | −4059.54 |
| Bayesian (BIC) | −1342.45 | −2826.65 | −3190.43 | −3474.54 | −3170.10 | −3595.81 | −3481.83 | −3647.061 |
| $\chi^2$ statistic | 3114.10 | 1912.41 | 1642.76 | 1399.24 | 1542.98 | 1199.75 | 1295.26 | 1169.95 |
| $\chi^2$ *df* | 275 | 274 | 272 | 269 | 249 | 248 | 248 | 246 |
| $\chi^2$ test of model fit | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Root mean square error of approximation | 0.163 | 0.124 | 0.114 | 0.104 | 0.115 | 0.099 | 0.104 | 0.098 |
| Comparative fit index | 0.484 | 0.702 | 0.751 | 0.795 | 0.765 | 0.827 | 0.810 | 0.832 |
| Standardized root mean square residual | 0.156 | 0.093 | 0.077 | 0.070 | 0.070 | 0.055 | 0.067 | 0.065 |



Several factors could explain our lack of adequate model fit, including the possibility of failing to model all of the sources of variation in our data, collecting noisy data, or using a moderately sized sample. This misfit implies that we should interpret the structure and meaning of our teacher-level factors with caution. We interpret the identified factors tentatively as corresponding to the constructs they were theoretically derived from while remaining realistic that they likely capture some variation that is unrelated to the construct of interest.

Model fit comparisons indicate that both general and content-specific dimensions are needed to describe variation across teachers. Of the nonbifactor models, Model 4, which includes four total factors without any overlap across instruments, appears to have the best fit according to the AIC and BIC statistics. Similarly, for bifactor models, Model 8, which has three total factors in addition to the two instrument factors, has the best fit. In both Models 4 and 8, Errors and Imprecisions and Classroom Organization form their own factors, even though this was not true in the three-factor solution from the EFA. In that analysis, we found that all items from the CLASS instrument clustered together to form a single factor. This finding about Errors and Imprecisions and Classroom Organization aligns with one of our hypotheses and highlights at least one area in which a general instrument such as the CLASS may be limited in its ability to capture some "pure" content-specific elements of instruction.

There also are important differences between Model 4 and Model 8, namely, that the latter allows items from the CLASS and the MQI to cluster onto the same factor, whereas the former does not. Unlike before, when aiming to determine which of these structures has a better fit to our data, we are not able to compare fit indices. This is because we expect the AIC and BIC indices to be lower in the bifactor model than in the nonbifactor models with similar numbers of instructional factors, given that the former includes two instrument-specific factors intended to capture more of the variability in our data than could be captured by nonbifactor models.

Therefore, we examine factor loadings from these two models to determine which has stronger substantive backing (see Table 12 for loadings from Model 4 and Table 13 for loadings from Model 8). Similar to the results from the EFA, we find substantive support for Model 4, where items have statistically significant loadings on their respective factors, generally above 0.40. Two item loadings for Generalizations and Mathematical Language fall just below this threshold; we observed similar issues in the EFA. Factor loadings in Model 8 are less clean. We hypothesized that variation in a particular item should be accounted for both by an instrument factor and by the content of the item. This is true for the Classroom Organization factor, where all three items have loadings on both the instrument and the instructional factor above 0.40. Four items from the MQI specified to load on the common factor with items from the CLASS: Teacher Uses Student Productions, Student Explanations, Student Mathematical Questioning and Reasoning, and Enacted Task Cognitive

**Table 12.** Standardized Factor Loadings for CFA Model 4.

| Items | Factor 1 | Factor 2 | Factor 3 | Factor 4 |
|---|---|---|---|---|
| CLASS | | | | |
| Negative climate | 0.699*** | | | |
| Behavior management | −0.841*** | | | |
| Productivity | −0.883*** | | | |
| Positive climate | | 0.797*** | | |
| Teacher sensitivity | | 0.823*** | | |
| Respect for student perspectives | | 0.821*** | | |
| Instructional learning formats | | 0.673*** | | |
| Content understanding | | 0.831*** | | |
| Analysis and problem solving | | 0.780*** | | |
| Quality of feedback | | 0.856*** | | |
| Instructional dialogue | | 0.886*** | | |
| Student engagement | | 0.671*** | | |
| MQI | | | | |
| Linking and connections | | | 0.524*** | |
| Explanations | | | 0.759*** | |
| Multiple methods | | | 0.523*** | |
| Generalizations | | | 0.389*** | |
| Mathematical language | | | 0.368*** | |
| Remediation | | | 0.575*** | |
| Teacher uses student productions | | | 0.909*** | |
| Student explanations | | | 0.836*** | |
| SMQR | | | 0.746*** | |
| ETCA | | | 0.848*** | |
| Major errors | | | | 0.834*** |
| Language imprecisions | | | | 0.508*** |
| Lack of clarity | | | | 0.876*** |

*Note.* CLASS = Classroom Assessment Scoring System; MQI = Mathematical Quality of Instruction; SMQR = Student Mathematical Questioning and Reasoning; ETCA = Enacted Task Cognitive Activation.
***$p < .001$.



Activation—also meet this condition. However, all three items from Errors and Imprecisions load only onto the instructional factor at this threshold; loadings on the instrument factor are below 0.40, even though they are statistically significant. Conversely, for the rest of the items from the CLASS and MQI specified to load onto the common instructional factor, almost all of the variation loads onto the instrument factor. For example, Linking and Connections has a strong loading of 0.57 on the MQI instrument factor but a nonsignificant loading of 0.14 on the instructional factor. Items from the CLASS specified to load on this same instructional factor have loadings no higher than 0.35. One reason for this may be that there is a large degree of overlap of items between this instructional factor and the two instrument factors; that is, we specify that all but three CLASS items and all but three MQI items load onto the common instructional factor. Another explanation is that our sample size is small relative to recommended guidelines for stable parameter estimates. The rule of thumb is to have five to 10 observations per parameter estimated (R. B. Kline, 2011), yet we have only 390 observations for approximately 100 parameters.

Even when we look beyond these measurement challenges, we find suggestive evidence that Model 4 offers a better solution than Model 8. In Model 8, two items from CLASS—Positive Climate and Content Understanding—and five items from the MQI—Linking and Connections, Explanations, Generalizations, Language, and Remediation—have nonsignificant loadings below 0.20. One interpretation of these loadings is that most of the variance for these items is captured by the instrument factor. Of interest, though, five of these six items (excluding Positive Climate) are rooted in a content-specific view of instruction. Therefore, it is possible that these items might form a separate cross-instrument factor. We did not explore this as a possible factor structure, as this would lead us to the same general conclusion as Model 4, with two mathematics-specific factors and two general factors. As we hypothesized earlier, these patterns suggest that there are subtleties in a content-specific instrument such as the MQI that capture additional

**Table 13.** Standardized Factor Loadings for Confirmatory Factor Analysis Model 8.

| | Instrument Factors | | Instructional Factors | | |
|---|---|---|---|---|---|
| Items | CLASS | MQI | Factor 1 | Factor 2 | Factor 3 |
| CLASS | | | | | |
| Negative climate | –0.493*** | | –0.486*** | | |
| Behavior management | 0.451*** | | 0.836*** | | |
| Productivity | 0.619*** | | 0.559*** | | |
| Positive climate | 0.808*** | | | 0.100† | |
| Teacher sensitivity | 0.795*** | | | 0.201** | |
| Respect for student perspectives | 0.756*** | | | 0.333*** | |
| Instructional learning formats | 0.615*** | | | 0.266*** | |
| Content understanding | 0.855*** | | | 0.078 | |
| Analysis and problem solving | 0.719*** | | | 0.326*** | |
| Quality of feedback | 0.849*** | | | 0.180** | |
| Instructional dialogue | 0.820*** | | | 0.348*** | |
| Student engagement | 0.619*** | | | 0.237** | |
| MQI | | | | | |
| Linking and connections | | 0.573*** | | 0.137 | |
| Explanations | | 0.903*** | | 0.133 | |
| Multiple methods | | 0.486*** | | 0.245† | |
| Generalizations | | 0.428*** | | 0.084 | |
| Mathematical language | | 0.382*** | | 0.113 | |
| Remediation | | 0.704*** | | 0.050 | |
| Teacher uses student productions | | 0.604** | | 0.722*** | |
| Student explanations | | 0.617*** | | 0.578** | |
| SMQR | | 0.432* | | 0.654*** | |
| ETCA | | 0.636** | | 0.535* | |
| Major errors | | –0.238*** | | | –0.788*** |
| Language imprecisions | | –0.166** | | | –0.477*** |
| Lack of clarity | | –0.197** | | | –0.870*** |

*Note.* CLASS = Classroom Assessment Scoring System; MQI = Mathematical Quality of Instruction; SMQR = Student Mathematical Questioning and Reasoning; ETCA = Enacted Task Cognitive Activation.
†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.



variability in teachers' interactions with students around that content than are captured by a more general instrument such as the CLASS.

That said, there does appear to be some overlap between instructional components of the CLASS and MQI instruments that differs from our stated hypotheses. This is most evident for items related to students' cognitive engagement in class and the content. In Model 8, four items from the MQI—Teacher Uses Student Productions, Student Explanations, Student Mathematical Questioning and Reasoning, and Enacted Task Cognitive Activation—load onto the factor that we specified to include items from both instruments. If we consider a slightly lower threshold for factor loadings around 0.30, then three related items from the CLASS instrument—Respect for Student Perspectives, Analysis and Problem Solving, and Instructional Dialogue—also appear to load onto this same instructional factor. Thus, a general instrument may be able to capture some of the same variability in teachers' instruction as a content-specific instrument.

## Discussion and conclusion

Results from this study identify only a small degree of overlap between a general and a content-specific instrument when used to examine upper-elementary teachers' mathematics instruction. Although we find some overlap between elements of instruction captured by the CLASS and MQI instruments in Model 8 of the CFA, we also find strong evidence for factors that are distinct to each. At the extreme, Errors and Imprecisions is content specific and Classroom Organization is more general. We also find distinctions between a general instrument and a content-specific one even among items that ask about teachers' interactions with students and the extent to which these interactions develop intellectual challenge, as observed in Model 4 from the CFA. Aligned with older

(Brophy, 1986) and more recent (Grossman & MacDonald, 2008; Hamre et al., 2013) work, these findings provide empirical support for the argument that, when studying complex phenomena such as that of teaching, we need to take both types of perspectives—a general and a content-specific viewpoint—into consideration if we are to better understand and capture the quality of instruction experienced by students in the classroom (see Charalambous & Kyriakides, 2017, for a related discussion).

These general conclusions also align with related empirical work using classroom data scored on multiple observation instruments. Lockwood et al. (2015) identified two distinct teaching factors, one focused on instructional practice and another focused on classroom management. In the MET study, researchers found similar substantive factors within instruments and very little overlap between instruments (Kane & Staiger, 2012; McClellan et al., 2013). At the same time, we also note areas of disagreement. Our analyses demonstrate support for between three and four instructional factors to describe variation across teachers, slightly more than the number identified by Lockwood et al. (2015) and far fewer than the number identified by McClellan et al. (2013). One reason for this latter difference likely is the fact that the MET data included scores from five observation instruments, whereas ours include scores from two. It is possible that we might find more factors if we were to score the same instruction on additional instruments. Another plausible reason for these discrepancies might be due to the models tested in each study. Our work used bifactor models that have the potential to account for any instrument-related variance, thus reducing the number of unique factors. Lockwood et al. (2015) did not account for variance due to instruments but did find fewer factors when accounting for other sources of construct-irrelevant variation (i.e., segments, lessons, and raters). Of course, it is possible that analyses that simultaneously account for all possible sources of variation—that is, raters, segments, lessons, instruments, and teachers—may come to slightly different conclusions. The fact that we are not able to do so is a limitation of this study.

We believe that our findings have important implications for the development and refinement of the two observational instruments, as well as for policy and practice. First, we note that our final factor structures do not align completely with those presented by instrument developers, thus requiring additional consideration of the groups of teaching skills that these instruments capture. Compared to four dimensions on the MQI instrument, we identify two; this is true both in Model 4 of the CFA and in Model 8, where one of these dimensions overlaps with items from the CLASS. The first of these factors, Errors and Imprecisions, is consistent with the original instrument. The second factor looks a lot like what many have described as "ambitious instruction" (Cohen, 2011; Lampert, 2001), referring to instruction that is "intellectually ambitious, uncertain, and contested" (Cohen & Ball, 1999, p. 6). Thus, we call the second factor Ambitious Mathematics Instruction, capturing the depth of the mathematics provided to students, the quality of teachers' interactions with students around this content, and opportunities for students to derive meaning about mathematical ideas. Compared to three dimensions on the original CLASS instrument, our empirical evidence points to two: Classroom Organization and a separate factor that appears to capture teachers' support—both instructional and emotional—for students in the classroom. These results are similar to other two-factor solutions identified in studies that also used bifactor or hierarchical CFA (Hamre et al., 2013; McCaffrey et al., 2015) but differ from other analyses that identify three factors (Hafen et al., 2015; Hamre et al., 2013). In light of the substantial work already conducted to date on the CLASS instrument, we leave further discussion about these differences and possible names for factors that emerge from a more parsimonious structure up to developers of this instrument.

The multidimensional nature of instruction—including both general and content-specific practices—requires evaluation systems that reflect this complex structure. Current discussion around teacher evaluation often advocate for the use of "multiple measures" in a way that implies consensus around the complexity of teaching. However, in practice, evaluation systems often assume a unidimensional or simplified criterion for evaluation (Rothstein & Mathis, 2013). One reason for this is the fact that current processes generally evaluate teachers on just one observation instrument, thus likely masking important variability within and across teachers. Similarly, both researchers and policymakers suggest

creating a single weighted composite of teachers' overall effectiveness by averaging across multiple metrics (e.g., dimensions within a given observation instrument, value-added to student test scores, student surveys; Center on Great Teachers and Leaders, 2013; Mihaly, McCaffrey, Staiger, & Lockwood, 2013). We recognize that simplifying the evaluation process—both by using a single observation instrument and by averaging across measures—lends itself to a systematized process for making binary decisions such as whether to grant teachers tenure. At the same time, such an approach cannot account for the full complexity of teachers' skill and classroom practice. Take, for instance, the case of a teacher who might do poorly in terms of establishing emotional rapport with her students but does an outstanding job in terms of creating a mathematically rich learning environment for her students. Or conversely, take a teacher who might perform extremely well in managing his classroom but whose instruction is replete with major mathematical errors. Both teachers would earn an overall score in the middle of the distribution of teacher effectiveness even though each is someone who might be an appropriate target for (different forms of) professional development. If the goal of using observation instruments is instructional improvement, which many argue it is (Hill & Grossman, 2013; Papay, 2012), then it is important to be able to provide teachers and school personnel with distinct dimension-specific scores that lead to individualized support targeted at skills and areas where they are lacking. Without this sort of information, evaluation scores may be useful only for generalized one-size-fits-all professional development, which has not proven effective at increasing teachers' instructional quality or student achievement (Hill, 2007; Yoon, Duncan, Lee, Scarloss, & Shapley, 2007), or for dismissal or promotion.

Finally, the results of this study also have practical implications for selecting and training raters to score teachers' instruction. Even though prior work highlights the ability of principals, peers, and other school leaders to accurately identify teachers who are effective at raising student achievement (Jacob & Lefgren, 2008; Rockoff & Speroni, 2010; Rockoff, Staiger, Kane, & Taylor, 2012), other work indicates that specific types of instruction—particular in a content area—require raters attuned to these elements. For example, Hill, Charalambous, Blazar, et al. (2012) show that raters who are selectively recruited due to a background in mathematics or mathematics education and who complete initial training and ongoing calibration score more accurately on the MQI than those who are not selectively recruited. Therefore, calls to identify successful teachers through evaluations that are "better, faster, and cheaper" (Gargani & Strong, 2014) may not prove useful across all instructional dimensions. Instead, along with Good and Lavigne (2015), we are more supportive of a "festina-lente" approach, in which intensive training helps raters to deeply understand the different dimensions of a given observational instrument and to apply specific items accurately and knowledgeably. This, in turn, raises concerns as to whether principals and school leaders have the capacity and the expertise to appropriately evaluate content-related dimensions of instructional quality in several subject matters, as is the case when supervising generalist primary school teachers.

That said, we are not dismissive of the central tension between the need for content-specific observation and the logistics-related challenges of selecting such instruments, training observers, and managing these systems. There is a range of possible ways to resolve this tension. One possibility would be to deputize content-specific district staff to take over a portion of teacher evaluations. The drawbacks of this proposal would be the mixing of professional developer and evaluator roles, which some point to as a problematic feature in current teacher evaluation systems (Herman & Baker, 2009). The benefit might be enhanced ability for that staff to provide customized feedback to teachers, a feature of successful coaching programs focused either on content-specific or general instructional dimensions (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Blazar & Kraft, 2015; P. F. Campbell & Malkus, 2011), and potentially increased coordination between teachers' needs and available professional development. A second possibility would be to overlay content-specific instructional guidance onto content-generic observation instruments, essentially providing both verbiage and observer training that would allow for a more nuanced understanding of what instructional activities like asking high-level questions, for instance, looks like in mathematics classrooms. Observers could be trained in both the general instrument and the subject matter

areas, thus aligning sources of instructional guidance and increasing the specificity of guidance to teachers. A third possibility may be to identify which of the dimensions of teaching practice are most predictive of student outcomes that we care about, and to focus on a parsimonious set of skills. This sort of predictive validity evidence has been important for advocates of value-added measures of teacher effectiveness that estimate the contribution of teachers to student outcomes (Chetty, Friedman, & Rockoff, 2014). However, Rothstein and Mathis (2013) pointed out that different student outcomes may be predicted by different combinations of teaching practices, which is borne out in several studies (Blazar & Kraft, 2017; Downer, Rimm-Kaufman, & Pianta, 2007; Hamre & Pianta, 2001; Luckner & Pianta, 2011). These findings make this approach less tractable than one where multiple teaching skills are taken into account.

Understanding and better measuring instructional quality is a particularly complex yet necessary endeavor. Although our results suggest that no fast and cheap solutions seem to exist, successfully undertaking this task appears necessary for improving instructional quality in the years to come.



## Acknowledgment

We thank the teachers who participated in this study, and the anonymous reviewers who provided helpful feedback on earlier drafts.

## Funding

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C090023 to the President and Fellows of Harvard College to support the National Center for Teacher Effectiveness. Additional funding comes from the National Science Foundation Grant 0918383. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

## References

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.

Ball, D. L. (1988). *Research on teaching mathematics: Making subject matter knowledge part of the equation*. East Lansing, MI: National Center for Research on Teacher Education.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching what makes it special? *Journal of Teacher Education*, *59*, 389–407.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., & Pianta, R. C. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2–3), 62–87.

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review*, *48*, 16–29.

Blazar, D., & Kraft, M. A. (2015). Exploring mechanisms of effective teacher coaching: A tale of two cohorts from a randomized experiment. *Educational Evaluation and Policy Analysis*, *37*, 542–566.

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis*, *39*, 146–170.

Blazar, D., Litke, E., & Barmore, J. (2016). What does it mean to be ranked a "high" or "low" value-added teacher? Observing differences in instructional quality across districts. *American Educational Research Journal*, *53*, 324–359.

Brophy, J. (1986). Teaching and learning mathematics: Where research should be going? *Journal for Research in Mathematics Education*, *17*, 323–346.

Brophy, J. E., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). New York, NY: Macmillan.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Campbell, P. F., & Malkus, N. N. (2011). The impact of elementary mathematics coaches on student achievement. *The Elementary School Journal*, *111*, 430–454.

Center on Great Teachers and Leaders. (2013). *Databases on state teacher and principal policies*. Retrieved from http://resource.tqsource.org/stateevaldb

Charalambous, C. Y., & Kyriakides, E. (2017). Working at the nexus of generic and content-specific teaching practices: An exploratory study based on TIMSS secondary analyses. *Elementary School Journal*, *117*(3), 423–454.
Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J.-P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, *80*, 219–251.
Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*, 2633–2679.
Cohen, D. K. (2011). *Teaching and its predicaments*. Cambridge, MA: Harvard University Press.
Cohen, D. K., & Ball, D. L. (1999). *Instruction, capacity, and improvement* (CPRE Research Report Series RR–43). Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania Graduate School of Education.
Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.
Danielson Group. (2013a). *The Framework for Teaching Evaluation Instrument*. Princeton, NJ: Author.
Danielson Group. (2013b). *General questions about the framework*. Retrieved from https://www.danielsongroup.org/questions-about-the-framework-for-teaching/
Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA: Association for Supervision and Curriculum Development.
Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. New York, NY: Teachers College Press.
Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the United States and abroad*. Palo Alto, CA: National Staff Development Council and The School Redesign Network, Stanford University.
Darling-Hammond, L., Wise, A. E., & Pease, S. R. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research*, *53*, 285–328.
Downer, J. T., Rimm-Kaufman, S., & Pianta, R. C. (2007). How do classroom conditions and children's risk for school problems contribute to children's behavioral engagement in learning? *School Psychology Review*, *36*, 413–432.
Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299.
Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). London, England: Sage.
Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*, 915–945.
Gargani, J., & Strong, M. (2014). Can we identify a successful teacher better, faster, and cheaper? Evidence for innovating teacher observation systems. *Journal of Teacher Education*, *65*, 389–401.
Good, T. L., & Lavigne, A. L. (2015). Rating teachers cheaper, faster, and better: Not so fast. *Journal of Teacher Education*, *66*, 288–293.
Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added. *American Journal of Education*, *199*, 445–470.
Grossman, P., & McDonald, M. (2008). Back to the future: Directions for research in teaching and teacher education. *American Educational Research Journal*, *45*, 184–205.
Gustafsson, J., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*, 407–434.
Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., & Pianta, R. C. (2015). Teaching through interactions in secondary school classrooms: Revisiting the factor structure and practical application of the Classroom Assessment Scoring System–Secondary. *Journal of Early Adolescence*, *35*(5–6), 651–680. doi:10.1177/0272431614537117
Hamre, B., Hatfield, B., Pianta, R., & Jamil, F. (2014). Evidence for general and domain-specific elements of teacher–child interactions: Associations with preschool children's development. *Child Development*, *85*, 1257–1274.
Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., … Brackett, M. A. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *The Elementary School Journal*, *113*, 461–487.
Hamre, B. K., & Pianta, R. C. (2001). Early teacher–child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development*, *72*, 625–638.
Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*, 191–205.
Herman J. L., & Baker, E. L. (2009). Assessment policy: Making sense of the Babel. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook of educational policy research* (pp. 176–190). New York, NY: Routledge.
Hill, H. C. (2007). Learning in the teacher workforce. *Future of Children*, *17*, 111–127.
Hill, H. C. (2010). The nature and predictors of elementary teachers' mathematical knowledge for teaching. *Journal for Research in Mathematics Education, 41*(5), 513–545.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*, 430–511.

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

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researchers*, *41*, 56–64.

Hill, H. C., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, *83*, 371–384.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*, 794–831.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Seattle, WA: Measures of Effective Teaching Project, Bill and Melinda Gates Foundation.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *20*, 101–136.

Kaiser, H. (1974). An index of factorial simplicity. *Psychometrika*, *39*, 31–36.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (Research paper). Seattle, WA: Measures of Effective Teaching Project, Bill and Melinda Gates Foundation.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, *46*, 587–613.

Kline, P. (1994). *An easy guide to factor analysis*. London, UK: Routledge.

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. New York, NY: Guilford Press.

Lampert, M. (2001). *Teaching problems and the problems of teaching*. New Haven, CT: Yale University Press.

Leinhardt, G. (1993). On teaching. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol.4, pp. 1–54). Hillsdale, NJ: Erlbaum.

Little, J. (2001). Professional development in the pursuit of school reform. In A. Lieberman & L. Miller (Eds.), *Teachers caught in the action* (pp. 23–44). New York, NY: Teachers College Press.

Lockwood, J. R., Savitsky, T. D., & McCaffrey, D. F. (2015). Inferring constructs of effective teaching from classroom observations: An application of Bayesian exploratory factor analysis without restrictions. *The Annals of Applied Statistics*, *9*, 1484–1509.

Luckner, A. E., & Pianta, R. C. (2011). Teacher–student interactions in fifth grade classrooms: Relations with children's peer behavior. *Journal of Applied Developmental Psychology*, *32*, 257–266.

Mathematical Quality of Instruction. (2014). Retrieved from http://isites.harvard.edu/icb/icb.do?keyword=mqi_training&pageid=icb.page394761

McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2015). Uncovering multivariate structure in classroom observations in the presence of rater errors. *Educational Measurement: Issues and Practice*, *34*(2), 34–46.

McClellan, C., Donoghue, J., & Park, Y. S. (2013). *Commonality and uniqueness in teaching practice observation*. Retrieved from http://www.clowderconsulting.com/wp-content/uploads/2016/01/Commonality-and-Uniqueness-in-Teaching-Practice-Observation_paper.pdf

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching*. Seattle, WA: Bill & Melinda Gates Foundation, Measures of Effective Teaching Project.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art: Teacher effectiveness and professional learning. *School Effectiveness and School Improvement*, *25*, 231–256.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics. (2014). *Principles to actions: Ensuring mathematical success for all*. Reston, VA: Author.

Odden, A. (2004). Lessons learned about standards-based teacher evaluation systems. *Peabody Journal of Education*, *79*, 126–137.

Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*, 123–141.

Pianta, B., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal*, *45*, 365–387.

Pianta, R., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*, 109–119.

Pianta, R. C., Hamre, B. K., & Mintz, S. (2010). *Classroom Assessment Scoring System (CLASS) Manual: Elementary 4–6 Manual*. Retrieved from http://teachstone.com/classroom-assessment-scoring-system-class/

Pohl, S., & Steyer, R. (2010). Modeling common traits and method effects in multitrait–multimethod analysis. *Multivariate Behavioral Research*, *45*, 45–72.

Rockoff, J. E., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *American Economic Review, 100*(2), 261–266.

Rockoff, J. E., Staiger, D. O., Kane, T. J., & Taylor, E. S. (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review*, *102*, 3184–3213.

Rothstein, J., & Mathis, W. J. (2013). *Review of two culminating reports from the MET project*. National Education Policy Center, University of Colorado at Boulder. Retrieved from http://nepc.colorado.edu/thinktank/review-MET-final-2013

Sandilos, L. E., DiPerna, J. C., & Family Life Project Key Investigators. (2014). Measuring quality in kindergarten classrooms: Structural analysis of the Classroom Assessment Scoring System (CLASS K–3). *Early Education and Development*, *25*, 894–914.

Savitsky, T. D., & McCaffrey, D. F. (2014). Bayesian hierarchical multivariate formulation with factor analysis for nested ordinal data. *Psychometrika*, *79*, 275–302.

Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research*, *46*, 553–611.

Sivo, S. A., Fan, X., Witta, E. L., & Willse, J. T. (2006). The search for "optimal" cutoff properties: Fit index criteria in structural equation modeling. *The Journal of Experimental Education*, *74*, 267–288.

Stiggins, R. J., & Chappuis, J. (2012). *An introduction to student-involved assessment for learning*. Boston, MA: Pearson.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). New York, NY: HarperCollins.

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher professional development: Motives and methods. *Educational Researcher*, 37, 469–479.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. Retrieved from https://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest.

Yuan, K.-H., & Bentler, P. M. (2004). On chi-square difference and z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement*, *64*, 737–757.