UC Riverside

UC Riverside Previously Published Works

Title

Converging on quality: Examining multiple measures of teaching effectiveness

Permalink

https://escholarship.org/uc/item/26r211d4

Journal

Journal of School Psychology, 74

ISSN

0022-4405

Authors

Sandilos, Lia E
Sims, Wesley A
Norwalk, Kate E
et al.

Publication Date

2019-06-01

DOI

10.1016/j.jsp.2019.05.004

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Journal of School Psychology 74 (2019) 10–28

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

Converging on quality: Examining multiple measures of teaching effectiveness

Lia E. Sandilos[a,*], Wesley A. Sims[b], Kate E. Norwalk[c], Linda A. Reddy[d]

[a] Temple University, United States of America
[b] University of California Riverside, United States of America
[c] North Carolina State University, United States of America
[d] Rutgers University, United States of America

ARTICLE INFO

Action Editor: Lisa Sanetti

Keywords:
Teacher evaluation
Classroom assessment
Student achievement

ABSTRACT

The present study explores the convergent and predictive validity for several widely used measures of teaching quality from the Measures of Effective Teaching Project (Bill and Melinda Gates Foundation, 2009-2011). Specifically, the Classroom Assessment Scoring System (CLASS; Pianta, Hamre, & Mintz, 2012), the Framework for Teaching (FFT; Danielson Group, 2013), and the Tripod Student Perceptions Scale (Tripod; Ferguson, 2008) were examined. Correlations among measures were assessed by developmental level and content area (elementary mathematics $N = 70$; elementary English language arts $N = 101$; middle school mathematics $N = 291$, middle school English language arts $N = 280$). Both average scores and score variability (i.e., coefficient of variation) for the CLASS, FFT, and Tripod were used to predict value-added models (VAM), a high-stakes measure of students' academic growth. For elementary mathematics and ELA, findings indicated the CLASS and FFT exhibited moderate convergent validity while divergent validity was found between the Tripod and the CLASS and FFT. Across content areas in middle school grades, the CLASS, FFT, and Tripod exhibited moderate to high-moderate convergent validity. Average student and observer scores were positively related to VAM scores, whereas variability in scores demonstrated negative relations to VAM scores. Implications of findings for teacher evaluation and professional development are discussed.

For decades, practitioners, researchers, and policy makers have endeavored to generate measures that capture “effective teaching” (Stronge, Ward, & Grant, 2011). Teachers have the potential to play a pivotal role in the academic and social-emotional development of their students (Pianta, 1999), yet education research indicates there is considerable variation in the quality of instruction students receive within and across classrooms (Chetty, Friedman, & Rockoff, 2011; Cohen & Goldhaber, 2016; Cohen, Ruzek, & Sandilos, 2018). In the United States, evaluation of effective teaching has propelled forward with the adoption of federal policies that provide incentives based on teacher qualifications and student achievement, such as the Teacher Incentive Fund (Heyburn, Lewis, & Ritter, 2010) and Race to the Top (U.S. Department of Education, 2009). More recently, the Every Student Succeeds Act (ESSA, 2015) stipulated that teacher evaluation systems should include multiple measures of teacher effectiveness that inform instructional planning and professional growth opportunities, underscoring the value of accumulating convergent and predictive psychometric evidence for various measures of instructional quality.

Despite increased national attention, consensus about how to best measure teacher effectiveness has yet to be established

* Corresponding author at: Temple University, 268 Ritter Annex, 1301 Cecil B. Moore Avenue, Philadelphia, PA 19122, United States of America.
E-mail address: lia.sandilos@temple.edu (L.E. Sandilos).

https://doi.org/10.1016/j.jsp.2019.05.004
Received 23 July 2018; Received in revised form 1 April 2019; Accepted 3 May 2019
Available online 20 May 2019
0022-4405/ © 2019 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.

(Newton, Darling-Hammond, Haertel, & Thomas, 2010). Although some agreement is noted in *which* practices are likely to be effective (Stronge et al., 2011), there is limited understanding of the specific tools and metrics that accurately and fairly capture these practices (Blazar, 2015) as well as how practices might differentially influence academic outcomes depending on the context or consistency of use of those practices. Moreover, school districts use a range of measures to assess effective teaching, and there is inconsistency in the ability of these measures to relate to important student educational outcomes (e.g., Hill, Kapitula, & Umland, 2011; Kane, McCaffrey, Miller, & Staiger, 2013). Thus, the purpose of the present study was to examine the convergent and predictive validity of several widely used measures of effective teaching practice across elementary and middle school grades and content areas (mathematics & English language arts [ELA]). In addition, we explored the influence of different metrics on student learning gains. Specifically, we examined the influence of both central tendency in scores (i.e., average score a teacher receives on a particular construct measured on multiple occasions) and variability in scores (i.e., spread of, or deviation in, scores a teacher receives on a construct across multiple occasions; Bedeian & Mossholder, 2000). These associations were examined using data from the landmark Measures of Effective Teaching (MET) Project (Bill & Melinda Gates Foundation, 2009-2011), the largest study of effective teaching conducted to date.

1. Elements of effective teaching practice

Existing research suggests that effective teaching can be described, in part, through multidimensional assessments of teachers' warmth, rigor, classroom management, and provision of instructional supports (e.g., Danielson Group, 2013; Hamre et al., 2013; Hattie, 2009; Marzano, 1998). These characteristics and practices are associated with positive academic and social-emotional student outcomes and are hypothesized to be effective across developmental levels and content areas (Hamre et al., 2013; Ruzek et al., 2016). Warmth reflects teachers' facilitation of a classroom environment in which positive and respectful interactions occur regularly (Reyes, Brackett, Rivers, White, & Salovey, 2012). Within a warm classroom environment, students feel more comfortable expressing themselves and taking academic risks (Ferguson, 2010; Rimm-Kaufman & Chiu, 2007). Teachers' skill at fostering warm and emotionally supportive classrooms is critical to high-quality instruction (National Research Council, 2004; Pianta & Hamre, 2009) and has been found to predict increased student motivation and academic engagement (Reyes et al., 2012; Ruzek et al., 2016).

In addition to warmth, a well-managed and academically rigorous classroom also supports student learning. Effective classroom management results in a productive environment in which learning time is maximized and behavioral expectations are clear (Woolfolk Hoy & Weinstein, 2011). Proactive behavior management strategies enhance students' self-regulation skills and support their academic achievement (e.g., Dudek, Reddy, & Lekwa, 2018; Ponitz, Rimm-Kaufman, Grimm, & Curby, 2009). Rigor involves setting high expectations for student effort as well as cultivating an environment in which learning is valued (Goddard, Sweetland, & Hoy, 2000; Lee, 2012; Lee & Smith, 1999). Prior research indicates that teachers' level of academic rigor and expression of high expectations directly influences students' achievement (Goddard et al., 2000; Rubie-Davies, 2007), particularly for students from historically marginalized groups (McKown & Weinstein, 2003; Sandilos, Rimm-Kaufman, & Cohen, 2017).

There are also a multitude of specific instructional strategies that promote learning across contexts, such as supporting learning through questioning techniques, providing specific academic and behavioral feedback (praise, correction), providing varying modalities for learning and different assessments of knowledge, and relating content to prior learning and real-world experience (Allington, 2002; Kamil et al., 2008; Kazemi & Stipek, 2009; Pianta & Hamre, 2009; Stronge et al., 2011). A key aspect of "effectiveness" has long been regarded as a teacher's ability to seamlessly and consistently implement instruction and positive management practices across lessons and adapt practices based on student needs and instructional context (Reddy, Glover, Kurz, & Elliott, 2019; Stronge, 2008). Presently, there are several published measures that strive to capture various aspects of the aforementioned elements of teaching effectiveness (warmth, rigor, classroom management, and instructional strategies) through classroom observations or student ratings (Ferguson & Danielson, 2013; Pianta & Hamre, 2009).

1.1. Methods of measurement

Teacher effectiveness is often evaluated using teacher inputs such as qualifications or aspects of instructional delivery (e.g., assessment of instructional practices), and outputs such as student achievement, or some composite of these elements (Stronge et al., 2011; Tyler, 2011). With regard to the use of inputs (e.g., qualifications, instructional practices) and outputs (e.g., gains in student learning), the most prominent evaluations of teacher effectiveness occur via classroom observation measures and value-added models (VAM). A VAM estimate is interpreted as a teacher's ability to facilitate student academic growth while accounting for a variety of student background variables and prior achievement. As of late, assessment efforts have also expanded to include student perspective ratings of teacher performance in classrooms (Downer, Stuhlman, Schweig, Martínez, & Ruzek, 2014; Polikoff, 2014).

1.1.1. Classroom observations

Classroom observation is the most common measurement approach to assessing instructional qualities and informing professional development plans. Observation tools assess content-neutral practices (i.e., universal core practices that cut across content areas) or discipline-specific practices (e.g., assessment of quality practices specific to a content area, such as mathematics). Two of the most widely used content-neutral instruments are the Classroom Assessment Scoring System (CLASS; Pianta, Hamre, & Mintz, 2012) and The Framework for Teaching (FFT; Danielson, 2007; Danielson Group, 2013). These measures are multi-dimensional and are used in preschool through 12th grade classrooms worldwide.

The CLASS is grounded in an ecological perspective emphasizing the influence of interactions in the classroom on students'

educational experiences and performance (Pianta & Hamre, 2009). The CLASS was initially developed for use in pre-kindergarten classrooms (Hamre, Pianta, Mashburn, & Downer, 2007) and has since been adapted for later grades. The CLASS captures the nature of teacher-child interactions by assessing teachers' efforts to foster a warm and positive classroom climate and use effective classroom management strategies such as proactively addressing behaviors and establishing productive routines. Additionally, the provision of specific instructional supports is assessed through teachers' use of constructive feedback and scaffolding, varied modalities for learning, modeling of novel vocabulary, and fostering of analytical thinking skills (Pianta et al., 2012). The CLASS has undergone rigorous empirical evaluation and is most frequently used in early childhood and elementary settings (e.g., Downer, Sabol, & Hamre, 2010; Hafen et al., 2015; Sandilos, Wollersheim Shervey, DiPerna, Lei, & Cheng, 2017), with a smaller body of research conducted at the secondary level (Allen, Pianta, Gregory, Mikami, & Lun, 2011). Direct relations between CLASS and student outcomes are generally moderate (Zaslow et al., 2016), spurring continued exploration of the ways in which the measurement of high-quality teacher-child interactions may relate to student achievement.

The FFT (Danielson Group, 2013) was designed to guide work around measuring and promoting teaching practices associated with desirable student outcomes. The FFT assesses aspects of the general classroom environment such as teachers' efforts to create a safe, respectful, and warm classroom climate and their skill at responding to student behavior in a manner that supports learning. With regard to specific instructional practices, the FFT examines an array of strategies including teachers' ability to communicate with students and their use of questions, discussion, and assessment. Like CLASS, emerging evidence supports modest relations between FFT scores and student outcomes (Kane, Taylor, Tyler, & Wooten, 2011; Sartain, Stoelinga, & Brown, 2011).

Though varied in the exact terminology used, the theoretical or conceptual overlap between the CLASS and FFT is particularly evident in their emphasis on evidence-based instructional strategies, use of effective classroom management practices, and the cultivation of a warm and emotionally supportive classroom climate.

1.1.2. Student ratings

Student perception of the instructional environment has the potential to yield valuable information about teachers' pedagogical practices and to facilitate understanding of the diversity of views students have about their classroom experience (Brock, Nishida, Chiong, Grimm, & Rimm-Kaufman, 2008; Kane & Staiger, 2012). Although student surveys are more common in secondary and higher education, the use of student rating scales in upper elementary grades is gaining popularity as a useful method for understanding teacher quality (e.g., Downer et al., 2014; Polikoff, 2014).

The Tripod rating scale (Ferguson, 2008), originally developed as part of a larger study aimed at harnessing student feedback to cultivate school improvement, is one example of a student perspective measure that is used in both upper elementary and secondary grades. The Tripod assesses seven categories that capture students' beliefs about the degree to which their teacher cares for them, the level of rigor in the classroom, and their teachers' control over behavioral issues. The scale also inquires about specific aspects of instruction, such as the teachers' ability to help students understand challenging content and connect new information to prior learning (Ferguson, 2010). Notably, the authors of the Tripod and FFT published a conceptual crosswalk demonstrating that the two measures theoretically capture similar constructs, which tap into warmth, classroom management, rigor, and instructional strategies (Ferguson & Danielson, 2013). The alignment between the Tripod and FFT ratings has the potential to offer a richer appraisal of teacher effectiveness using different sources of data (i.e., observer & student) and may generate meaningful information for teacher professional improvement.

1.1.3. Value-added models (VAMs)

In the last decade, estimating teacher and school effectiveness based on student standardized test gains has become common practice in research and school-based evaluation (Chetty et al., 2011; Hill et al., 2011). VAMs are algorithm-based measures of teachers' ability to facilitate student growth on standardized assessments, taking into account a large set of student covariates (e.g., gender, socio-economic status, prior achievement, disability status). VAMs are controversial because they depart from traditional use of teacher qualifications (e.g., years of experience, professional certification, degree attainment) to determine effectiveness, and because administrators and policy makers have used VAM scores to make high-stakes decisions (e.g., promotion, compensation, and dismissal of teachers; Hill, 2009; Paige, 2012). Some evidence links VAM scores to long-term student outcomes (e.g., graduation rates, college attendance, job salary; Chetty et al., 2011; Jackson, 2018). Yet, significant measurement flaws in VAM scores exist, such as the potential instability of scores from year to year (Braun, 2005). Despite the lack of empirical clarity on the use of VAM scores, they have strong practical relevance given that they are widely used as a part of a comprehensive teacher evaluation approach in more than half of the large districts in the U.S. and the use of these models continues to grow in prevalence (Steinberg & Donaldson, 2016). Thus, continued empirical research that examines how the measurement of unique and complementary constructs representing effective teaching contribute to higher VAM scores across grade levels and content areas is needed (Kane et al., 2013).

1.2. Considering developmental level and content area

As students progress through school, so too do the demands placed on them and on the adults charged with their education (Cleary & Chen, 2009; Midgley & Edelin, 1998). In the U.S., this development is frequently marked by three school transition points: the transition to kindergarten, to middle or junior high school, and to high school (Eccles, Lord, Roeser, Barber, et al., 1997; Rimm-Kaufman & Pianta, 2000). The transition from elementary to middle school is noted as a particularly impactful change in the students' learning experiences (Eccles et al., 1997). Academically, this transition is characterized by a shift from a supportive, mastery-based orientation typical of elementary schools to a performance-based orientation emphasizing academic productivity expectations, more

intensive teacher-directed instruction, normative comparison, and high-stakes outcomes (Cleary & Chen, 2009; Midgley & Edelin, 1998; Schunk & Miller, 2002; Zimmerman, 2002). These changes in the instructional environment can also alter students' perspectives of school (Gentry & Springer, 2002). As such, another critical component of quality instruction is the continued exploration of teaching practices that enhance the learning experience and outcomes of students at various developmental levels.

Similarly, the instructional content area also influences the assessment of effective teaching. Content-neutral measures, such as those examined in the present study, capture effective practice across academic subjects and are intended to offer school districts a fair and balanced assessment for all teachers (Hill & Grossman, 2013). Yet, certain disciplines may afford greater opportunities for some content-neutral practices to be observed or rated. For example, there may be more opportunities for on-going dialogue during an ELA lesson as compared to a mathematics lesson. Moreover, as Cohen and Goldhaber (2016) note, it is possible that certain seemingly content-neutral practices will actually serve as "essential components" of teaching in particular content domains and, in turn, result in larger academic gains in certain subjects than others. The MET Project provides a rich landscape for this nuanced examination of various content-neutral measures of effective teaching across developmental levels and content areas.

1.3. Building on the MET findings

Prior to the MET project, no evaluation of teaching quality had ever included such as vast number of teachers and instructional quality measures. As a particularly distinctive contribution of this study, MET researchers also attempted to reduce sorting bias (Blazer, 2015) by randomly assigning class rosters to teachers in the second year of the study. In a large report of MET findings produced by Kane and colleagues, the authors describe the individual and collective impact of teaching effectiveness measures on student learning gains. Results from the larger study indicated that there was a generally moderate positive relation between quality teaching indicators and VAM scores across grade levels. MET researchers concluded that multiple measures provide a more robust estimate of teaching quality compared to a singular assessment, supporting the value of understanding convergence across measures of effectiveness (Kane et al., 2013). The current study uses this seminal data set and builds on the initial descriptive MET findings in two unique ways: exploring variability in scores and examining theoretical groupings of constructs.

1.3.1. Variability in scores

First, within existing MET research, observations and student ratings were studied largely in terms of mean scores averaged across multiple data collection points to represent teaching effectiveness (Kane et al., 2013). Emerging research suggests that teachers tend to show variability in their practices across multiple observations (Cohen et al., 2018; Praetorius, Pauli, Reusser, Rakoczy, & Klieme, 2014). However, to date, we know far less about how this variability relates to student outcomes across contexts (Patrick & Mantzicopoulos, 2016), or whether variability in scores across measures is potentially an indicator of lower-quality instruction. One possibility is that variability in scores may be due to external factors such as changes in student learning needs, varied lesson formats, or rater differences (e.g., Cohen, Raudenbush, & Ball, 2003; Mantzicopoulos, French, Patrick, Watson, & Ahn, 2018; Praetorius et al., 2014). Yet another possibility is that variability in scores may reflect teachers' differential treatment of students (McKown & Weinstein, 2003) or inconsistent use of quality practices, negatively influencing the overall learning environment for students (Curby, Grimm, & Pianta, 2010). With regard to assessing student ratings, some initial evidence indicates that high variability in student ratings within a single classroom negatively relates to student achievement gains, and thus, may be indicative of less optimal instructional practice (Sandilos, Rimm-Kaufman, et al., 2017).

Although many educational researchers have argued that a single observation or rating scale is insufficient to understand the quality of teachers' instruction, the use of a singular measure to evaluate teaching effectiveness remains common place in routine educational practice and research (Brophy, 2006; Hill, Charalambous, & Kraft, 2012). Evidence that variability in teachers' quality scores over time influences student outcomes would provide additional support for the value of examining data from multiple time points and sources. As such, exploring the relation between variability in ratings and observations of teaching quality and student learning gains may offer beneficial information for improving the process of teacher evaluation.

1.3.2. Theoretical groupings

Second, much of the initial work on MET has focused on examining measures in isolation or using one large composite of effectiveness (e.g., combining FFT and CLASS scores into one observation total score metric; Kane et al., 2013). The present investigation extends these analyses and findings by incorporating subscales and dimensions across effectiveness measures to create theoretically grounded groupings based on the larger teacher effectiveness literature (i.e., *warmth*, *classroom management*, *rigor*, and *instructional strategies*). Given the growing emphasis on the use of multiple measures of effectiveness in teacher evaluation (ESSA, 2015; Kane et al., 2013; Reddy, Kettler, & Kurz, 2015), examination of these theoretical groupings provides information about the ways in which conceptually similar constructs across assessments might differentially relate to student learning gains.

To this end, the present investigation examined the convergent and predictive validity of three widely used measures (CLASS, FFT, and Tripod) of teacher effectiveness to extend existing MET findings and establish a more nuanced understanding of the interrelatedness of these measures. Specifically, this study contributes to existing literature by examining instructional quality scores, theoretically grouped, by developmental level (elementary vs. secondary grades) and content area (mathematics vs. ELA), and across score type (mean score vs. variability in scores). Research questions addressed were: (1) What is the convergent validity of three measures of teaching effectiveness (CLASS, FFT, and Tripod) in elementary (grades 4–5) and secondary classrooms (grades 6–8) and across two different content areas (mathematics vs. ELA)? (2) Across elementary and secondary grades, do average scores versus variability in scores on these theoretically grouped measures of teaching effectiveness predict teachers' VAM scores, a high-stakes

measure of teaching effectiveness in mathematics and ELA?

Although this study was largely exploratory in nature, several overarching hypotheses were generated by drawing from prior research. Based on inherent differences in assessment design (i.e., method, metrics and purpose), greater convergence was expected between CLASS and FFT across developmental levels and content areas, as compared to the Tripod. With regard to predicting VAM scores, it was hypothesized that score variability would demonstrate a negative relation to achievement in comparison to mean scores across contexts (Sandilos, Rimm-Kaufman, et al., 2017).

2. Method

2.1. Participants

The current study used data from the second year (2010 − 2011) of the MET project, a large-scale observational study of classroom instruction funded by the Bill and Melinda Gates Foundation. Participants in this study included fourth- through eighth-grade teachers from five large districts across the United States. Secondary teachers were almost exclusively content-area specialists (i.e., teacher who only taught mathematics or ELA, but not both). The elementary sample included a mix of teachers who taught in departmentalized instructional settings (i.e., teachers assigned to instruct in specific content areas) and teachers who taught multiple content areas (referred to as "specialists" and "generalists" in the MET Project, respectively; White & Rowan, 2013). To maintain consistency and independence across samples, this study analyzed data from specialist teachers across elementary and secondary grades. Additionally, the samples are also limited to elementary and secondary teachers who have video observation data because only a subset of teachers participating in the MET Project consented to have their video data analyzed (White & Rowan, 2013).

The 171 fourth- and fifth-grade specialist elementary teachers with either mathematics (n = 70) or ELA (n = 101) data came from 34 schools across the five districts and were largely female (86%). Their racial/ethnic makeup was White (70%), African American (20%), Hispanic (7%), and other races (3%), and approximately one-quarter (26%) had a master's degree or higher. Teachers had an average of 8 years of experience teaching in their district (< 1–35 years, SD = 6.57). Regarding classroom composition, on average, approximately 50% of students in classrooms were female, 12% were English language learners, and 12% received special education services. Average racial/ethnic demographics in classrooms were as follows: White (40%), African American (27%), Latino (24%), Asian (4%), and other races (5%). No statistically significant differences in teacher or classroom characteristics were found between the sample of ELA elementary teachers and the sample of mathematics elementary teachers, with the exception of gender such that more male teachers instructed in mathematics as compared to ELA (χ^2 = 6.50, p = .01).

The 571 secondary teachers in sixth through eighth grade with either mathematics (n = 291) or ELA (n = 280) data came from 86 schools in six districts and were majority female (78%). Their racial/ethnic composition was White (55%), African American (37%), Latino (6%), and other races (3%). Across the sample, approximately 30% of teachers had a master's degree or higher and had an average of 8 years of experience teaching in their district (< 1–41 years, SD = 7.43). On average, approximately 50% of students in classrooms were female, 14% were English language learners, and 7% received special education services. Average racial/ethnic demographics in classrooms were as follows: White (24%), African American (29%), Latino (37%), Asian (7%), and other races (3%). No statistically significant differences in teacher or classroom characteristics were found between the sample of ELA secondary teachers and the sample of mathematics secondary teachers, with the exception of gender, with more male teachers instructing in mathematics as compared to ELA (χ^2 = 7.34, p = .01).

2.2. Measures

2.2.1. Classroom assessment scoring system (CLASS – upper elementary & secondary)

The CLASS observation system (Pianta, Hamre, & Mintz, 2012) is designed to assess teacher-child interactions in the classroom environment. Trained observers rate teachers on 11 dimensions which are averaged to form three primary domains: Emotional Support, Classroom Organization, and Instructional Support (see Pianta, Hamre, & Mintz, 2012). The Emotional Support domain consists of Positive Climate, Teacher Sensitivity, and Regard for Student Perspectives dimensions. The Classroom Organization domain consists of Behavior Management, Productivity, and Negative Climate dimensions. The Instructional Support domain consists of Instructional Learning Formats, Content Understanding, Analysis and Problem Solving, Quality of Feedback, and Instructional Dialogue dimensions.

In the MET Project, teachers submitted four videotaped classroom lessons in either mathematics or ELA that were then coded by trained raters using the CLASS. The CLASS dimensions were scored in 15-min observation cycles using a 7-point scale ranging from *Low* (1–2), *Middle* (3–5), to *High* (6–7). In a given observation cycle, a teacher received one score on each dimension. Four 30-min classroom lessons, with two 15-min observation cycles per lesson, were scored in a given content area, resulting in a total of eight CLASS observation cycles coded for a teacher in either ELA or mathematics. Dimension scores within observation cycles were aggregated to form teachers' eight domain scores for that content area, which were then used to compute an overall average score or a coefficient of variation for each domain.

Inter-rater agreement within 1 point (i.e., adjacent agreement) was identified as an acceptable indicator of reliability for all the CLASS dimensions for the MET Project. Inter-observer adjacent agreement for the MET Project ranged from 68 to 86% (Pianta, Hamre, & Mintz, 2012). Prior factor analyses using MET data confirmed the three-factor structure of CLASS in elementary and secondary grades (Pianta, Hamre, & Mintz, 2012).

2.2.2. Framework for Teaching (FFT)

The FFT (Danielson Group, 2013) emphasizes the intellectual engagement of students through instructional practices espoused by a Constructivist Theory of learning. FFT ratings yield four major domain scores: Planning and Preparation, Classroom Environment, Instruction, and Professional Responsibilities. Only dimensions within the Classroom Environment and Instruction domains were used in the MET Project because they could be coded via videotaped observations. The Classroom Environment domain consists of four dimensions: Creating an Environment of Respect and Rapport, Establishing a Culture for Learning, Managing Classroom Procedures, and Managing Student Behavior. The four dimensions of Instruction consisted of Communicating with Students, Using Questioning and Discussion Techniques, Engaging Students in Learning, and Using Assessment in Instruction (Danielson Group, 2013).

The same four videotaped lessons that were rated using the CLASS were also rated using the FFT. For each lesson, 25-min observations were conducted; raters observed the first 15 min of the tape and then 10 additional min at the 25 to 35 min mark. Each of the eight dimensions received one score per lesson using a range of *Unsatisfactory* (1), *Basic* (2), *Proficient* (3), and *Distinguished* (4). The four scores for each dimension in a given content area were then used to create an overall average score or a coefficient of variation.

Inter-observer reliability of MET FFT observations was above 0.60 (Polikoff & Porter, 2014). Prior examination of the two domains used in the MET Project, Classroom Environment and Instruction, indicated that more than half (67%) of the variance in scores was attributable to teachers (Kane & Staiger, 2012).

2.2.3. Tripod 7Cs Student Perceptions Survey

The Tripod 7Cs survey (Ferguson, 2008) assesses students' perspectives of instructional quality and classroom processes using seven scales. All items are scored from 1 (*Totally Untrue*) to 5 (*Totally True*), and each scale is calculated by taking the mean of the item-level scores. The scales measured students' perceptions of their teacher during either ELA or mathematics instruction. Tripod surveys were administered to students in participating teachers' classrooms once per year. Across both elementary (E) and secondary (S) measures, a total of 36 items on seven scales were surveyed: Care (E = 7 items, S = 3 items) inquires about the degree to which students feel their teacher acts in a supportive and caring manner toward them; Control (E = 4 items, S = 7 items) measures a teacher's management of classroom behavior; Clarify (E = 8 items, S = 5 items) measures the teacher's skill at helping students gain a better understanding of difficult content; Challenge (E = 4 items, S = 8 items) measures the teacher's expectations for hard work and persistence; Captivate (E = 4 items, S = 4 items) measures the teacher's ability to maintain student engagement and interest in lessons; Confer (E = 7 items, S = 5 items) assesses a teacher's skill at eliciting students' perspectives during lessons; Consolidate (E = 2 items, S = 4 items) measures a teacher's skill at connecting learning topics coherently and checking for understanding (White & Rowan, 2013). On average, 19 students in a classroom completed Tripod surveys in elementary classrooms and 20 students completed Tripod surveys in secondary classrooms.

All internal consistency reliabilities of the seven subscales across elementary and secondary surveys were above 0.60. With regard to psychometric properties, Ferguson (2012) notes that the Tripod was developed through consultation with educators and item analyses were conducted on subscales. Although the Tripod is theoretically established and hypothesized to have two overarching constructs, *academic press* (i.e., rigor) and *support* (Ferguson, 2010), several studies of MET data have yielded differing factor structures in the secondary grades (e.g., Kuhfled, 2017; Wallace, Kelcey, & Ruzek, 2016). The present study examined the seven scales individually using either the overall mean score for each scale averaged across all student respondents or the coefficient of variation calculated across all student responses.

2.2.4. VAM scores

A VAM score is an estimate of teacher effectiveness frequently used as a high-stakes interpretation of a teacher's ability to facilitate growth in an academic content area, while accounting for the composition of students in a classroom. In the MET Project, an average VAM score in mathematics or ELA was computed for each teacher using the state standardized test as the measure of achievement. To create VAM scores in each content area, MET researchers used an algorithm that included student assessment data and background information. The VAM estimate was composed of (a) students' state test scores from the current year, (b) their state test scores from the year prior, (c) the average score on the state test in the students' classroom the year prior, (d) student demographic variables (i.e., ethnicity, ELL status, age, gender, special education status, gifted status, and receipt of free or reduced lunch) at the individual level and the classroom average (White & Rowan, 2013). For elementary VAMs, an average of 18 students in a classroom was included in math models and 20 students were included in ELA models. For secondary VAMs, an average of 22 students in a classroom was included in either mathematics or ELA models.

Standardized tests in mathematics and ELA varied based on the state in which each district was located. Across states, the standardized tests were typically constructed in multiple-choice formats. Within the MET data set, rank-based z-scores for the state standardized test measures are provided. The interpretation of state test scores converted into VAMs is as follows: (a) a standardized score of zero indicates that students in the teacher's class are performing as expected on the state test given prior achievement and background data, (b) a negative score indicates that students are performing lower than expected on the state test, and (c) a positive score indicates that students are performing higher than expected.

2.3. Procedure

To recruit districts, an opportunity-sampling procedure was employed in which MET researchers sought out large, urban districts

that had previously worked with the Gates Foundation in some capacity. This resulted in six large districts across the country agreeing to participate in the MET Project; however, the sample is not considered to be nationally representative. One district was not included in the elementary sample because it did not have participating fourth and fifth grade classrooms. Exclusions to the study included alternative schools (special needs, vocational schools, etc.), and teachers who engaged in team- or co-teaching situations in which it would be difficult to link students to one teacher.

Once districts were recruited, school principals were asked to identify eligible teachers. With each participating school, eligible teachers had to be part of a grade level or content area specialty that had at least two other teachers to form an "exchange group" in which students could be randomly assigned to one of three teachers in that grade level in the second year of the study. The within-grade randomization was conducted to reduce selection bias associated with the systematic sorting of students into certain classrooms, which in turn, influences measures of teaching effectiveness. The use of this randomization procedure, although not perfect, was a unique strength of the data collection in the second year of the MET Project (White & Rowan, 2013).

With regard to data collection timeline, the state standardized testing occurred between March and June 2011. Classroom observations and student surveys were conducted between February and June 2011. The FFT and CLASS observation rubrics were used to score the same set of four submitted videos (approximately one video per month) for each teacher. Although the Tripod represents an assessment of instructional quality from the perspective of multiple students, CLASS and FFT scores come from multiple lessons collected over the four months. Raters, who were randomly assigned to videos, went through a full training and certification in each measure. They completed a daily calibration on each assessment to offset rater drift, then watched videos and coded immediately after watching (White & Rowan, 2013).

To participate in the study, teachers agreed to a multitude of data collection tasks over two years including videotaped observations, several lengthy surveys, a content-area assessment, and a willingness to participate in within-grade level randomization in the second year. Given the demands on data collection, each teacher was offered $1500 as incentive for participation. However, teachers did not receive all $1500 at one time. Instead, they received $1000 at the beginning of the study and $500 at the end of the study. Each school also received $1500 over the course of the study, along with an additional $500 per year to pay for an in-house project coordinator.

MET data are restricted because the study files contain sensitive and potentially identifying information about school districts. The findings presented in this paper follow the secure data reporting policies outlined by MET (Bill & Melinda Gates Foundation, 2009-2011) and the Inter-University Consortium for Political and Social Research.

2.4. Data analysis

This study included specialist teachers in grades 4–8 who participated in year 2 of the MET project. The full sample of specialist teachers consisted of 234 elementary teachers and 832 secondary teachers. However, a portion of teachers in the elementary and secondary samples were excluded from the present study because they did not provide consent for their video observation data to be analyzed by researchers (White & Rowan, 2013). Differences between teachers included in the final samples (Elementary = 171, Secondary = 571) and teachers who were removed due to a lack of video consent were examined using *t*-tests and chi-square difference tests. No significant differences between these groups of teachers were found for any demographic variables (i.e., years of experience, master's degree, gender, and race/ethnicity).

To explore convergence across measures, Pearson correlations were conducted in SPSS version 25 to determine the relations among the CLASS, FFT, and Tripod. Correlations, conducted by developmental level (elementary and secondary grades) and by content area (mathematics or ELA), yielded four separate correlation matrices. Within the larger MET Project, grades 4 through 5 were coded as "elementary" and grades 6 through 8 were coded as "middle school" to reflect the general structure of grade levels in the United States (White & Rowan, 2013). For each measure of effectiveness, scores were coded as reflecting either mathematics or ELA instruction.

To examine how average scores versus variability in scores on CLASS, FFT, and Tripod relate to teachers' VAM scores, multiple regression analyses were conducted in M*plus* version 8. All regression models were estimated using clustered standard errors to account for the nesting of teachers within schools (i.e., Type = Complex). Full information maximum likelihood (FIML) was used to account for missing data. In addition to examining central tendency (i.e., mean scores), regression analyses also explored variability in teachers' effectiveness scores. Specifically, a coefficient of variation was computed, which is a statistical technique used to measure variation in an observed variable through the ratio of the standard deviation to the mean (Bedeian & Mossholder, 2000). This metric was chosen in order to compare score variation across measures. Regression models using mean scores and coefficient of variation scores were conducted to explore relations of CLASS, FFT, and Tripod to VAMs in ELA and mathematics.

A priori power analyses using G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009) indicated that samples of teachers across elementary (Math $n = 70$, ELA $n = 101$) and middle school grades (Math $n = 291$, ELA $n = 280$) were sufficiently large to detect a medium effect ($f^2 = 0.15$–0.26) with at least 80% power. Similar effect sizes have been identified in prior research, including other MET studies, examining gains in state test scores based on indicators of teaching quality after controlling for student characteristics and prior achievement (e.g., Allen et al., 2013; Darling-Hammond, 2000; Kane et al., 2013).

The regression models were estimated by grouping conceptually similar scales/dimensions/domains across measures. The grouping of variables occurred in several stages. First, the authors consulted a chapter by Ferguson and Danielson (2013), which grouped the Tripod and FFT subscales based on similarities across constructs. Next, because CLASS also was included in these analyses, the authors, who were trained on the CLASS and FFT, mapped the CLASS domains onto Ferguson and Danielson's existing conceptual crosswalk (see Table 1). To do so, a consensus-building task (Hennessy et al., 2016) was used whereby the current authors

Table 1
Theoretical grouping among CLASS, FFT and Tripod.

Constructs	CLASS	FFT	Tripod
Warmth	• Emotional Support (Positive Climate, Regard for Student Perspectives, Teacher Sensitivity)	• Creating an Environment of Respect and Rapport	• Care
Classroom management	• Classroom Organization (Behavior Management, Productivity, Negative Climate)	• Managing Classroom Procedures • Managing Student Behavior	• Control
Rigor	(no close alignment)	• Establishing a Culture for Learning	• Challenge
Instructional Strategies	• Instructional Support (Instructional Learning Formats, Content Understanding, Analysis and Problem Solving, Quality of Feedback, Instructional Dialogue)	• Using Questioning and Discussion Techniques • Using Assessment in Instruction • Engaging Students in Learning • Communicating with Students	• Confer • Captivate • Clarify • Consolidate

Note. This table is based on conceptual similarities proposed by Ferguson and Danielson (2013) with the addition of CLASS domains incorporated by the authors of the present study through a structured Q-Sort task.

discussed the individual scales/dimensions and the overlapping content across measures. Four key components of effective teaching emerged from the related constructs across measures: *warmth*, *classroom management*, *rigor*, and *instructional strategies*. The authors produced definitions for these four theoretical components using the larger effective teaching literature (e.g., Danielson Group, 2013; Dudek et al., 2018; Hamre et al., 2013; Hattie, 2009; Marzano, 1998; National Research Council, 2004; Pianta & Hamre, 2009; Ruzek et al., 2016). Finally, to confirm the theoretical groupings, the authors engaged in a structured Q-sort activity (Rimm-Kaufman, Storm, Sawyer, Pianta, & LaParo, 2006) in which each author individually sorted the three CLASS domains, eight FFT dimensions, and seven Tripod subscales into one of the four overarching components of effective teaching based on the specific definitions of subscales provided in published versions of the measures (Ferguson, 2008; The Danielson Group, 2013; Pianta, Hamre, & Mintz, 2012). The results of the Q-sort activity resulted in 100% agreement among the co-authors and also reflected the conceptual crosswalk proposed by the authors of the Tripod and FFT (i.e., Ferguson & Danielson, 2013).

For regression analyses, measure scales/dimensions/domains were grouped based on the agreed-upon theoretical groupings to determine if certain subcomponents of a group had more predictive power than others and explore trends across the groups. Each of the theoretical groupings of rating subscales and observation dimensions/domains (i.e., *warmth*, *classroom management*, *rigor*, and *instructional strategies*) was entered into a separate regression analysis to predict VAMs.

3. Results

Means, ranges, and standard deviations for the CLASS, FFT, and Tripod measures are reported by developmental level and content area in Table 2.

Table 2
Descriptive Statistics for CLASS, FFT, and Tripod Constructs in Mathematics and ELA across Elementary and Middle School Grades.

		Elementary Math		Elementary ELA		Middle school Math		Middle school ELA	
		Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range
CLASS	Emotional support	4.55 (0.35)	3.63–5.75	4.79 (0.37)	3.94–5.59	4.11 (0.52)	2.46–5.22	4.23 (0.51)	2.56–5.41
	Classroom Organization	5.49 (0.31)	4.46–6.00	5.56 (0.30)	4.46–6.13	5.06 (0.51)	3.06–6.04	5.09 (0.52)	2.60–6.10
	Instructional support	3.74 (0.40)	2.94–4.72	3.86 (0.43)	2.78–4.72	3.14 (0.54)	1.83–4.69	3.15 (0.57)	1.63–5.25
FFT	Respect & rapport	2.83 (0.30)	1.75–3.40	2.94 (0.22)	2.25–3.50	2.53 (0.42)	1.17–3.50	2.62 (0.42)	1.13–3.50
	Communicate with students	2.71 (0.32)	2.13–3.25	2.86 (0.31)	2.00–3.75	2.49 (0.36)	1.33–3.50	2.57 (0.36)	1.63–3.75
	Culture for learning	2.63 (0.31)	2.00–3.38	2.78 (0.28)	2.00–3.25	2.39 (0.39)	1.00–3.75	2.42 (0.41)	1.13–3.50
	Engage students in learning	2.53 (0.31)	2.00–3.13	2.72 (0.32)	1.75–3.50	2.31 (0.39)	1.00–3.25	2.40 (0.40)	1.25–3.50
	Classroom procedures	2.81 (0.33)	2.00–3.50	2.89 (0.25)	2.00–3.50	2.57 (0.41)	1.00–3.50	2.62 (0.40)	1.38–3.50
	Student behavior	2.92 (0.30)	1.75–3.50	2.96 (0.23)	2.00–3.25	2.65 (0.41)	1.00–3.50	2.71 (0.39)	1.25–3.50
	Assessments in instruction	2.43 (0.31)	1.75–3.00	2.50 (0.33)	1.75–3.25	2.24 (0.39)	1.20–3.25	2.19 (0.39)	1.25–3.50
	Questioning & discussion	2.28 (0.31)	1.50–3.00	2.46 (0.30)	1.75–3.25	2.10 (0.39)	1.20–3.25	2.14 (0.41)	1.00–3.25
Tripod	Care	4.13 (0.35)	2.54–4.64	4.17 (0.36)	2.49–4.80	3.42 (0.58)	1.70–4.66	3.54 (0.57)	1.81–4.93
	Control	3.61 (0.34)	2.70–4.37	3.64 (0.35)	2.41–4.56	3.36 (0.57)	1.91–4.70	3.40 (0.55)	1.99–4.78
	Clarify	4.22 (0.27)	2.80–4.70	4.19 (0.24)	3.06–4.79	3.78 (0.45)	2.22–4.97	3.85 (0.44)	2.31–5.00
	Challenge	4.14 (0.30)	3.08–4.58	4.25 (0.26)	3.69–4.80	4.13 (0.34)	3.05–4.97	4.12 (0.36)	2.79–4.97
	Captivate	3.59 (0.36)	2.52–4.56	3.60 (0.37)	2.69–4.72	3.54 (0.60)	1.80–4.98	3.65 (0.57)	1.74–4.94
	Confer	4.23 (0.26)	2.92–4.81	4.19 (0.27)	3.06–4.89	3.44 (0.43)	2.19–4.57	3.59 (0.46)	2.04–5.00
	Consolidate	3.82 (0.40)	2.29–4.65	3.88 (0.38)	2.58–4.82	5.02 (0.57)	3.12–6.14	5.14 (0.59)	3.14–6.67

Table 3
Correlations among CLASS, FFT, and Tripod in Mathematics in elementary grades (4–5, $n = 70$).

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
CLASS	1. Emotional Support	–	0.59** ***	0.79**	0.41**	0.42**	0.40**	0.41**	0.27*	0.23**	0.32**	0.31*	0.02	0.09	0.05	−0.18	−0.18	0.02	−0.12
	2. Classroom Organization		–	0.53**	0.57**	0.39**	0.54**	0.45**	0.61**	0.56**	0.44**	0.32*	−0.17	−0.08	−0.15	−0.11	−0.28*	−0.11	−0.18
	3. Instructional support			–	0.21	0.31**	0.30**	0.44**	0.21	0.09	0.31*	0.22	−0.05	0.04	−0.08	−0.28*	−0.11	−0.04	−0.12
FFT	4. Respect and rapport				–	0.57**	0.44**	0.37**	0.62**	0.65**	0.45**	0.48**	−0.07	0.03	−0.04	−0.14	−0.16	−0.10	−0.16
	5. Communicate with students					–	0.65**	0.58**	0.54**	0.45**	0.56**	0.63**	0.09	−0.02	−0.02	−0.22	−0.17	0.01	0.02
	6. Culture for learning						–	0.69**	0.50**	0.49**	0.64**	0.52**	−0.02	−0.05	−0.05	−0.08	0.23	0.02	0.02
	7. Engage students in learning							–	0.52**	0.42**	0.62**	0.62**	−0.05	−0.23	−0.08	−0.17	−0.15	−0.08	−0.12
	8. Classroom procedures								–	0.66**	0.53**	0.47**	−0.17	−0.12	−0.16	−0.02	−0.21	−0.16	−0.20
	9. Student behavior									–	0.51**	0.48**	−0.05	0.02	−0.13	−0.12	−0.32**	−0.13	−0.01
	10. Assessments in instruction										–	0.64**	−0.12	−0.14	−0.11	−0.15	−0.14	−0.14	−0.05
	11. Questioning & Discussion											–	0.05	−0.07	0.03	−0.16	−0.17	0.00	0.01
Tripod	12. Care												–	0.45**	0.80**	0.45**	0.55**	0.77**	0.73**
	13. Control													–	0.53**	0.32**	0.28**	0.46**	0.44**
	14. Clarify														–	0.58**	0.53**	0.82**	0.64**
	15. Challenge															–	0.42**	0.57**	0.46**
	16. Captivate																–	0.48**	0.50**
	17. Confer																	–	0.69**
	18. Consolidate																		–

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

Table 4
Correlations among CLASS, FFT, and Tripod in ELA in elementary grades (4–5, n = 101).

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
CLASS	1. Emotional support	–	0.65*****	0.82**	0.34**	0.51**	0.49**	0.52**	0.19	0.20*	0.55**	0.45**	0.10	0.22*	0.04	0.02	0.03	−0.02	−0.05
	2. Classroom organization		–	0.63**	0.42**	0.55**	0.37**	0.47**	0.43**	0.54**	0.53**	0.38**	0.08	0.29*	0.07	0.14	0.03	−0.03	−0.03
	3. Instructional support			–	0.27**	0.53**	0.37**	0.53**	0.19	0.25*	0.53**	0.45**	0.13	0.24*	0.04	0.07	0.04	−0.01	−0.07
FFT	4. Respect and rapport				–	0.49**	0.39**	0.44**	0.54**	0.51**	0.40**	0.42**	−0.06	0.15	−0.11	−0.13	−0.07	−0.14	−0.20*
	5. Communicate with students					–	0.60**	0.66**	0.51**	0.55**	0.64**	0.66**	0.07	0.13	0.03	0.11	0.03	0.01	−0.03
	6. Culture for learning						–	0.79**	0.31**	0.34**	0.65**	0.63**	0.10	0.15	0.03	0.15	0.04	−0.04	−0.12
	7. Engage students in learning							–	0.43**	0.43**	0.73**	0.64**	−0.01	0.11	−0.10	0.03	−0.08	−0.09	−0.10
	8. Classroom procedures								–	0.67**	0.39**	0.38**	−0.11	0.07	−0.06	0.01	−0.16	−0.14	−0.16
	9. Student behavior									–	0.44**	0.40**	−0.16	0.09	−0.08	0.12	−0.09	−0.14	−0.13
	10. Assessments in instruction										–	0.69**	0.03	0.16	−0.04	0.04	0.07	−0.01	−0.07
	11. Questioning & discussion											–	0.05	0.15	−0.05	0.03	0.00	−0.02	−0.14
Tripod	12. Care												–	0.53**	0.86**	0.30**	0.60**	0.78**	0.59**
	13. Control													–	0.63**	0.35**	0.36**	0.39**	0.25*
	14. Clarify														–	0.50**	0.67**	0.80**	0.69**
	15. Challenge															–	0.37**	0.40**	0.43**
	16. Captivate																–	0.67**	0.62**
	17. Confer																	–	0.68**
	18. Consolidate																		–

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

Table 5
Correlations among CLASS, FFT, and Tripod in Mathematics in Middle School Grades (6–8, n = 291).*, ***

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
CLASS	1. Emotional Support	–	0.72**	0.85**	0.59**	0.55**	0.60**	0.59**	0.48**	0.47**	0.52**	0.54**	0.41**	0.38**	0.39**	0.47**	0.40**	0.47**	0.44**
	2. Classroom organization		–	0.64**	0.64**	0.56**	0.58**	0.55**	0.69**	0.69**	0.43**	0.48**	0.29**	0.42**	0.35**	0.43**	0.31**	0.37**	0.39**
	3. Instructional support			–	0.46**	0.51**	0.51**	0.54**	0.43**	0.38**	0.56**	0.56**	0.34**	0.24**	0.33**	0.40**	0.31**	0.39**	0.38**
FFT	4. Respect and rapport				–	0.68**	0.72**	0.66**	0.67**	0.75**	0.54**	0.60**	0.25**	0.32**	0.26**	0.36**	0.25**	0.30**	0.31**
	5. Communicate with students					–	0.68**	0.68**	0.65**	0.58**	0.63**	0.67**	0.19**	0.22**	0.20**	0.30**	0.20**	0.26**	0.27**
	6. Culture for learning						–	0.81**	0.62**	0.65**	0.65**	0.67**	0.29**	0.33**	0.31**	0.41**	0.31**	0.34**	0.34**
	7. Engage students in learning							–	0.57**	0.58**	0.69**	0.69**	0.30**	0.33**	0.30**	0.41**	0.30**	0.33**	0.35**
	8. Classroom procedures								–	0.75**	0.48**	0.58**	0.15**	0.32**	0.22**	0.31**	0.17**	0.20**	0.27**
	9. Student behavior									–	0.43**	0.50**	0.18**	0.41**	0.26**	0.37**	0.21**	0.26**	0.28**
	10. Assessments in instruction										–	0.69**	0.26**	0.24**	0.24**	0.35**	0.26**	0.30**	0.29**
	11. Questioning & discussion											–	0.27**	0.23**	0.26**	0.34**	0.25**	0.33**	0.34**
Tripod	12. Care												–	0.51**	0.87**	0.80**	0.88**	0.87**	0.88**
	13. Control													–	0.59**	0.62**	0.59**	0.58**	0.59**
	14. Clarify														–	0.87**	0.92**	0.86**	0.91**
	15. Challenge															–	0.81**	0.84**	0.88**
	16. Captivate																–	0.85**	0.87**
	17. Confer																	–	0.86**
	18. Consolidate																		–

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

Table 6
Correlations among CLASS, FFT, and Tripod in ELA in middle school grades (6–8, n = 280).

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
CLASS	1. Emotional support	–	0.69** ***	0.76**	0.60**	0.54**	0.62**	0.60**	0.49**	0.48**	0.55**	0.56**	0.29**	0.33**	0.24**	0.27**	0.32**	0.27**	0.22**
	2. Classroom organization		–	0.70**	0.65**	0.53**	0.61**	0.59**	0.68**	0.66**	0.51**	0.53**	0.22**	0.38**	0.20**	0.27**	0.21**	0.22**	0.16**
	3. Instructional support			–	0.53**	0.56**	0.55**	0.59**	0.51**	0.43**	0.59**	0.59**	0.27**	0.23**	0.24**	0.31**	0.29**	0.25**	0.22**
FFT	4. Respect and rapport				–	0.68**	0.68**	0.61**	0.70**	0.78**	0.58**	0.57**	0.21**	0.32**	0.19**	0.22**	0.22**	0.19**	0.18**
	5. Communicate with students					–	0.71**	0.70**	0.58**	0.60**	0.67**	0.67**	0.24**	0.32**	0.23**	0.27**	0.28**	0.22**	0.24**
	6. Culture for learning						–	0.78**	0.62**	0.66**	0.69**	0.73**	0.21**	0.35**	0.19**	0.25**	0.29**	0.19**	0.21**
	7. Engage students in learning							–	0.62**	0.58**	0.75**	0.75**	0.19**	0.32**	0.19**	0.23**	0.28**	0.18**	0.19**
	8. Classroom procedures								–	0.75**	0.54**	0.53**	0.15**	0.33**	0.13**	0.21**	0.12*	0.17**	0.11
	9. Student behavior									–	0.54**	0.52**	0.17**	0.41**	0.19**	0.24**	0.19**	0.18**	0.18**
	10. Assessments in instruction										–	0.73**	0.20**	0.25**	0.19**	0.24**	0.25**	0.19**	0.19**
	11. Questioning & discussion											–	0.21**	0.30**	0.21**	0.27**	0.28**	0.18**	20**
Tripod	12. Care												–	0.51**	0.89**	0.84**	0.87**	0.86**	0.89**
	13. Control													–	0.60**	0.57**	0.57**	0.63**	0.55**
	14. Clarify														–	0.88**	0.90**	0.85**	0.90**
	15. Challenge															–	0.81**	0.80**	0.88**
	16. Captivate																–	0.82**	0.84**
	17. Confer																	–	0.85**
	18. Consolidate																		–

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

Table 7
Main effects of Tripod, CLASS, and FFT constructs to predict teachers' value-added scores in Mathematics and ELA across elementary and middle school grades.

Construct	Elementary VAMs (Mathematics n = 70, ELA n = 101)				Middle School VAMs (Mathematics n = 291, ELA n = 280)			
	Model 1 Math (mean)	Model 2 Math (coef. var)	Model 1 ELA (mean)	Model 2 ELA (coef. var)	Model 1 Math (mean)	Model 2 Math (coef. var)	Model 1 ELA (mean)	Model 2 ELA (coef. var)
Warmth								
[1]Emotional support	0.10	−0.24	0.15*	−0.20	0.03	−0.03	0.01	−0.02
[2]Environment of respect & rapport	0.17	−0.13	0.05	−0.14	0.10	−0.05	0.04	−0.17***
[3]Care	−1.00	0.06	0.07	−0.14	0.11	−0.12	0.01	−0.05
Classroom management								
[1]Classroom organization	0.19*	−0.18	0.15	−0.05	0.10	−0.04	0.06	−0.05
[2]Managing student behavior	0.01	−0.12	0.22**	−0.15	−0.08	−0.04	0.24**	−0.10
[2]Managing procedures	0.09	−0.04	−0.15	0.06	0.09	0.00	−0.25**	0.11
[3]Control	0.08	−0.10	0.01	−0.13	0.26***	−0.21***	−0.01	0.01
Rigor								
[2]Culture for learning	0.19	−0.12	0.08	−0.01	0.10	−0.01	−0.01	−0.02
[3]Challenge	−0.06	0.09	0.29**	−0.22	0.24***	−0.25***	0.12	−0.11
Instructional strategies								
[1]Instructional support	0.07	−0.38**	0.11	−0.10	−0.13	0.09	0.05	−0.01
[2]Assessments in instruction	−0.07	−0.28*	0.24	0.07	0.04	−0.05	0.01	0.06
[2]Questioning & discussion	−0.20	−0.10	−0.07	0.04	0.08	0.01	−0.05	−0.09
[2]Engage students in learning	0.38	−0.11	−0.05	−0.09	0.04	0.00	−0.13	0.03
[2]Communicate with students	0.04	0.05	0.13	−0.11	0.09	−0.01	0.16	−0.06
[3]Clarify	0.02	−0.17	−0.11	0.07	0.61***	−0.02	0.05	0.04
[3]Confer	−0.09	0.09	0.22	−0.28	0.07	−0.07	−0.03	0.01
[3]Consolidate	−0.01	0.29	0.09	−0.02	−0.11	−0.10	0.04	−0.16
[3]Captivate	0.06	−0.07	0.05	−0.08	−0.33*	−0.07	−0.05	0.08

Superscripts indicate the measure associated with each subscale = [1]CLASS, [2]FFT, [3]Tripod. Coef. Var. = Coefficient of Variation. Standardized beta weights are reported in the table. Separate models were run for each of the four overarching constructs (i.e., warmth, classroom management, rigor, and other instructional supports).

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

3.1. Convergence across observation systems and rating scale

Correlation analyses are reported in Tables 3-6. Across content areas in elementary grades, the FFT and CLASS measures exhibited low to moderate convergent validity as measured through bivariate correlations. The Tripod exhibited divergence from the two observation scales, with many correlations being small in magnitude or negative, particularly for mathematics. For elementary mathematics (Table 3), the CLASS and FFT correlations ranged from 0.09 to 0.61. The correlations between the Tripod and the two observation scales ranged widely from −0.32 to 0.23. For elementary ELA (Table 4), the CLASS and FFT correlations ranged from 0.19 to 0.55. The correlations between the Tripod and the two observation scales ranged widely from −0.20 to 0.29.

Across content areas in middle school grades, the CLASS and FFT exhibited moderate to high-moderate convergent validity. The Tripod demonstrated greater convergent validity with the two observation scales, as compared to elementary grades. For middle school mathematics (Table 5), CLASS and FFT were correlated at a range of 0.38 to 0.69. Correlations between the Tripod and the two observation scales ranged from 0.15 to 0.47. Similarly, for middle school ELA (Table 6), correlations between CLASS and FFT ranged from 0.43 to 0.68. Correlations between the Tripod and the two observation scales ranged from 0.11 to 0.41.

3.2. Prediction of VAMs

Regression findings are reported in Table 7. For each measure (CLASS, FFT, Tripod), Model 1 shows the mean scores of each measure predicting VAM estimates, and Model 2 displays the coefficient of variation predicting VAM estimates. In both types of models, each theoretical grouping was estimated as a separate regression analysis. Across analyses, mean scores tended to be more positively predictive of VAM scores, whereas greater variability in scores tended to demonstrate a negative relation to VAM scores. With regard to specific subscales and dimensions that predicted VAMs, findings were varied across measure type (observation vs. student ratings) and construct (warmth, classroom management, rigor, instructional strategies).

In regression models predicting elementary mathematics VAM scores, mean CLASS Classroom Organization ($B = 0.19, p < .05$), a measure of teachers' *classroom management*, was positively predictive of VAM scores. In contrast, score variability in FFT Assessments in Instruction ($B = -0.28, p < .05$) and CLASS Instructional Support ($B = -0.38, p < .01$), both of which assess *instructional strategies,* were negatively predictive of mathematics VAM scores.

In models predicting elementary ELA VAM scores, mean scores on *warmth, rigor,* and *classroom management* exhibited positive predictive power. Specifically, mean scores on CLASS Emotional Support ($B = 0.15, p < .05$), FFT Managing Student Behavior ($B = 0.22, p < .05$), and Tripod Challenge ($B = 0.29, p < .05$) were predictive of VAM scores. No constructs were predictive of ELA VAM when examining variation in scores.

In models predicting middle school mathematics scores, mean scores on the Tripod assessing *rigor, classroom management*, and *instructional strategies* were positively predictive of outcomes. Specifically, mean scores on Tripod Control ($B = 0.26, p < .001$), Tripod Challenge ($B = 0.24, p < .001$), and Tripod Clarify ($B = 0.61, p < .001$) were positively predictive of VAM scores. In contrast, a mean score on Tripod Captivate ($B = -0.33, p < .05$), a measure of *instructional strategies* was negatively predictive of VAM scores. Score variability in Tripod Control ($B = 0.21, p < .001$) and Challenge ($B = 0.25, p < .001$) were negatively predictive of mathematics VAM scores.

For models predicting middle school ELA, only FFT measures of *classroom management* and *warmth* were significantly associated to VAM scores. A mean score on FFT Managing Student Behavior ($B = 0.24, p < .01$) was positively predictive of VAM scores, but a mean score on FFT Managing Procedures ($B = -0.25, p < .01$) was negatively predictive of the outcome. Score variability in FFT Creating an Environment of Respect and Rapport ($B = -0.17, p < .001$) was negatively predictive of ELA VAM scores.

4. Discussion

As school districts increasingly adopt evaluation approaches that use multiple content-neutral measures for assessing teaching effectiveness and informing professional development (e.g., direct training, consultation, coaching), it is crucial to understand the utility of widely used measures of teacher effectiveness across grade levels and content areas. This study examined the convergent validity of three widely used measures of teaching effectiveness, the CLASS, the FFT, and the Tripod, across two developmental levels (i.e., elementary and middle school grades) and content areas (i.e., mathematics and ELA). We also assessed whether mean scores versus variability in scores on assessments of teaching quality predicted gains in student mathematics and ELA test scores via VAM estimates. Overall, findings offer some promising evidence for convergence among measures of teacher effectiveness and score inferences of direct classroom observation and student ratings to student achievement.

In general, convergent validity between study measures was stronger in middle school grades than in elementary grades, a finding that was consistent across both mathematics and ELA. At the elementary level, some evidence of convergent validity was found for the FFT and CLASS observation measures, as evidenced by positive correlations that were low to moderate in strength; however, the majority of correlations between scores on these measures and the Tripod were small or negative. Conversely, at the middle school level correlations between scores on each of these measures were generally stronger in magnitude. These findings appear to support the argument of many researchers in that while seemingly important, student ratings of teacher effectiveness should not be used as the primary source of information when evaluating teacher effectiveness, especially in elementary grades. These concerns are rooted in potential bias and lack of knowledge about what it means to be an effective teacher (Peterson, Wahlquist, & Bone, 2000; Worrell & Kuterbach, 2001). However, student ratings may offer helpful information for teacher self-reflection and possible professional improvement decisions.

Some interesting patterns emerged within regression models examining predictive associations between CLASS, FFT, Tripod and VAM scores. The most consistent trend was evident in the comparison of mean scores versus variability in scores on these measures. In almost all cases, higher score variability was negatively predictive of VAM estimates, whereas, with few exceptions, mean scores were positively predictive of VAMs. Although not directly measured in this study, there are several possible reasons for this discrepancy. Higher score variability may reflect inconsistency in practices on the part of teachers. Whereas a teacher's instructional quality is often viewed by school administrators as a stable construct (Cohen & Goldhaber, 2016), recent studies highlight that instructional quality can actually vary across time, classroom composition, or lesson format and content (e.g., Cohen & Goldhaber, 2014; Curby et al., 2011). Variability in the student ratings may also reflect differential treatment of students in the same classroom by the teacher, whether intentional or unintentional (McKown & Weinstein, 2003). For example, teachers may form warm and supportive relationships with some students, but have strained relationships with others, or provide more challenging tasks for some based on preconceived expectations for success.

Some predictive patterns associated with specific measures also emerged across developmental levels and content areas. The CLASS measure only exhibited predictive power in the elementary sample, but not in the middle school sample. This is not altogether surprising given that the CLASS framework emerged from research on early childhood (Hamre et al., 2007; Pianta, La Paro, & Hamre, 2008). In current empirical literature, the research base pertaining to CLASS pre-kindergarten and elementary grades is generally stronger than in middle school grades (Pianta, Hamre, & Allen, 2012; Virtanen et al., 2018). Given the limited research on CLASS in the middle school grades, it is possible that CLASS domains better reflect high-quality instruction during the earlier years of schooling and thus warrant further investigation in later grades.

Conversely, the Tripod demonstrated greater predictive power in the middle school sample. Research on the reliability and validity of student perception surveys in K-12 settings is still emerging, but some findings suggest that student ratings are more stable in secondary grades than in elementary grades (Polikoff, 2014). Consequently, the Tripod may have more predictive utility at the middle school level. Some dimensions of the FFT were modestly predictive of VAMs across developmental levels, but the FFT was

most consistently predictive of middle school ELA. The FFT was developed to align with a constructivist approach to teaching, which emphasizes the role of students in developing their own understanding of concepts by interpreting new experiences through the lens of their existing worldviews. The role of teachers, according to a constructivist viewpoint, is to structure learning activities in such a way that students are able to create their own knowledge (Windschitl, 2002). As such, one possibility for stronger predictive power of FFT in middle school grades is that there may be more opportunities to observe constructivist teaching practices in ELA classrooms with older students. Polikoff and Porter (2014) also identified a relation between FFT and ELA outcomes. The authors used MET data to study a subsample of 4th through 8th grade generalist and specialist teachers across all participating districts and found that the interaction between FFT scores and degree of alignment of teachers' instruction with state test content was positively predictive of VAM scores across grade levels in ELA, but not in mathematics. Additionally, within the larger MET Project, both FFT and CLASS were found to demonstrate stronger correlations with ELA-specific observation measures, as compared to mathematics-specific observation tools (Kane & Staiger, 2012), indicating that there may be more alignment between content-neutral observation measures and measures of ELA instruction as compared to mathematics.

With regard to the theoretical constructs representing teaching effectiveness, in elementary grades, mean scores on measures of *classroom management* positively predicted VAM scores across content areas. In other words, teachers' observed skill at managing student behavior and leading a productive classroom related to their ability to cultivate achievement growth in both ELA and mathematics. The importance of a well-managed classroom for student achievement has been identified in other studies of elementary grades (Freiberg, Huzinec, & Templeton, 2009; Pianta, Hamre, & Allen, 2012). *Classroom management* was the only significant construct for elementary mathematics, but there was greater diversity in constructs predicting elementary ELA, with *warmth* and *rigor* also relating to VAM scores. These findings are somewhat aligned with prior literature. The link between classroom management and mathematics also emerged from the 2011 Trends in International Mathematics and Science Study (TIMSS), which identified a strong relation between teacher report of classroom behavioral issues and lower average mathematics achievement (Mullis, Martin, Foy, & Arora, 2012). Additionally, prior examinations of early childhood and elementary samples have found links between rigor and warmth and ELA outcomes (Reyes et al., 2012; Sandilos, Rimm-Kaufman, et al., 2017).

Interestingly, variability in elementary teachers' *instructional strategies* had a negative relation to VAM scores in mathematics, but not ELA. This finding indicates that greater levels of inconsistency in teachers' use of instructional strategies was associated with less growth in mathematics skill in their classrooms. Indeed, consistency and coherence of instruction have long been considered essential components of content area teaching in mathematics (Ball, Lubienski, & Mewborn, 2001).

When examining theoretical constructs at the middle school level, several mean scores from the Tripod tapping into *classroom management* (Control), *rigor* (Challenge), and *instructional strategies* (Clarify) were predictive of higher VAM scores in mathematics, with a particularly strong effect for the Clarify subscale. This latter dimension represents a teacher's ability to help students grasp difficult instructional concepts. This instructional strategy may be especially important for mastery of mathematics in the middle school grades, when students are expected to learn increasingly abstract and difficult mathematics content (Schielack & Seeley, 2010). The unique predictive power of Control and Challenge was consistent with findings from other MET studies (Ferguson & Danielson, 2013; Sandilos, Rimm-Kaufman, et al., 2017). One unexpected finding was that higher mean scores on the Tripod Captivate dimension were associated with lower VAM scores in mathematics. Captivate represents teacher behaviors that are perceived by students as stimulating rather than boring, including efforts to make instruction interesting and relevant to students. It is possible that instruction perceived as more captivating by students may not necessarily be directly aligned with components of mathematics that are assessed on standardized tests.

For middle school grades, mean scores on FFT Managing Student Behavior were positively predictive of VAM scores, whereas FFT Managing Procedures mean scores were negatively predictive, both of which measure aspects of *classroom management.* The transition from elementary to middle school grades is associated with a number of contextual changes, many of which conflict with the developmental needs of students (Eccles et al., 1993). One such change, which may help to explain these mixed findings, is an increased emphasis on teacher control and discipline, during a developmental period in which students have an increasing desire for autonomy.

4.1. Limitations and future directions

The results of this study should be interpreted in light of limitations. First, several factors may limit the generalizability of the findings. Participants in the MET Project sample were not representative of the national population of teachers and students within U.S. schools (White & Rowan, 2013). The present sample represented only a subset of the specialist teachers that participated in the original MET Project, further limiting the generalizability of these findings. Video recordings were also available for a subsample of generalist elementary teachers, but were not utilized in the present study.

Second, the observation and student survey measures from the MET Project were administered under highly controlled conditions (i.e., randomization of classroom rosters) and by highly trained research assistants; thus, these results may not generalize to teacher evaluation systems as they are typically implemented in state and local school districts. However, the controlled conditions provide some insight into relations among measures when potential sources of error are reduced (Blazer, 2015). Additional research is needed with representative samples in less controlled settings to determine if these results replicate within other teacher evaluation systems as they are currently implemented in practice.

Despite the controlled study conditions, all measures in the MET Project demonstrated some degree of measurement error (Ho & Kane, 2013). Consequently, in addition to potential inconsistency in teachers' use of quality practices, variability in scores over time may represent error in the measurement tools. Further research exploring sources of variability in scores (e.g., observer vs. teacher)

would shed light on this issue. Relatedly, there were differences inherent to the measures (i.e., multiple students vs. multiple observations) and study procedures (four FFT scores vs. eight CLASS scores) for the CLASS, FFT, and Tripod that resulted in varying numbers of data points included in the mean scores and coefficient of variation scores computed for each measure. As such, the consistency (or lack thereof) in teachers' scores also may have been influenced by the number of data points available for each measure. Additionally, the amount of time students spent with their assigned teachers remains unknown in this investigation, constituting a possible limitation which warrants further investigation.

An additional limitation reflects the modest psychometric properties of the Tripod. Although the measure is theoretically established (Ferguson, 2010), several studies of MET data have yielded differing factor structures for the Tripod in the secondary grades (e.g., Kuhfeld, 2017; Wallace et al., 2016) bringing into question the validity of examining seven distinct subscales across developmental levels. However, given that the Tripod student surveys have been administered in their current form to more than a million students over the past 15 years (Ferguson, 2012), the seven subscales retain practical relevance and warrant further psychometric exploration.

Finally, there are also a number of limitations associated with using VAM scores as a metric for student achievement. A number of statistical and educational experts have cautioned against the use of VAMs, citing limitations to reliability and validity, and their susceptibility to sources of error and bias (Braun, 2005). Nonetheless, in response to federal and state policy initiatives to improve student achievement, school districts are increasingly adopting standards-based evaluation systems that include VAMs (Steinberg & Donaldson, 2016). Continued research is needed to determine the degree to which VAMs, which primarily measure student growth on standardized achievement tests, relate to the constructs reflected in observation and student rating measures as well as to long-term academic outcomes.

4.2. Implications for practice and research

Using data from the seminal MET Project, this study represents an important contribution in assessing and understanding core aspects of teacher quality (warmth, classroom management, rigor, and instructional strategies) in relation to student achievement. With the passage of ESSA in 2015, which granted significantly more flexibility to states in developing teacher accountability systems, there have been calls to redefine teacher evaluation as a data-driven process that leads to professional growth, rather than a basis on which to determine employment decisions alone (Connally & Tooley, 2016). As states move to reconceive teacher evaluation systems, many are starting to employ multiple measures and evaluation methods including classroom observations, student ratings, and VAMs or other standardized measures of student academic growth (e.g., ETS, Pearson & Measured Progress, 2016). Evaluation systems that integrate multiple measures of instructional process and student educational outcomes have the potential to offer a more balanced and fair approach for determining effectiveness. Moreover, subscale and dimension scores among the three measures in this investigation were not so highly correlated as to indicate redundancy; the measures provide complementary information for educator improvement.

Findings from this study underscore the notion that no singular measure or theoretical construct can fully capture effective teaching across contexts or outcomes (Kane et al., 2013). It is important for schools to consider limitations of existing measures of teaching effectiveness when determining the extent to which they will use such assessments to make high-stakes decisions (Cohen & Goldhaber, 2016). Furthermore, the findings highlight the importance of considering score type (i.e., mean scores versus variability of scores) when inferring classroom performance in relation to student achievement. The value of examining variability among scores on measures of teaching effectiveness is important and warrants further investigation. For example, mean scores on student surveys represent aspects of the classroom teaching environment that are assumed to be shared by all students, whereas variability better reflects the experiences of individual students. Given the negative trend between variability and VAMs, gaining a better understanding of sources of variability in observations and ratings of teachers' instructional quality could help school practitioners target areas for professional growth. For example, some sources of variability that can be included in future measurement modeling may include instructional conditions (opportunities to learn, content rigor, task difficulty), as well as classroom- and/or student-related factors (classroom disruption, student academic engagement).

Teachers' warmth, rigor, classroom management skills, and provision of instructional supports to meet student needs are regarded as essential for a positive learning environment for all students (e.g., Danielson Group, 2013; Hamre et al., 2013; Hattie, 2009; Marzano, 1998; Ruzek et al., 2016). However, findings from this study suggest that the measurement of these core constructs of teacher quality (i.e., warmth, classroom management, rigor, and instructional strategies) should be considered in light of development level (elementary versus middle schools) and content areas taught. In this study, we found that the CLASS, FFT and Tripod offer a synthesis of these core assets of effective teaching and yield some evidence of validity in relation to student mathematics and ELA performance in elementary and middle school. It is our hope that this investigation serves as a springboard for continued studies examining convergence across methods of teacher assessment.

References

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher–student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring Secondary. *School Psychology Review, 42*, 76.

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.

Allington, R. L. (2002). What I've learned about effective reading instruction: From a decade of studying exemplary elementary classroom teachers. *Phi Delta Kappan, 83*, 740–747.

Ball, D. L., Lubienski, S. T., & Mewborn, D. S. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. *Handbook of research on teaching. Vol. 4. Handbook of research on teaching* (pp. 433–456).
Bedeian, A. G., & Mossholder, K. W. (2000). On the use of coefficient of variation as a measure of diversity. *Organizational Research Methods, 3*, 285–297.
Bill and Melinda Gates Foundation (2009-2011). *The measures of effective teaching (MET) project*.
Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review, 48*, 16–29.
Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models.* Princeton, NJ: Educational Testing Service.
Brock, L. L., Nishida, T. K., Chiong, C., Grimm, K. J., & Rimm-Kaufman, S. E. (2008). Children's perceptions of the classroom environment and social and academic performance: A longitudinal analysis of the contribution of the Responsive Classroom approach. *Journal of School Psychology, 46*, 129–149.
Brophy, J. (2006). Observational research on generic aspects of classroom teaching. In P. A. Alexander, & P. H. Winne (Eds.). *Handbook of educational psychology* (pp. 755–780). Mahwah, NJ: Erlbaum.
Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood* (No. w17699)National Bureau of Economic Research.
Cleary, T. J., & Chen, P. P. (2009). Self-regulation, motivation, and math achievement in middle school: Variations across grade level and math context. *Journal of School Psychology, 47*, 291–314.
Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25*, 119–142.
Cohen, J., & Goldhaber, D. (2014). Observations on evaluating teacher performance. In J. A. Grissom, & P. Youngs (Eds.). *Improving teacher evaluation measures: Making the most of multiple measures*. New York, NY: Teachers College Press.
Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*, 378–387.
Cohen, J., Ruzek, E., & Sandilos, L. (2018). Does teaching quality cross subjects? Exploring consistency in elementary teacher practice across subjects. *AERA Open, 4*, 1–16.
Connally, K., & Tooley, M. (2016). Beyond ratings: Re-envisioning state teacher evaluation systems as tools for professional growth. New Americahttps://static.newamerica.org/attachments/12744-beyond-ratings 3/NA_BeyondRatingsPaper.deba47a82ff04af2833cebdbeed0c3ab.pdf.
Curby, T. W., Grimm, K. J., & Pianta, R. C. (2010). Stability and change in early childhood classroom interactions during the first two hours of a day. *Early Childhood Research Quarterly, 25*, 373–384.
Curby, T. W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Mooney, L., Downer, J., ... Pianta, R. C. (2011). Within-day variability in the quality of classroom interactions during third and fifth grade: Implications for children's experiences and conducting classroom observations. *The Elementary School Journal, 112*, 16–37.
Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.
Danielson Group (2013). Framework for teaching evaluation instrument. Retrieved from http://www.danielsongroup.org/theframeteach.htm.
Darling-Hammond, L. (2000). Teacher quality and student achievement. *Education Policy Analysis Archives, 8*, 1.
Downer, J. T., Sabol, T., & Hamre, B. K. (2010). Teacher-child interactions in the classroom: Toward a theory of within- and cross-domain links to children's developmental outcomes. *Early Education and Development, 21*, 699–723.
Downer, J. T., Stuhlman, M., Schweig, J., Martínez, J. F., & Ruzek, E. (2014). Measuring effective teacher-student interactions from a student perspective: A multi-level analysis. *Journal of Early Adolescence, 35*, 722–758.
Dudek, C. M., Reddy, L. A., & Lekwa, A. (2018). Measuring teacher practices to inform student achievement in high poverty schools: A predictive validity study. *Contemporary School Psychology,* 1–14.
Eccles, J. S., Lord, S. E., Roeser, R. W., Barber, B. L., et al. (1997). The association of school transitions in early adolescence with developmental trajectories through high school. In J. Schulenberg, J. L. Maggs, & K. Hurrelmann (Eds.). *Health risks and developmental transitions during adolescence* (pp. 283–320). New York, NY, US: Cambridge University Press.
Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., & Mac Iver, D. (1993). Development during adolescence: The impact of stage-environment fit on young adolescents' experiences in schools and in families. *American Psychologist, 48*, 90.
ESSA (2015). *Every Student Succeeds Act of 2015.* (Pub. L. No. 114–95 § 114 Stat. 1177 (2015–2016)).
ETS, Pearson, & Measured Progress (2016). *PARCC: Final technical report for 2015 administration.* Princeton, NJ: Author.
Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160.
Ferguson, R., & Danielson, C. (2013). How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.). *Designing teacher evaluation systems*. San Francisco, CA: Jossey-Bass.
Ferguson, R. F. (2008). *The tripod project framework.* Cambridge, MA: Harvard University.
Ferguson, R. F. (2010, October). *Student perceptions of teaching effectiveness. (discussion brief).* National Center for Teacher Effectiveness and the Achievement Gap Initiative, Harvard University.
Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan, 94*, 24–28.
Freiberg, H. J., Huzinec, C. A., & Templeton, S. M. (2009). Classroom management—A pathway to student achievement: A study of fourteen inner-city elementary schools. *The Elementary School Journal, 110*, 63–80.
Gentry, M., & Springer, P. M. (2002). Secondary student perceptions of their class activities regarding meaningfulness, challenge, choice, and appeal: An initial validation study. *Journal of Secondary Gifted Education, 13*, 192–204.
Goddard, R. D., Sweetland, S. R., & Hoy, W. (2000). Academic emphasis of urban elementary schools and student achievement in reading and mathematics: A multilevel analysis. *Educational Administration Quarterly, 36*, 682–702.
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. *The Journal of Early Adolescence, 35*, 651–680.
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., Mashburn, A. J., & Downer, J. T. (2007). Building a science of classrooms: Application of the CLASS framework in over 4,000 early childhood and elementary classrooms. New York: Foundation for Child Development. Retrieved from http://fcd-us.org/resources/building-science-classrooms-application-class-framework-over-4000-us-early-childhood-and-e?destination=resources%2Fsearch%3Ftopic%3D0%26authors%3DHamre%26keywords%3D.
Hattie, J. (2009). *Visible learning: A synthesis of meta-analyses relating to achievement.* New York, NY: Routledge.
Hennessy, S., Rojas-Drummond, S., Higham, R., Márquez, A. M., Maine, F., Ríos, R. M., & Barrera, M. J. (2016). Developing a coding scheme for analyzing classroom dialogue across educational contexts. *Learning, Culture and Social Interaction, 9*, 16–44.
Heyburn, S., Lewis, J., & Ritter, G. (2010). *Compensation reform and design preferences of Teacher Incentive Fund grantees.* RAND Corporation.
Hill, H. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management, 28*, 700–709.
Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*, 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. MET project.* Bill & Melinda Gates Foundation.
Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy, 126*, 2072–2107.
Kamil, M. L., Borman, G. D., Dole, J., Kral, C. C., Salinger, T., & Torgesen, J. (2008). Improving adolescent literacy: Effective classroom and intervention practices. *IES*

Practice Guide. NCEE 2008–4027. National Center for Education Evaluation and Regional Assistance.
Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. Retrieved from Bill & Melinda Gates Foundationhttps://www.rand.org/pubs/external_publications/EP50156.html.
Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teachers: Combining high-quality observations with student surveys and achievement gains. Retrieved from Policy and practice brief prepared for the Bill and Melinda Gates Foundationhttp://metproject.org/downloads/MET_Gathering_Feedback_Practioner_Brief.pdf.
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.
Kazemi, E., & Stipek, D. (2009). Promoting conceptual thinking in four upper-elementary mathematics classrooms. *Journal of Education, 189*, 123–137.
Kuhfeld, M. (2017). When students grade their teachers: A validity analysis of the tripod student survey. *Educational Assessment, 22*, 253–274.
Lee, J. (2012). The effects of the teacher-student relationship and academic press on student engagement and academic performance. *International Journal of Educational Research, 53*, 330–340.
Lee, V., & Smith, J. B. (1999). Social support and achievement for young adolescents in Chicago: The role of school academic press. *American Education Research Journal, 36*, 907–945.
Mantzicopoulos, P., French, B. F., Patrick, H., Watson, J. S., & Ahn, I. (2018). The stability of kindergarten teachers' effectiveness: A generalizability study comparing the framework for teaching and the classroom assessment scoring system. *Educational Assessment, 23*, 24–46.
Marzano, R. J. (1998). *A theory-based meta-analysis of research on instruction.* (Retrieved from ERIC database. (ED427087).
McKown, C., & Weinstein, R. (2003). The development and consequences of stereotype consciousness in middle childhood. *Child Development, 74*, 498–515.
Midgley, C., & Edelin, K. C. (1998). Middle school reform and early adolescent well-being: The good news and the bad. *Educational Psychologist, 33* (195–106).
Mullis, I. V., Martin, M. O., Foy, P., & Arora, A. (2012). TIMSS 2011 international results in mathematics. Retrieved from International Association for the Evaluation of Educational Achievementhttps://eric.ed.gov/?id=ED544554.
National Research Council (2004). Engaging schools: Fostering high school students' motivation to learn. Retrieved from http://www.nap.edu/catalog.php?record_id=10421.
Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives, 18*, 1–27.
Paige, M. (2012). Using VAM in high-stakes employment decisions. *Phi Delta Kappan, 94*, 29–32.
Patrick, H., & Mantzicopoulos, P. (2016). Is effective teaching stable? *The Journal of Experimental Education, 84*, 23–47.
Peterson, K. D., Wahlquist, C., & Bone, K. (2000). Student surveys for school teacher evaluation. *Journal of Personnel Evaluation in Education, 14*, 135–153.
Pianta, R. C. (1999). *Enhancing relationships between children and teachers.* Washington, DC: American Psychological Association.
Pianta, R. C., & 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., & Allen, J. P. (2012). Teacher-student relationships and engagement: Conceptualizing, measuring, and improving the capacity of classroom interactions. In S. Christenson, A. Reschly, & C. Wylie (Eds.). *Handbook of research on student engagement.* Boston, MA: Springer.
Pianta, R. C., Hamre, B. K., & Mintz, S. (2012). Upper elementary and secondary CLASS technical manual. Retrieved from cdn2.hubspot.net/hubfs/336169.
Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system manual, K-3.* Baltimore: Brookes Publishing Co.
Polikoff, M. S. (2014). The stability of observational and student survey measures of teaching effectiveness. *The American Journal of Education, 121*, 183–212.
Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis, 36*, 399–416.
Ponitz, C. C., Rimm-Kaufman, S. E., Grimm, K. J., & Curby, T. W. (2009). Kindergarten classroom quality, behavioral engagement, and reading achievement. *School Psychology Review, 38*, 102.
Praetorius, A. K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31*, 2–12.
Reddy, L. A., Glover, T., Kurz, A., & Elliott, S. N. (2019). Assessing the effectiveness and interactions of instructional coaches: Initial psychometric evidence for the instructional coaching assessments–teacher forms. *Assessment for Effective Intervention, 44*, 104–119.
Reddy, L. A., Kettler, R. J., & Kurz, A. (2015). School-wide educator evaluation for improving school capacity and student achievement in high-poverty schools: Year 1 of the school system improvement project. *Journal of Educational and Psychological Consultation, 25*, 90–108.
Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology, 104*(3), 700–712. https://doi.org/10.1037/a0027268.
Rimm-Kaufman, S. E., & Chiu, Y. J. (2007). Promoting social and academic competence in the classroom. *Psychology in the Schools, 44*, 397–413.
Rimm-Kaufman, S. E., & Pianta, R. C. (2000). An ecological perspective on the transition to kindergarten: A theoretical framework to guide empirical research. *Journal of Applied Developmental Psychology, 21*, 491–511.
Rimm-Kaufman, S. E., Storm, M. D., Sawyer, B. E., Pianta, R. C., & LaParo, K. M. (2006). The Teacher Belief Q-Sort: A measure of teachers' priorities in relation to disciplinary practices, teaching practices, and beliefs about children. *Journal of School Psychology, 44*, 141–165.
Rubie-Davies, C. M. (2007). Classroom interactions: Exploring the practices or high- and low-expectation teachers. *British Journal of Educational Psychology, 77*, 289–306.
Ruzek, E. A., Hafen, C. A., Allen, J. P., Gregory, A., Mikami, A. Y., & Pianta, R. C. (2016). How teacher emotional support motivates students: The mediating roles of perceived peer relatedness, autonomy support, and competence. *Learning and Instruction, 42*, 95–103.
Sandilos, L. E., Rimm-Kaufman, S. E., & Cohen, J. J. (2017). Warmth and demand: The relation between students' perceptions of the classroom environment and achievement growth. *Child Development, 88*, 1321–1337.
Sandilos, L. E., Wollersheim Shervey, S., DiPerna, J. C., Lei, P., & Cheng, W. (2017). Structural validity of CLASS K–3 in primary grades: Testing alternative models. *School Psychology Quarterly, 32*, 226.
Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations. *Principal-teacher conferences, and district implementation. Research report* (pp. 60637). 1313 East 60th Street, Chicago, IL: Consortium on Chicago School Research.
Schielack, J., & Seeley, C. L. (2010). Transitions from elementary to middle school math. *Teaching Children Mathematics, 16*, 358–362.
Schunk, D. H., & Miller, S. D. (2002). Self-efficacy and adolescents' motivation. *Academic Motivation of Adolescents, 2*, 29–52.
Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy, 11*, 340–359.
Stronge, J. H. (2008). *Qualities of effective teachers* (3rd ed.). Alexandria, VA: ASCD.
Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education, 62*, 339–355.
Tyler, J. H. (2011). Designing high quality evaluation systems for high school teachers: Challenges and potential solutions. Retrieved from Center for American Progresshttps://www.americanprogress.org/issues/education-k-12/reports/2011/11/29/10614/designing-high-quality-evaluation-systems-for-high-school-teachers/.
U.S. Department of Education (2009, November). Race to the top executive summary. Retrieved from https://www2.ed.gov/programs/racetothetop/executive-summary.pdf.
Virtanen, T. E., Pakarinen, E., Lerkkanen, M. K., Poikkeus, A. M., Siekkinen, M., & Nurmi, J. E. (2018). A validation study of classroom assessment scoring system–secondary in the Finnish school context. *The Journal of Early Adolescence, 38*(6), 849–880.
Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the tripod student perception survey. *American Educational Research Journal, 53*, 1834–1868.
White, M., & Rowan, B. (2013). *User guide to measures of effective teaching longitudinal database.* Ann Arbor, MI: Inter-University Consortium for Political and Social Research, The University of Michigan.

Windschitl, M. (2002). Framing constructivism in practice as the negotiation of dilemmas: An analysis of the conceptual, pedagogical, cultural, and political challenges facing teachers. *Review of Educational Research, 72*, 131–175.

Woolfolk Hoy, A., & Weinstein, C. S. (2011). Student and teacher perspectives on classroom management. In C. M. Evertson, & C. S. Weinstein (Eds.). *Handbook of classroom management: Research, practice, contemporary issues* (pp. 181–223). Malwah, NJ: Lawrence Erlbaum Associates.

Worrell, F. C., & Kuterbach, L. D. (2001). The use of student ratings of teacher behaviors with academically talented high school students. *Journal of Secondary Gifted Education, 12*, 236–247.

Zaslow, M., Anderson, R., Redd, Z., Wessel, J., Daneri, P., Green, K., ... Martinez-Beck, I. (2016). I. Quality thresholds, features, and dosage in early care and education: Introduction and literature review. *Monographs of the Society for Research in Child Development, 81*(2), 7–26.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*, 64–70.