American Educational Research Journal
October 2020, Vol. 57, No. 5, pp. 2021–2058
DOI: 10.3102/0002831219891409
Article reuse guidelines: sagepub.com/journals-permissions
© 2019 AERA. http://aerj.aera.net

The Predictive Validity of Classroom Observations: Do Teachers' Framework for Teaching Scores Predict Kindergarteners' Achievement and Motivation?

Helen Patrick
Panayota Mantzicopoulos
Purdue University
Brian F. French
Washington State University

We used multilevel analysis to examine the predictive validity of scores from the Framework for Teaching (FFT), the observation measure used most often to evaluate teachers' instruction. We investigated how well 81 kindergarten teachers' FFT scores for eight reading and eight mathematics lessons observed throughout the year predicted students' year-end achievement and motivation in reading and mathematics, controlling for students' sex, ethnicity, and achievement entering kindergarten. Standardized reading and mathematics achievement were each predicted by FFT scores; however, they accounted for very little of the overall variance in students' achievement: 2.5% for reading and 1.3% for mathematics. Neither students' end-of-year criterion-referenced achievement nor motivation were predicted by FFT scores.

Helen Patrick is a professor of educational psychology in the Department of Educational Studies at Purdue University, 100 North University Street, West Lafayette, IN 47907-2098; e-mail: *hpatrick@purdue.edu*. Her research includes examining associations of classroom contexts and teacher practices with student motivation, engagement, and achievement.

Panayota Mantzicopoulos is a professor of educational psychology in the Department of Educational Studies at Purdue University. Her research focuses on the links between teacher practices in the early years of school and young children's learning and engagement.

Brian F. French is a professor of educational psychology at Washington State University. He investigates the properties of test scores used to make decisions about individuals and groups. His research includes both methodological and applied studies in the area of educational and psychological measurement.

Keywords: classroom research, elementary schools, hierarchical linear modeling, instructional practices, learning environments, mixed methods, observational research, school/teacher effectiveness, teacher assessment

Despite a plethora of initiatives intended to ensure high academic achievement for students across the United States, results have been disappointing. Attributions for low student achievement vary; however, a commonly cited claim is that ineffective teachers are in large part responsible (Weisberg, Sexton, Mulhern, & Keeling, 2009). Over the past decade, politicians and policy advocates argued for the need to shift attention to teachers' instruction. They proposed that student achievement would be increased if teachers were observed and evaluated in their use of instructional practices associated with achievement, and those with low scores were dismissed (The New Teacher Project, 2010; Weisberg et al., 2009). These arguments influenced subsequent policies in the United States.

Observation-based measures of instruction (OMI) are now ubiquitous in teacher evaluation systems throughout the United States (Cohen & Goldhaber, 2016; Steinberg & Kraft, 2017). Although it is crucial that their use is grounded in robust empirical evidence, there is little published, independently reviewed research supporting the validity of using OMI scores to evaluate teachers. This includes research that examines the rationale for their use—that scores predict students' achievement growth (Weisberg et al., 2009). Therefore, research that addresses associations between teacher OMI scores and student achievement is critically needed. This is especially the case in the early elementary grades because, in the absence of student achievement tests, teachers in kindergarten through second grade are evaluated primarily on their observed instruction (Whitehurst, Chingos, & Lindquist, 2014).

We sought to address this need by examining the predictive validity of scores from the Framework for Teaching (FFT; Danielson, 2013), the most commonly used observation measure in U.S. schools (Center on Great Teachers and Leaders, 2013). We focused on kindergarten teachers not only because the use of OMI in the early grades is understudied but also because there is evidence that teachers' effectiveness, as measured by FFT scores, varies across grade levels (Mihaly & McCaffrey, 2014). We used data from 81 teachers and 1,296 lessons to investigate how well their FFT scores predicted student achievement and motivation in two key content areas—reading and mathematics.

Teacher Accountability for Student Achievement

Teacher evaluation in the United States underwent a major overhaul a decade ago, when the initial widespread public support for the No Child Left Behind (NCLB) legislation faltered. NCLB was intended to raise achievement by encouraging schools to hire highly qualified teachers and by administering substantial rewards and sanctions to schools based on student test scores (U.S. Department of Education [USDOE], 2007; Weisberg et al., 2009). However, near the time when all students were expected to meet grade-level standards, it was clear that the NCLB policies had not resulted in universally high achievement.

Policymakers and political commentators responded to NCLB's failure with another recommendation for increasing student achievement. Their argument involved leveraging student achievement by (a) focusing on teachers' instruction directly, rather than indirectly via teacher qualifications and student test scores; (b) shifting rewards and penalties from schools to individual teachers; and (c) changing how instruction was evaluated (Toch & Rothman, 2008; Weisberg et al., 2009). Proponents argued that there was "a culture of indifference about the quality of instruction in each classroom" (Weisberg et al., 2009, p. 2) and school administrators could not be counted on to identify and respond to differential teacher quality. Therefore, teachers needed to be held personally accountable for their practices and their students' achievement. However, purportedly, schools' inability to assess instruction accurately stymied teacher accountability. Existing evaluation systems were lambasted as superficial, lax, capricious, and unremittingly rosy, yielding unreliable results that failed to differentiate among teachers. Therefore, teacher evaluation needed to be reconstructed (Kane & Staiger, 2012; Weisberg et al., 2009).

Political commentators argued that teachers should be evaluated with a system that accurately differentiates among them in terms of their students' achievement growth. A crucial part of this assessment system involved observing teachers' instruction for evidence they were using effective practices (Chait, 2010; Toch & Rothman, 2008; Weisberg et al., 2009).

Federal legislation after NCLB did overhaul teacher evaluation (USDOE, 2009, 2011). Policies stipulated that teachers be evaluated with multiple measures, including student test scores. Furthermore, high-stakes outcomes (e.g., salary increases, contract terminations) were attached to individual teacher's evaluations. States generally chose to evaluate teachers with a combination of student standardized test scores, teacher value added, and observed instruction; some more recently also use student surveys (Steinberg & Kraft, 2017). In practice, though, test scores, and hence value added, are not feasible for the 70% of teachers who teach grade levels or subjects that are not part of states' standardized testing programs (Steinberg & Kraft, 2017). Furthermore, students in the early grades are unlikely to be asked to rate their teacher. Thus, OMI became the most widespread means of evaluating teachers (Garrett & Steinberg, 2015).

The emphases on student achievement and teacher evaluation using multiple measures continues with the Every Student Succeeds Act (ESSA; USDOE, 2016). Greater recognition of the importance of student outcomes beyond achievement, such as motivation and engagement, led to ESSA also requiring that districts evaluate at least one nonacademic outcome. A significant change to previous federal legislation is that ESSA gives states independence in how they evaluate teachers, including whether evaluations are based on student achievement (USDOE, 2016). A predominant post-ESSA change at the state level has been to propose or institute legislation that prohibits or postpones the use of value-added measures (Close, Amrein-Beardsley, & Collins, 2018; Croft, Guffy, & Vitale, 2018). Similar actions have not addressed OMI, suggesting that their use may continue.

Evaluating Instruction With Observational Measures

There is considerable support among teachers and school administrators for using OMI for accountability purposes (Jiang, Sporte, & Luppescu, 2015), in addition to the previously mentioned support of the lay public, including politicians. This practice is compatible with the long-standing tradition of school administrators observing teachers. Moreover, protocols generally have high face validity (Jiang et al., 2015; Kimball, 2002), consistent with the view that these measures "must be based on aspects of teaching that excellent teachers recognize as characteristic of their practice" (Bill & Melinda Gates Foundation, 2010, p. 4).

In contrast to face validity, there is little compelling evidence that OMI have the predictive validity fundamental to evaluating individual teachers. Rather, there is considerable evidence that they are not associated consistently with student achievement (Brophy, 2006; Lavigne & Good, 2014). Issues of low validity can be understood from evidence of observation scores' low reliability (i.e., stability). A wealth of research spanning decades indicates that OMI do not capture stable instructional patterns (e.g., Brophy, 1973; Brophy, Coulter, Crawford, Evertson, & King, 1975; Emmer, Evertson, & Brophy, 1979; Good, 1979; Good & Grouws, 1977; Meyer, Linn, & Hastings, 1991; Rosenshine, 1970; for reviews, see Good & Lavigne, 2015; Lavigne & Good, 2014). For example, Brophy (1973) found that over three consecutive years only 14% of teachers were consistently rated highly effective and 14% were consistently rated ineffective. Widely accepted conclusions from this body of research included that "optimal teacher behavior . . . varies with the nature of the students and the goals of the instructional activities" (Brophy, 2006, p. 765) and that most teachers do not exhibit stable patterns of instruction (Good & Lavigne, 2015).

Political commentators' and policy advocates' arguments for adopting observation systems were not accompanied with caution. There was no mention of the paucity of empirical evidence showing that OMI scores are robust predictors of students' achievement growth. Neither was there acknowledgment that the scores of teachers' observed practices are highly variable (Lavigne & Good, 2014). Rather, arguments were based heavily on rhetoric, particularly regarding the prevalence of ineffective teachers and the difficulty firing them (Bill & Melinda Gates Foundation, 2010; Griffith & McDougald, 2016; Weisberg et al., 2009).

The long history of researchers using observation procedures to identify teacher practices related to achievement (Brophy, 2006) may be construed as validating the decision to harness these procedures for evaluation systems, including assessing the effectiveness of individual teachers. However, proponents have not addressed crucial differences between the use of OMI scores for research and for individual teacher evaluation purposes. Specifically, although researchers using OMI identified broad patterns of instructional practices that differentiated groups of teachers, results were acknowledged as not necessarily applying to any particular teacher (Brophy, 2006). That is, despite yielding statistically significant findings, researchers cautioned expressly that results should not be used to prescribe practices, nor were they intended for identifying individual teachers or portrayed as being sufficiently robust to support either purpose (Brophy, 1988, 2006). However, using researchers' procedures for teacher evaluation assumes that OMI scores (a) are necessarily relevant and accurate for each individual teacher and (b) meet the levels of reliability and validity required for high-stakes decision making, which are considerably higher than needed for research purposes (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

At the time that federal and state legislation prompted use of OMI for high-stakes accountability, empirical support for this practice from independently reviewed research published in scholarly journals was scarce. This situation has not improved appreciably since then. Notably, most studies are disseminated as unpublished reports, often by foundations with political agendas. These reports do not typically provide the level of methodological detail necessary for publication, leaving results open to questions, and thus allowing for statements not supported by data to remain unchecked.

There is a critical need for research that examines whether using OMI for teacher evaluation serves the intended purpose; that is, that scores identify teachers whose instruction leads to increased achievement for their students. Such research is especially important for early elementary teachers because, as we have noted, they are evaluated predominantly with OMI (Dee & Wyckoff, 2015; Whitehurst et al., 2014).

In the present study, we respond to the need for research addressing the predictive validity of using OMI scores for teacher evaluation in the early grades. Specifically, we investigate the extent to which kindergarten teachers' FFT scores predict their students' end-of-year achievement in reading and mathematics.

Framework for Teaching

The FFT is the most prominent and endorsed OMI in the United States, recommended by 26 states and the District of Columbia for evaluating teachers (Center on Great Teachers and Leaders, 2013). It stems from the Praxis

III, a performance assessment developed by ETS for evaluating beginning teachers and used for licensure (Danielson, 2007; Dwyer, 1998). Danielson (2007), who contributed to developing the Praxis III, modified it and renamed it the FFT, which was then promoted as a measure of in-service teachers' effectiveness. It is purportedly appropriate for lessons in any content area and for Grades K–12 (Danielson, 2007, 2013). Its developers have not published evidence of the FFT's validity, however.

The FFT comprises four domains of practice: Preparation and Planning, Classroom Environment, Instruction, and Professionalism (Danielson, 2013). In the present study, we focus on the two observation-based domains—Classroom Environment and Instruction. In the absence of consistent findings about the structure of FFT scores, there is some variability in how researchers create scores. Most often, researchers combine teachers' Classroom Environment and Instruction scores to create a composite observation measure (e.g., Garrett & Steinberg, 2015; Martinez, Schweig, & Goldschmidt, 2016; Polikoff & Porter, 2014), which mirrors the single evaluation score that teachers receive. However, researchers have used other combinations of FFT components (i.e., items) to create scores; these include computing separate scores for Classroom Environment and Instruction (e.g., Muñoz & Dossett, 2016), aggregating scores across all four domains (e.g., Milanowski, 2004), and using other configurations of components (e.g., Kimball, White, Milanowski, & Borman, 2004).

In the present study, we take two approaches to examining the FFT. First, we consider Classroom Environment and Instruction scores separately, in line with Halpin and Kieffer's (2015) arguments that "effective teaching requires the skillful coordination of multiple practices . . . and teachers' practices are not well described in terms of a single construct" (p. 263). Thus, it is possible that the two domains may be associated differentially with student outcomes. If so, this information would be valuable to both researchers and school administrators. We also create a Total FFT score by aggregating the two domains, as is most typical, therefore allowing comparison with other studies.

Associations Between FFT Scores and Student Achievement

As we have noted, there is little published research documenting associations between teacher FFT scores and their students' achievement growth. Of the studies we located, none addressed the early grades (i.e., Grades K–2). One reason for this may be because researchers typically measure achievement growth with value-added metrics, which require state standardized test scores that are not available in kindergarten through second grade. An exception is Kimball et al.'s study (2004); they used scores from the state's criterion-referenced achievement tests, which were aligned to the state academic content standards.

The few published studies investigating whether teacher FFT observation scores predict student achievement growth in upper elementary through high school produced mixed results. A Classroom Environment and Instruction aggregate was not related significantly to either math or reading achievement in one study (Polikoff & Porter, 2014), but in another associations were statistically significant, albeit small (Tyler, Taylor, Kane, & Wooten, 2010). Results of other studies were inconsistent across grade levels, content areas, or both. Specifically, associations with FFT scores were statistically significant for growth in mathematics, but not reading, achievement (Muñoz & Dossett, 2016). They were also significant for reading and math achievement in fifth but not third grade, and for fourth-grade reading, but not mathematics, achievement (Kimball et al., 2004). Finally, partial correlations varied widely, with no discernable pattern, across grade levels and content areas (Milanowski, 2004). In the latter study, however, the FFT score was an aggregate of all four domains rather than just the two observation-based domains, making comparisons with the other studies difficult.

The most comprehensive data about the FFT emanate from the high-profile and large-scale Measures of Effective Teaching (MET) Project, developed "to improve the quality of information about teaching effectiveness" (Bill & Melinda Gates Foundation, 2010, p. 3). Despite the prominence of the MET Project, we do not present results contained in its unpublished reports; findings were not independently peer-reviewed and some results have been revised (e.g., Kane, McCaffrey, Miller, & Staiger, 2013, p. 3). One of the project's central claims—that "all five observation instruments [of which the FFT was one] were positively associated with student achievement gains" (Kane & Staiger, 2012, p. 6)—did not hold when student randomization was considered. That is, when only scores of teachers with students who complied with randomization were examined, FFT scores did not predict student achievement gains (Garrett & Steinberg, 2015). With respect to its observation data, there is concern that MET project OMI raters were held to an "unacceptably low" (White, 2018, p. 497) standard of reliability; it was 0.26 for the FFT, in contrast to the "0.70 rule of thumb" (p. 497) for research. This low requirement for rater reliability arguably undermines confidence in OMI ratings and, consequently, results with MET project observation data.

Early Elementary Grades

The absence of research from the early grades that examines associations between FFT scores and student achievement may be construed as indifference to early education, or as a view that it is less important than education in the middle and high school years. It is disconcerting because it runs contrary to the crucial nature of mastering content taught in the early school grades. First, teaching reading has been noted as "a most fundamental and important issue facing schools . . . particularly in the early grades" (National Research Council [NRC], 1998, p. 172). Second, and equally important, teachers' mathematics practices are critical to supporting young students' mathematical thinking (National Council of Teachers of Mathematics, 2013; NRC, 2001, 2009). Third, for young students, mastering reading and math skills forms the basis for all later learning. Learners with deficits in either reading or mathematics will likely experience poor academic outcomes throughout their school years and beyond (NRC, 1998, 2009). Therefore, because teaching practices in each content area have both immediate and far-reaching consequences for students' academic success, it is important to document the links between FFT scores and student achievement separately by content area.

Content Area Differences

Instruction is linked inextricably to the content being taught; accordingly, there is considerable variation in how different disciplines are taught. Teachers' practices differ depending on instructional resources (e.g., curricular materials), which typically vary by content area (Grossman, Stodolsky, & Knapp, 2004). Policies targeting specific disciplines, such as the intensive time commitment to reading necessary for enacting Reading First, affect the allocation of instructional time to other subjects. Accordingly, even exceptionally high-quality instruction in a content area that is allocated little time in teachers' schedules may not lead to substantial gains in student achievement. Additionally, beliefs about the best ways to teach specific content and skills are components of pedagogical content knowledge, which necessarily differs among content areas (Ball, Thames, & Phelps, 2008; Shulman, 1986). Thus, practices included in the FFT may be more central for some content areas than others.

Given the range of content they are required to teach, it is likely that elementary school teachers are not equally effective across all subject areas. In general, teachers' knowledge varies across the areas they teach, as does their enthusiasm for and confidence in teaching it (Grossman et al., 2004). Teachers in the elementary grades tend to be more passionate about and confident in teaching reading compared with mathematics (Grossman et al., 2004). Elementary teachers also typically receive considerably more professional development in reading instruction compared with other content areas (Rouge, Hansen, Muller, & Chien, 2008). Thus, the typical content area differences in teachers' subject-specific knowledge, pedagogical content knowledge, interest, and confidence may manifest in differential, and differentially effective, instructional practices (e.g., extending or enriching curricula, inviting students' questions, encouraging extended discussion).

The scant research investigating associations between FFT scores and achievement growth has not addressed possible content area differences explicitly. Most data come from middle or high school teachers, who typically do not teach a range of content areas, which precludes examining the same teachers' FFT scores for different disciplines. Of the two published studies we identified that included only elementary school teachers, one aggregated FFT scores for English and mathematics lessons (Martinez et al., 2016). The second, which used FFT scores supplied by the school district, did not refer to the content area observed (Kimball et al., 2004). The paucity of attention to possible content area differences involving FFT scores may stem from the claim that the FFT is "a generic instrument, applying to all disciplines" (Danielson, 2013, p. 6), thus fostering an assumption that teachers' instructional quality is consistent across different content. There is little, if any, empirical support for this claim however, and none from the early grades. Therefore, our analyses of teachers' FFT scores and students' achievement and motivation in reading were conducted separately from analyses with comparable teacher and student mathematics data.

The Importance of Student Motivation as an Indicator of Effective Teaching

Although achievement is the educational outcome receiving the greatest public attention, there is overwhelming evidence that students' motivation is equally important for their current and future success. From the earliest grades, motivation plays an important role in promoting students' learning and achievement. When students are motivated to learn, they engage with their teacher and the content during lessons, seek to extend their knowledge and skills, expend effort, take on challenges, express interest and enthusiasm, are thoughtful about what they are learning, and persist when experiencing difficulty. These behaviors generally lead to learning and achievement, which encourage motivational beliefs (e.g., perceived competence or efficacy, interest, and enjoyment) and support students' continued engagement (Wigfield et al., 2015). The reciprocal patterns of motivation and achievement become increasingly stable during elementary school (Alexander & Entwisle, 1988; Gottfried, Fleming, & Gottfried, 2001).

Although young students are typically enthusiastic and confident in their abilities to learn, as early as kindergarten a discernable number develop maladaptive motivational beliefs (e.g., low perceived competence, disliking learning activities; Patrick, Mantzicopoulos, Samarapungavan, & French, 2008) and behaviors (e.g., giving up easily, exhibiting anxiety or helplessness, avoiding or resisting difficult activities; Dweck, 2002; Hirvonen, Tolvanen, Aunola, & Nurmi, 2012; Patrick et al., 2008). Poor motivation tends to perpetuate during the early school years, and its reciprocal associations with poor achievement also tend to become cumulative (Hirvonen et al., 2012). For children in the early school years, the consequences of low motivation are greater than they are for students in high school (Schwinger, Wirthwein, Lemmer, & Steinmayr, 2014).

Current social cognitive theories of motivation identify teachers as being central to student motivation (e.g., Ames, 1992; Reeve, 2002; Wigfield & Eccles, 2000). Considerable evidence to support this premise comes from a wealth of studies representing numerous motivation theories that, together, address diverse teacher practices, involve both experimental and correlational designs, and span the range of grade levels. Furthermore, researchers have considered practices that are related positively to student motivation, in addition to those associated negatively (for reviews, see Kaplan & Patrick, 2016; Karabenick & Urdan, 2014; Perry, Turner, & Meyer, 2006).

Given the association between teachers' practices and their students' motivation, we believe that evaluations of teacher quality must consider students' motivation in addition to their achievement. Furthermore, we argue that instruction that raises achievement while undermining motivation should not be considered effective, or even satisfactory.

There is little evidence yet of whether OMI designed to measure instructional practices that lead to student achievement will also be sensitive to teacher-level differences that are linked to student motivation. However, many of the practices that promote achievement are also central to supporting motivation (Stipek, 2002). Therefore, OMI that identify teachers who increase student achievement may also identify teachers who foster their students' motivation. In the present study, we considered the extent to which teachers' FFT scores were related to student motivation. We included measures of both positive and negative motivation: students' interest and effort for learning, and their need for teacher support or encouragement to engage in lessons.

The Present Study

We used the FFT to rate eight reading and eight mathematics lessons (four per content area in fall and in spring) taught by 81 kindergarten teachers. We then used multilevel analyses to investigate how well teachers' subject-specific FFT scores predicted students' end-of-year achievement (progress toward meeting state standards and standardized achievement) and motivation (interest and need for support) in both reading and mathematics, controlling for student characteristics (sex, ethnicity, free or reduced-cost lunch [FRL] status, and subject-specific Fall achievement).

Method

Participants

This study is part of a larger, multiyear project that examines a range of teacher observation measures. Data for this study were collected during two consecutive years.

Teachers

There were 81 kindergarten teachers: 79 female and 2 male, 79 White and 2 Hispanic. The teacher sample comprised of 41 teachers in the first year and 40 in the second. Teachers' experience ranged from 1 to 40 years (*M* = 13.1, *SD* = 9.5 years).

Schools

Teachers from 22 public schools throughout Indiana participated. Schools' composition of students' ethnicity and FRL status varied considerably. Across schools, 93.7% to 19.6% of students were White; 52.6% to 0.0% were Black, and 51.4% to 2.9% were Hispanic, whereas between 87.4% and 25.4% of students received FRL. Schools also varied in terms of size (279–976 students), average student achievement (state's report card grade ranged from A to F), and locale (located in rural areas, small towns, and the urban fringe of a large city).

Students

We received informed consent for 1,455 kindergarteners, representing 84% of students in the participating classrooms. However, data from fall to spring on the variables used in the study were available for 1,302 students (643 boys [49.4%], 659 girls). Of the 153 students with incomplete data, 86 moved out of the classroom during the school year, 36 could not be tested during one or both data collection periods due to communication difficulties (e.g., students had limited English knowledge or special needs such as autism), and 31 had missing data either because they were absent, enrolled after the fall testing period, or the teacher forms were not complete.

According to school records, the majority of participating students (65.4%) were White; 13.7% were Hispanic, 13.1% were Black, and 7.8% were Multiracial or Other. This distribution of student ethnicity is similar to that within the state, where 67.6% of students are White, 12.3% Hispanic, and 12.3% Black (Indiana Department of Education [DOE], 2019). As a measure of family socioeconomic status (SES), 590 students (45.3%) received FRL; 47.4% of students in the state receive FRL. We coded students' sex (males = 0, females = 1), and students' FRL status (0 = receiving FRL, 1 = paid lunch) for use in the multilevel models. We also created dummy codes for the three minority ethnicities, whereby White was the comparison group. That is, Black = 1, non-Black = 0; Hispanic and Other were coded similarly.

Lessons

We issued teachers with an iPad and stand, which they used to record one reading and one mathematics lesson per week for 20 weeks (10 weeks in the fall semester and 10 in the spring). We told them that we were interested in seeing their regular lessons and did not want them to do anything special for us. Our only requirement was that the entire lesson was at least 20 minutes long, although different activities may be, and usually were, included within the lesson. In addition to assuring teachers that their lessons would not be viewed by anyone outside our project, we encouraged them to upload lessons even if lessons did not occur as anticipated or teachers were dissatisfied with them, because it would approximate the process of an observer visiting. Teachers uploaded their lessons to a secure project website. Teachers knew we would be applying different observation protocols to their lessons. Although we do not know whether they were familiar with the FFT, the state's recommended teacher evaluation instrument (RISE; Indiana DOE, n.d.-a) is modeled closely on the FFT and has been mapped to its components (Indiana DOE, n.d.-b). School districts could modify the RISE, however, or use another state-approved instrument.

For the present study, we selected eight reading and eight math lessons per teacher randomly (four from each semester) for a total of 1,296 lessons. Reading lessons averaged (denoted as minutes:seconds) 24:41 (*SD* = 5:58) and math lessons averaged 24:44 (*SD* = 5:04).

Classroom Measure and Procedure

The FFT's (Danielson, 2013) observation measure of teacher practices is purported to promote improved student learning across contexts (e.g., grade level, content area). It has two domains (i.e., Classroom Environment and Instruction) with four components in each. The *Classroom Environment* components are (1) creating an environment of respect and rapport, (2) establishing a culture for learning, (3) managing classroom procedures, and (4) managing student behavior. The *Instruction* components are (1) communicating with students, (2) using questioning and discussion techniques, (3) engaging students in learning, and (4) using assessment in instruction. At the end of the observation period, raters score each component on a 4-point scale (1 = *unsatisfactory*, 2 = *basic*, 3 = *proficient*, 4 = *distinguished*), then average component scores within each domain.

The FFT's developer claims that its two domains are distinct; however, the developer does not report psychometric data. Studies have reported inconsistent factor structures for the FFT, which, depending on the study, vary from several factors, to two, or one factor (e.g., Garrett & Steinberg, 2015; Kane & Staiger, 2012; Lash, Tran, & Huang, 2016; Lockwood, Savitsky, & McCaffrey, 2015). Typically, however, researchers and teacher evaluators combine the two domain scores to create a *Total* score. Given lack of consensus about the FFT's structure, we conducted analyses with both the domain scores and the total score. In the present study, internal consistency reliabilities of scores for reading and mathematics lessons were, respectively, .92 and .90 for Classroom Environment, .80 and .84 for

Instruction, and .90 and .92 for the Total score. The relative reliability (*G*) estimates from generalizability theory analyses, calculated with kindergarten data from the first year of our project, were high. They were .95 and .96 for reading and mathematics Classroom Environment, .87 and .79 for reading and mathematics Instruction, and .94 and .92 for the reading and mathematics Total score (Patrick, French, & Mantzicopoulos, 2019).

Rater Training and Calibration

Lessons were scored by a team of nine raters. Seven raters completed Teachscape's Focus for Observers, an online, self-paced training and certification program for the FFT (Teachscape, n.d.). The certification component was administered by ETS and involved two online tests of approximately three hours each; tests involved viewing and scoring classroom videos and answering multiple-choice questions about the measure. Results are reported as Proficient or Not Proficient, but proficiency criteria were not available. MET project raters, who underwent the same certification process, were required to achieve "at least 50 percent exact match" and "no more than 25 percent [of] ratings "discrepant" (i.e., scores two or more off . . .)" (Kane & Staiger, 2012, p. 22). When two new raters joined the project, the online training and certification were not available; they received similar training from project members. After training, and certification as available, and before coding lessons, project members established interrater agreement by independently viewing and rating four video-recorded lessons. Exact agreement was 78%.

Observing and Scoring Lessons

Each rater scored videos following a unique, order-specific schedule of randomly selected lessons, thus controlling for order effects. Raters also scored reading and mathematics lessons from each teacher, to prevent systematic rater × teacher bias. We checked for rater drift throughout the scoring process by assigning a second team member to score approximately 15% of lessons. Exact interrater agreement was 80%.

Student Measures and Procedure

Achievement

We assessed students' reading and mathematics achievement with two types of measures. Although neither measure is equivalent to state standardized tests—which are not an option for the early grades—they provide complementary views of students' knowledge and skills. The standardized, norm-referenced instrument is administered external to the teacher; like other standardized tests, it reflects broad competencies but is distal to the curriculum (Hickey, Zuiker, Taasoobshirazi, Schafer, & Michael, 2006).

Thus, it shares some similarities with the state achievement tests used in later grades. We also used a teacher-rated, standards-based measure of students' progress toward mastering state standards, consistent with evaluation systems that assess achievement with criterion-referenced learning objectives (Steinberg & Kraft, 2017). Standards-based assessments likely reflect the goals of classroom instruction and the competencies that teachers are likely to target, because the curricula teachers follow are linked closely to state standards.

Standardized achievement. Researchers administered four subtests of the Woodcock-Johnson Tests of Achievement III (WJ-III; Woodcock, McGrew, & Mather, 2001) to students in individual sessions twice in the year: early Fall (approximately six weeks into the school year) and late Spring (April and May). We created a *Reading* composite by averaging the WJ-III Letter Word Recognition and Passage Comprehension subtest scores and a *Math Reasoning* composite by averaging the Applied Problems and Quantitative Concepts subtest scores, as outlined in the technical manual (McGrew & Woodcock, 2001). With 5- and 6-year old children, the subtests' median internal consistency reliabilities range from .88 to .93 (McGrew & Woodcock, 2001).

Standards-based achievement. We developed criterion-referenced measures of student progress toward meeting the state standards for reading and mathematics by creating items to address each of the state's standards for kindergarten reading and mathematics. For example, the item *Understand structural elements of text (e.g., identify genre, define author & illustrator)* represented the standard "Structural elements and organization" (Indiana DOE, 2014, p. 3). Items were rated from 1 (= *does not demonstrate yet*) through 5 (= *independent mastery*). Teachers rated students' progress in late fall (November) and late spring (May).

Reading standards-based achievement (SBAch) was measured with 10 items addressing concepts about print, phonological awareness, vocabulary knowledge, writing conventions, reading fluency, speaking and listening, and understanding structural elements of texts. The internal consistency reliabilities ranged from .95 to .96 (fall and spring of both years). Mathematics SBAch was assessed with seven items addressing number sense, solving real-world problems with numbers, concepts of time, measurement, geometry, and data analysis skills. The internal consistency reliabilities ranged from .89 to .94.

Motivation

Teachers rated students' motivation to learn reading and mathematics in late spring using two subscales of the Teacher Rating Scale of Children's Motivation (Mantzicopoulos, French, & Patrick, 2019). The items refer to key behavioral indicators of students' motivation, which all currently prominent social-cognitive motivation theories seek to explain (Schunk, Meece, & Pintrich, 2014; Wentzel & Brophy, 2014). Items were rated on a 1 (= *not at all*) through 5 (= *a great deal*) scale. We developed the Teacher Rating Scale of Children's Motivation by adapting the Teacher Rating Scale of Children's Motivation for Science (Mantzicopoulos, Patrick, & Samarapungavan, 2013), which was developed for kindergarten, by substituting the word "science" to "reading," "writing," or "math." Evidence of validity for the science version includes significant correlations of both subscales with kindergarteners' reported perceived competence in and liking of science and their researcher-assessed science knowledge; correlations were positive with Interest and negative with Need for Support (Mantzicopoulos et al., 2013).

Interest. The Interest subscale measures students' effort, enthusiasm, and interest in either learning to read and write or do mathematics (e.g., "How excited or enthusiastic is she or he about reading [writing, doing math]?"). Reading Interest contains eight items (four each for reading and writing; αs = .96–.97 across fall and spring of both years) and Math Interest contains four items (αs = .93–.95). Items were identical except for the content area referred to.

Need for Support. The Need for Support subscale includes items about students' frustration, giving up when work is hard, and their need for encouragement (e.g., "How likely is she or he to give up when reading [writing, math] is hard?"). Need for Support in reading contains eight items (four reading and four writing; αs = .92–.95) and Need for Support in mathematics has four items (αs = .88–.92). Items were also identical except for the content area named.

Analysis Plan

To avoid repetition here and in the Results section, we note that we conducted all analyses twice—once for reading and once for mathematics.

We averaged each teacher's ratings across their eight lessons to create composite Classroom Environment, Instruction, and Total scores. Averaging scores across multiple observations produces estimates that are more stable compared with those derived from single observations (Whitehurst et al., 2014). As we noted in the previous section, we created scores for both the *Classroom Environment* and *Instruction* domains, in addition to the *Total* FFT score.

We first examined descriptive statistics and correlations separately for the FFT (Level 2) and student measures (Level 1). We then conducted a series of multilevel models, with students nested within teachers, to identify relations between teacher scores and students' achievement and motivation, controlling for students' sex, ethnicity, FRL, entering standardized achievement, and fall teacher-rated SBAch. We did not include teacher experience in the models because it was not correlated with FFT scores (see Results section).

Multilevel Models

We estimated a series of models for each of the student end-of-year achievement (i.e., criterion-referenced and standardized) and motivation (interest and need for support) outcomes. This involved first estimating the *unconditional or null model* for each outcome measure. We used this model to calculate the intraclass correlation, or percentage of variance in the outcome accounted for by differences between teachers (i.e., at Level 2).

Next, we investigated associations between teachers' FFT scores and their students' achievement and motivation. This involved estimating four sets of conditional models (i.e., one set of models for each student outcome).

Model 1. We included only the set of student-level covariates—sex, ethnicity, FRL, and fall content-specific achievement, both standardized and criterion-referenced—in Model 1. These estimates indicate the extent to which student demographics and fall achievement are related to student- and teacher-level differences in end-of-year achievement and motivation.

Models 2 to 4. We estimated the teacher-level FFT scores for *Classroom Environment, Instruction*, and *Total* scores separately, in three different models, because these scores were highly correlated. Specifically, to the student covariates (Model 1) we added Classroom Environment scores only in Model 2, Instruction scores only in Model 3, and the Total score only in Model 4. We compared Models 2 to 4 with the covariate-only Model 1. For Models 1 to 4, we also examined the change in between-teacher variance explained relative to the null model. We did not include teacher experience because, as we report in the Results section, it was not associated with FFT scores; this finding is consistent with those of others (e.g., Kimball et al., 2004).

We entered the teacher-level FFT scores and student-level covariates as fixed effects. All variables were grand mean centered, because our research questions were focused on Level 2 variables (Enders & Tofighi, 2007). We used restricted maximum likelihood estimation to obtain the parameter estimates and employed maximum likelihood estimation to obtain the deviance estimates for model comparison purposes (Snijders & Bosker, 2012).

We used the following criteria to identify the best fitting model: (1) change in within and between variance estimates; (2) deviance statistics; (3) R^2 values for both within and between variance for the models, as defined by Raudenbush and Bryk (2002); (4) the significance of the predictors; and (5) effect sizes associated with significant effects. In the two-level model, R^2 between and within values represent the change in the available variance among teachers (Level 2) and students (Level 1), respectively. The R^2 statistic is useful in the model building process because it provides a way to compare increases in the within and between variances explained from one model to another. The effect sizes allowed for understanding the association of a 1 standard deviation increase in FFT scores with the *SD* increase in student outcomes. To judge if these effects were meaningful, we were guided by Hattie's (2008) extensive meta-analytic findings that "teachers average an effect of $d = .20$ to $d = .40$ on student achievement" (p. 31). We also compared the effect sizes with those found in other studies of the FFT.

Results

Descriptive Statistics and Correlations

Teachers' Scores

We show the descriptive statistics and correlations for teachers' Classroom Environment, Instruction, and Total scores in reading and mathematics in Table 1. Mean scores were between 2.23 and 2.82, indicating teachers were, on average, classified as Developing. Total reading and mathematics scores were correlated highly ($r = .83$), as were domain scores both within and across content areas. Correlations between Classroom Environment and Instruction scores were .69 for reading and .78 for mathematics, respectively. Reading and mathematics Classroom Environment scores were correlated, $r = .84$; $r = .74$ for Instruction. Teacher experience was not related to FFT scores; *r*s ranged from −.03 to .10.

We examined whether there were either differences between teachers' Classroom Environment and Instruction scores in the same content area or differences between scores for the same domain across content areas. We conducted four paired-samples *t* tests, and, because we performed multiple comparisons, used the Dunn-Bonferonni correction to keep the Type I error at $\alpha = .05$; therefore, we evaluated each of the four comparisons at $\alpha = .0125$. Classroom Environment scores were significantly higher than those for Instruction in both reading ($t = 26.03$, $df = 80$, $p < .0001$, $d = 2.26$) and mathematics ($t = 30.20$, $df = 80$, $p < .0001$, $d = 2.24$). In addition, Classroom Environment scores were higher for reading than for mathematics ($t = 3.19$, $df = 80$, $p = .002$, $d = 0.17$), as were Instruction scores ($t = 3.51$, $df = 80$, $p = .001$, $d = 0.29$).

Students' Achievement and Motivation

We show the descriptive statistics for and correlations between student achievement, motivation, and covariates in Table 2. Fall and spring standardized achievement scores were correlated significantly (*r*s = .75 and .77, $p < .01$, for reading and mathematics, respectively), as was teacher-rated SBAch

Table 1

Descriptive Statistics and Correlations for FFT Domain and Total Scores for Reading and Mathematics and Teacher Experience

					Reading			Mathematics		
FFT Domain Component	*M*	*SD*	Minimum Score	Maximum Score	1	2	3	4	5	6
Reading										
1. Classroom Environment	2.82	0.22	2.06	3.13	—					
2. Instruction	2.30	0.24	1.66	2.72	.69**	—				
3. Total	2.56	0.21	1.86	2.86	.91**	.93**	—			
Mathematics										
4. Classroom Environment	2.78	0.25	1.91	3.03	.84**	.60**	.78**	—		
5. Instruction	2.23	0.24	1.59	2.78	.69**	.74**	.78**	.78**	—	
6. Total	2.51	0.23	1.75	2.89	.81**	.71**	.83**	.95**	.94**	—
Years of teaching experience	13.07	9.46	1.00	40.00	.10	.05	.08	.09	−.03	.04

Note. FFT = Framework for Teaching
*p < .05. **p < .01.

Table 2

Descriptive Statistics and Correlations for Student-Level Demographics and Reading and Mathematics Achievement and Motivation

				Fall		Spring			
Variable	1	2	3	4	5	6	7	8	9
1. Sex[a]	—	−.01	.01	.02	.08**	.06*	.11*	.24**	−.17**
2. FRL[b]	−.01	—	.23**	.31*	.27**	.32**	.25**	.20**	−.24**
3. Ethnicity[c]	.01	.23**	—	.07**	.11**	.08**	.11**	.03	.10**
Fall									
4. Standardized achievement	−.00	.33**	.29**	—	.56**	.75**	.44**	.30**	−.35**
5. Standards-based achievement	.04	.27**	.12**	.49**	—	.56**	.50**	.39**	−.39**
Spring									
6. Standardized achievement	.01	.32**	.26**	.77**	.45**	—	.58**	.41**	−.46**
7. Standards-based achievement	.01	.25**	.15**	.53**	.43**	.58**	—	.69**	−.73**
8. Interest	.04	.19**	.07*	.40**	.35**	.44**	.69**	—	−.75
9. Need for support	−.02	−.22**	−.12**	−.45**	−.35**	−.49**	−.72**	−.74**	—
Reading, *M* (*SD*)				8.76 (4.17)	2.60 (0.93)	18.42 (6.12)	4.12 (0.89)	3.95 (1.00)	2.24 (1.12)
Mathematics, *M* (*SD*)				12.14 (3.18)	2.48 (0.90)	16.27 (2.71)	4.23 (0.86)	4.18 (0.91)	1.95 (1.07)

Note. Coefficients above the diagonal refer to reading and coefficients below the diagonal refer to mathematics. FRL = free or reduced-cost lunch.

[a] Male = 0, female = 1.

[b] FRL = 0, self-pay lunch = 1.

[c] Black, Hispanic, or Other = 0, White = 1.

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

(*r*s = .50 and .43, $p < .01$, for reading and mathematics, respectively). In both semesters, and for both content areas, standardized achievement and SBAch were also correlated significantly (*r*s = .49–.58, *p*s < .01). Other correlations were also in the expected direction and strength; none were large enough to raise multicollinearity concerns.

Predicting Student Achievement and Motivation From Teacher FFT Scores

As we noted in the analysis plan, we considered the association of students' content-specific achievement and motivation with teachers' FFT scores in the corresponding subject. Specifically, we estimated eight sets of multilevel models (i.e., two achievement and two motivation outcomes, each for reading and mathematics), nesting students in teachers.

Null Models

Student Achievement

The intraclass correlations, estimated from the null models, indicated that 26.78% and 21.00% of the variance in student standardized reading and mathematics achievement, respectively, were at the teacher level. Between-teacher variance in SBAch was 16.06% for reading and 25.47% for mathematics.

Student Motivation

Approximately one fifth of the variance in student motivation was between teachers. Specifically, between-teacher variance in student interest was 18.30% for reading and 18.73% for mathematics. The Level 2 variance in students' need for support was 20.96% for reading and 24.90% for mathematics.

Multilevel Models

We present the associations of teacher reading and mathematics FFT scores with student achievement (Tables 3 and 4) and motivation (Tables 5 and 6), while accounting for students' sex, FRL, ethnicity, and entry standardized and fall SBAch in the corresponding content area. In each table, we show the unstandardized reading and mathematics estimates for four models: Model 1 includes student-level covariates only; Model 2 adds Classroom Environment to the covariates; Model 3 adds Instruction, without Classroom Environment, to the covariates; and Model 4 includes the Total score and covariates, without either domain score.

Standardized Achievement

As shown in Table 3, students' end-of-year composite reading achievement was related to their standardized and standards-based reading

Table 3

Parameter Estimates for Student Background and Teacher FFT Scores Predicting Students' Year-End Standardized Achievement

	Reading				Mathematics			
Variable	Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
Fixed parameters								
Intercept	18.20** (0.26)	18.20** (0.25)	18.20** (0.25)	18.20** (0.25)	16.17** (0.09)	16.17** (0.09)	16.17** (0.09)	16.17** (0.09)
Student variables								
Sex	0.10 (0.20)	0.09 (0.20)	0.09 (0.20)	0.09 (0.20)	0.01 (0.09)	0.01 (0.09)	0.01 (0.09)	0.01 (0.09)
FRL	0.65** (0.22)	0.63** (0.22)	0.62** (0.22)	0.62** (0.22)	0.21* (0.10)	0.20* (0.10)	0.20* (0.10)	0.19† (0.10)
Black	0.25 (0.34)	0.26 (0.34)	0.24 (0.34)	0.26 (0.34)	0.11 (0.16)	0.11 (0.16)	0.10 (0.16)	0.10 (0.16)
Hispanic	−0.45 (0.31)	−0.41 (0.31)	−0.44 (0.31)	−0.41 (0.31)	−0.14 (0.15)	−0.13 (0.15)	−0.14 (0.15)	−0.14 (0.15)
Other minority	0.45 (0.39)	0.45 (0.39)	0.44 (0.39)	0.45 (0.39)	−0.06 (0.18)	−0.07 (0.18)	−0.08 (0.18)	−0.07 (0.18)
Fall Stdz Ach	0.70** (0.03)	0.70** (0.03)	0.70** (0.03)	0.70** (0.03)	0.56** (0.02)	0.56** (0.02)	0.55** (0.02)	0.56** (0.02)
Fall SBAch	2.16** (0.17)	2.17** (0.17)	2.17** (0.17)	2.17** (0.17)	0.50** (0.08)	0.50** (0.07)	0.51** (0.08)	0.50** (0.08)
FFT scores								
Classroom Environment		2.60* (1.15)				0.82* (0.36)		
Instruction			3.05** (1.04)				0.66† (0.38)	
Total				3.37** (1.74)				0.84* (0.39)
Variance components								
Student (within)	11.74** (0.48)	11.74** (0.48)	11.74** (0.48)	11.73** (0.48)	2.44** (0.10)	2.44** (0.10)	2.44** (0.10)	2.44** (0.10)
Teacher (between)	4.52** (0.90)	4.31** (0.86)	4.10 (0.83)	4.13** (0.83)	0.48** (0.11)	0.45** (0.10)	0.47** (0.11)	0.46** (0.10)
Model fit								
−2LL (deviance)	6971.513	6966.396	6963.031	6963.460	4861.734	4856.538	4858.614	4857.088
R^2 within	58.10%[a]	0.00%[b]	0.00%[b]	0.09%[b]	58.50%[a]	0.00%[b]	0.00%[b]	0.00%[b]
R^2 between	—	4.64%[b]	9.29%[b]	8.63%[b]	—	6.25%[b]	2.08%[b]	4.17%[b]

Note. Standard errors are in parentheses. FRL = free or reduced-cost lunch; FFT = Framework for Teaching; Stdz Ach = standardized achievement; SBAch = standards-based achievement; −2LL = minus 2 log likelihood.

[a] Value is compared with the null model (not shown).

[b] Value represents the reduction in variance when Model 1 values are compared with those in Models 2 to 4.

[c] Value not calculated when R^2 is higher compared with Model 1 (i.e., variance is not reduced).

$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$.

Table 4

Parameter Estimates for Student Background and Teacher FFT Scores Predicting Students' Year-End Standards-Based Achievement

	Reading				Mathematics			
Variable	Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
Fixed parameters								
Intercept	4.11** (0.06)	4.11** (0.06)	4.11** (0.06)	4.11** (0.06)	4.21** (0.06)	4.21** (0.06)	4.21** (0.06)	4.21** (0.06)
Student variables								
Sex	0.11** (0.03)	0.10** (0.03)	0.11** (0.03)	0.10** (0.03)	0.02 (0.03)	0.01 (0.03)	0.02 (0.03)	0.02 (0.03)
FRL	0.14** (0.04)	0.14** (0.04)	0.14** (0.04)	0.14** (0.04)	0.12** (0.04)	0.12** (0.04)	0.12** (0.04)	0.12** (0.04)
Black	−0.03 (0.06)	−0.03 (0.06)	−0.03 (0.06)	−0.03 (0.06)	0.12* (0.06)	0.13* (0.06)	0.12* (0.06)	0.12* (0.06)
Hispanic	−0.11* (0.05)	−0.10† (0.05)	−0.10† (0.05)	−0.10† (0.05)	0.02 (0.05)	0.02 (0.05)	0.02 (0.05)	0.02 (0.05)
Other minority	−0.01 (0.07)	−0.01 (0.07)	−0.01 (0.07)	−0.01 (0.07)	0.07 (0.06)	0.07 (0.06)	0.07 (0.06)	0.07 (0.06)
Fall Stdz Ach	0.02** (0.01)	0.02** (0.01)	0.02** (0.01)	0.02** (0.01)	0.11** (0.01)	0.11** (0.01)	0.11** (0.01)	0.11** (0.01)
Fall SBAch	0.59** (0.03)	0.59** (0.03)	0.59** (0.03)	0.59** (0.03)	0.34** (0.03)	0.34** (0.03)	0.34** (0.03)	0.34** (0.03)
FFT scores								
Classroom Environment		0.28 (0.26)				0.22 (0.23)		
Instruction			0.35 (0.24)				0.02 (0.24)	
Total				0.37 (0.27)				0.14 (0.25)
Variance components								
Student (within)	0.35** (0.01)	0.35** (0.01)	0.35** (0.01)	0.35** (0.01)	0.29** (0.01)	0.29** (0.01)	0.29** (0.01)	0.29** (0.01)
Teacher (between)	0.23** (0.04)	0.23** (0.04)	0.23** (0.04)	0.23** (0.04)	0.23** (0.04)	0.23** (0.04)	0.23** (0.04)	0.23** (0.04)
Model fit								
−2LL (deviance)	2526.488	2525.319	2524.307	2524.541	2276.598	2275.657	2276.592	2276.278
R^2 within	47.76%[a]	0.00%[b]	0.00%[b]	0.00%[b]	48.21%[a]	0.00%[b]	0.00%[b]	0.00%[b]
R^2 between	—	0.00%[b]	0.00%[b]	0.00%[b]	—	0.00%[b]	0.00%[b]	0.00%[b]

Note. Standard errors are in parentheses. FRL = free or reduced-cost lunch; FFT = Framework for Teaching; Stdz Ach = standardized achievement; SBAch = standards-based achievement; −2LL = minus 2 log likelihood.

[a] Value is compared with the null model (not shown).

[b] Value represents the comparison of Model 1 values with those in Models 2 and 3.

$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$.

Table 5

Parameter Estimates for Student Background and Teacher FFT Scores Predicting Students' Year-End Interest

	Reading				Mathematics			
Variable	Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
Fixed parameters								
Intercept	3.93** (0.06)	3.93** (0.06)	3.93** (0.06)	3.93** (0.06)	4.17** (0.05)	4.17** (0.05)	4.17** (0.05)	4.17** (0.05)
Student variables								
Sex	0.41** (0.04)	0.41** (0.04)	0.41** (0.04)	0.41** (0.04)	0.09* (0.04)	0.09* (0.04)	0.09* (0.04)	0.09* (0.04)
FRL	0.19** (0.05)	0.18** (0.05)	0.18** (0.05)	0.18** (0.05)	0.15** (0.05)	0.15** (0.05)	0.15** (0.05)	0.15** (0.05)
Black	0.07 (0.07)	0.07 (0.07)	0.07 (0.07)	0.07 (0.07)	0.17* (0.07)	0.17* (0.07)	0.17* (0.07)	0.17* (0.07)
Hispanic	−0.04 (0.07)	−0.04 (0.07)	−0.04 (0.07)	−0.04 (0.07)	0.11 (0.07)	0.10 (0.07)	0.11 (0.07)	0.10 (0.07)
Other minority	0.13 (0.09)	0.13 (0.09)	0.13 (0.09)	0.13 (0.09)	0.16* (0.08)	0.16* (0.08)	0.16* (0.08)	0.16* (0.08)
Fall Stdz Ach	$0.01^{\dagger}$ (0.01)	$0.01^{\dagger}$ (0.01)	$0.01^{\dagger}$ (0.01)	$0.01^{\dagger}$ (0.01)	0.10** (0.01)	0.10** (0.01)	0.10** (0.01)	0.10** (0.01)
Fall SBAch	0.48** (0.04)	0.48** (0.04)	0.48** (0.04)	0.48** (0.04)	0.26** (0.04)	0.26** (0.04)	0.26** (0.04)	0.26** (0.04)
FFT scores								
Classroom Environment		0.07 (0.28)				−0.15 (0.22)		
Instruction			0.24 (0.25)				−0.32 (0.23)	
Total				0.19 (0.29)				−0.26 (0.24)
Variance components								
Student (within)	0.57** (0.02)	0.57** (0.02)	0.57** (0.02)	0.57** (0.02)	0.49** (0.02)	0.49** (0.02)	0.49** (0.02)	0.49** (0.02)
Teacher (between)	0.25** (0.05)	0.25** (0.05)	0.25** (0.05)	0.25** (0.05)	0.20** (0.04)	0.21** (0.04)	0.20** (0.04)	0.20** (0.04)
Model fit								
−2LL (deviance)	3123.935	3123.869	3123.006	3123.482	2891.945	2891.460	2890.020	2890.736
R^2 within	31.33%^{a}	0.00%^{b}	0.00%^{b}	0.00%^{b}	28.99%^{a}	0.00^{b}	0.00%^{b}	0.00%^{b}
R^2 between	—	0.00%^{b}	0.00%^{b}	0.00%^{b}	—	—^{c}	0.00%^{b}	0.00%^{b}

Note. Standard errors are in parentheses. FRL = free or reduced-cost lunch; FFT = Framework for Teaching; Stdz Ach = Standardized achievement, SBAch = standards-based achievement; −2LL = minus 2 log likelihood.

^{a}Value is compared with the null model (not shown).

^{b}Value represents the comparison of Model 1 values with those in Models 2 to 4.

^{c}Value not calculated when R^2 is higher compared with Model 1 (i.e., variance is not reduced).

$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$.

Table 6

Parameter Estimates for Student Background and Teacher FFT Scores Predicting Students' Year-End Need for Support

	Reading				Mathematics			
Variable	Model 1	Model 2	Model 3	Model 4	Model 1	Model 2	Model 3	Model 4
Fixed parameters								
Intercept	2.27** (0.07)	2.27** (0.07)	2.27** (0.07)	2.27** (0.07)	1.98** (0.06)	1.98** (0.07)	1.98** (0.07)	1.98** (0.07)
Student variables								
Sex	−0.33** (0.05)	−0.32** (0.05)	−0.32** (0.05)	−0.32** (0.05)	−0.08† (0.04)	−0.08† (0.04)	−0.08† (0.04)	−0.08† (0.04)
FRL	−0.22** (0.05)	−0.22** (0.05)	−0.22** (0.05)	−0.22** (0.05)	−0.15** (0.05)	−0.15** (0.05)	−0.15** (0.05)	−0.15** (0.05)
Black	0.00 (0.08)	−0.00 (0.08)	0.00 (0.08)	0.00 (0.08)	−0.15† (0.08)	−0.15† (0.08)	−0.15† (0.08)	−0.15† (0.08)
Hispanic	0.01 (0.07)	0.00 (0.07)	0.01 (0.07)	0.00 (0.07)	−0.12 (0.07)	−0.12 (0.07)	−0.12 (0.07)	−0.12 (0.07)
Other minority	−0.04 (0.09)	−0.04 (0.09)	−0.04 (0.09)	−0.04 (0.09)	−0.08 (0.09)	−0.08 (0.09)	−0.08 (0.09)	−0.08 (0.09)
Fall Stdz Ach	−0.02** (0.01)	−0.02** (0.01)	−0.02** (0.01)	−0.02** (0.01)	−0.13** (0.01)	−0.13** (0.01)	−0.13** (0.01)	−0.13** (0.01)
Fall SBAch	−0.57** (0.04)	−0.57** (0.04)	−0.57** (0.04)	−0.57** (0.04)	−0.27** (0.04)	−0.27** (0.04)	−0.27** (0.04)	−0.27** (0.04)
FFT scores								
Classroom Environment		−0.23 (0.33)				−0.09 (0.27)		
Instruction			−0.05 (0.30)				0.07 (0.28)	
Total				−0.15 (0.34)				−0.02 (0.29)
Variance components								
Student (within)	0.65** (0.03)	0.65** (0.03)	0.65** (0.03)	0.65** (0.03)	0.59** (0.02)	0.59** (0.02)	0.59** (0.02)	0.59** (0.02)
Teacher (between)	0.37** (0.07)	0.37** (0.07)	0.37** (0.07)	0.37** (0.07)	0.30** (0.05)	0.30** (0.05)	0.30** (0.05)	0.30** (0.05)
Model fit								
−2LL (deviance)	3306.718	3306.226	3306.688	3306.512	3139.854	3139.733	3139.793	3139.850
R^2 within	34.33%^{a}	0.00%^{b}	0.00%^{b}	0.00%^{b}	31.40%^{a}	0.00%^{b}	0.00%^{b}	0.00%^{b}
R^2 between	—	0.00%^{b}	0.00%^{b}	0.00%^{b}	—	0.00%^{b}	0.00%^{b}	0.00%^{b}

Note. Standard errors are in parentheses. FRL = free or reduced-cost lunch; FFT = Framework for Teaching; Stdz Ach = Standardized achievement; SBAch = standards-based achievement.

[a] Value is compared with the null model (not shown).

[b] Value represents the comparison of Model 1 values with those in Models 2 to 4.

$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$.

achievement at the start of the year (γs = 0.70 and 2.16, respectively; *p*s < .01) and paid lunch status (γ = 0.65, p < .01). These covariates explained 58.10% of the available student-level variance in reading achievement (i.e., 42.5% of the total variance). Classroom Environment (Model 2) predicted standardized reading achievement (γ = 2.60, p < .05) and explained 4.64% of the available between-teacher variance. Instruction (Model 3) was also related significantly to reading achievement (γ = 3.05, p < .01); it explained 9.29% of the between-teacher variance, and 2.49% of the total variance, in standardized reading achievement. The FFT Total score (Model 4) also predicted reading achievement (γ = 3.37, p < .01) but at 8.63% explained less Level 2 variance than did Model 3. The effect sizes associated with these analyses were 0.09 (Classroom Environment), 0.12 (Instruction), and 0.13 (FFT Total). Thus, a 1-*SD* increase in the FFT Classroom Environment, Instruction, and Total score is associated with increases in students' standardized reading achievement scores well below the average range of effect sizes (0.20–0.40) reported by Hattie (2008). The differences in the magnitude of the effect sizes and variance accounted for did not suggest that one model was "best fitting" compared with the others.

Students' standardized math reasoning at the end of the year was related positively to their standardized and standards-based mathematics achievement at the start of the year (γs = 0.56 and 0.50, respectively; *p*s < .01) and to paid lunch status (γ = 0.21, p < .05). These covariates explained 58.50% of the student-level variance in math reasoning (i.e., 46.2% of the total variance). Classroom Environment (Model 2) scores predicted math reasoning (γ = 0.82, p < .05); it explained 6.25% of the teacher-level variance and 1.31% of the total variance in standardized math reasoning. The Total score (Model 4) also predicted math achievement (γ = 0.84, p < .05), and explained 4.17% of the Level 2 variance. The association between students' math reasoning and teacher Instruction (Model 3) approached significance (γ = 0.66, p < .10), explaining 2.08% of the between-teacher variance in students' math reasoning. The effect sizes associated with these analyses were 0.08 (Classroom Environment), 0.06 (Instruction), and 0.07 (FFT Total). Thus, a 1-*SD* increase in the FFT Classroom Environment, Instruction, and Total score is associated with negligible increases in students' standardized mathematics achievement scores. Like reading achievement, the slight differences in the magnitude of the effect sizes and variance accounted for did not suggest that one model was "best fitting" compared with others.

Standards-Based Achievement

As shown in Table 4, students' end-of-year reading SBAch was predicted by their entering standardized reading achievement and fall standards-based reading achievement (γs = 0.02 and 0.59, respectively; *p*s < .01). Additionally, it was associated with being female (γ = 0.11, p < .01) and paid lunch status (γ = 0.14, $p < .01$), and related negatively to being Hispanic ($\gamma = -0.11$, $p < .05$). These covariates accounted for 47.76% of the available student level variance, and 40.1% of the total variance, in reading SBAch. Neither teachers' Classroom Environment, Instruction, nor Total scores (in Models 2, 3, and 4, respectively) were associated significantly with reading SBAch; γs = 0.28, 0.35, and 0.37, respectively. Each of the FFT scores explained 0.00% of the available Level 2 variance, which was 16.6%. None of the models' parameter estimates were statistically significant, and the model fit indices were comparable. That is, none of the FFT scores significantly predicted between-teacher differences in students' standards-based reading achievement.

Similarly, students' end-of-year standards-based math achievement was predicted by their standardized and standards-based math achievement at the beginning of the year (γs = 0.11 and 0.34, respectively; ps < .01) and self-paid lunch status (γ = 0.12, $p < .01$). Additionally, it was related to being Black (γ = 0.12, $p < .05$). These covariates accounted for 48.21% of the student-level variance (35.9% of the total variance in this outcome). As with reading, neither teachers' Classroom Environment, Instruction, nor Total scores (in Models 2, 3, and 4, respectively) predicted mathematics SBAch; the teacher-level variance explained by each was 0.0%.

Student Interest

As shown in Table 5, students' end-of-year interest in reading was related significantly to their fall standards-based reading achievement (γs = 0.48, $p < .01$), being female (γ = 0.41, $p < .01$), and self-paid lunch status (γ = 0.19, $p < .01$). These covariates explained 31.33% of the available student level variance in reading interest (i.e., 25.6% of the total variance). Adding the Classroom Environment, Instruction, or Total scores (Models 2, 3, and 4, respectively) did not explain any between-teacher variance in reading interest.

Students' end-of-year interest in mathematics was related significantly to their fall math SBAch (γ = 0.26; $p < .01$), standardized achievement (γ = 0.10, $p < .01$), paid lunch (γ = 0.15 $p < .01$), and being female (γ = 0.09, $p < .05$), Black (γ = 0.17, $p < .05$), and Other minority (γ = 0.16, $p < .05$). The student covariates explained 29.0% of the Level 1 variance in math interest (i.e., 23.6% of the total variance). Neither teachers' Classroom Environment, Instruction, nor Total scores (in Models 2, 3, and 4, respectively) predicted interest in mathematics.

Student Need for Support

As shown in Table 6, students' end-of-year need for support in reading was related negatively to their fall reading SBAch ($\gamma = -0.57$, $p < .01$), standardized achievement entering kindergarten ($\gamma = -0.02$, $p < .01$), being male ($\gamma = -0.33$, $p < .01$), and receiving FRL ($\gamma = -0.22$, $p < .01$). These covariates explained 34.33% of the student-level variance in need for support in reading (i.e., 27.1% of the total variance). As with interest, the FFT scores did not contribute to the Level 2 variance, which was 24.9% for this outcome. Neither teachers' Classroom Environment, Instruction, nor Total scores (in Models 2, 3, and 4, respectively) predicted students' need for support in reading, as rated by their teachers.

Students' need for support in mathematics at the end of the year was associated negatively with their fall math SBAch ($\gamma = -0.27, p < .01$), standardized math achievement entering kindergarten ($\gamma = -0.13$, $p < .01$), and receiving FRL ($\gamma = -0.15, p < .01$). The covariates explained 31.4% of the student-level variance in math need for support (i.e., 23.6% of the total variance). Teachers' Classroom Environment, Instruction, and Total scores (in Models 2, 3, and 4, respectively) were not related to students' need for support in mathematics and did not contribute to the teacher-level variance in this outcome.

Discussion

Our objective with the present study was to address the predictive validity of using FFT scores for teacher evaluation. Our results from kindergarten classrooms suggest that FFT scores capture small, yet measurable, associations between ratings of teachers' practices and their students' end-of-year standardized achievement, a finding that is consistent with results from another sample of kindergarten teachers (Patrick, Mantzicopoulos, & French, 2019). Beyond statistical significance, however, only a very small proportion of the total variability in student standardized achievement was related to ratings of teacher instructional practices—at most, 2.5% for reading achievement and 1.3% for mathematics achievement. The effect sizes (0.06–0.13) are well below the average range (0.20–0.40) reported by Hattie (2008). They are also at the low end of the range of effect sizes found in other FFT studies (i.e., 0.11–0.25; Garrett & Steinberg, 2015; Polikoff & Porter, 2014; Tyler et al., 2010). Together, the findings suggest that the FFT is not sufficiently sensitive to identify the average effect of teachers' instruction on student achievement. Additionally, for either content area, FFT scores did not predict any of the teacher-rated outcomes (i.e., teacher-rated criterion-referenced achievement, students' positive or negative motivation—interest and need for encouragement and support to engage in learning). Thus, the practical significance of the FFT scores for kindergarten appears minimal, at best.

In contrast to the negligible predictive validity of teacher FFT scores, students' family SES (inferred from FRL status), incoming achievement, and fall mastery of content standards—and sometimes sex and ethnicity—accounted for a large proportion of variance in year-end student outcomes: 36% to 46% of the total variance in reading and math achievement. This finding suggests that a more effective route to increasing student achievement may be through high-quality preschool experiences that are available to students regardless of family SES.

Our results are important, given that (1) scores from the FFT are used widely throughout the United States to evaluate teachers' effectiveness, (2) the rationale for including classroom observations like the FFT for teacher evaluation was that they differentiate teachers in terms of their ability to increase students' achievement (e.g., Weisberg et al., 2009), (3) early elementary teachers' observation scores cannot be combined with student achievement data to increase the accuracy of evaluations, and (4) there is very little published evidence, and none from the early grades, indicating that FFT scores predict robust gains in student achievement or motivation with sufficient accuracy to warrant their use for teacher evaluation. Our results are also important because they indicate that the large body of research, particularly from the 1970s and 1980s, showing that OMI scores are not reliably associated with student achievement (see Lavigne & Good, 2014) continue to apply to education today. This consistency in findings is despite many differences, including "the increasing changes in the curriculum, diversity of student populations, changes in the teaching force, and arguably better statistical measures of teachers' impact on student learning" (Good & Lavigne, 2015, p. 4), in addition to different OMI.

Possible Explanations for the Negligible or No Predictive Validity of FFT Scores

We found that teachers' FFT scores explained little-to-no variance in their students' end-of-year achievement or motivation. There are several possible explanations for these findings.

Limited Range of Teacher Measures

One reason may be that the highest of the FFTs four ratings—Distinguished—is not appropriate for kindergarten (Kimball, 2002). Specifically, to achieve this rating students must show initiative and self-monitoring by managing their task time and behavior as they improve their own work, as well as monitor and support other students' engagement and learning (Danielson, 2013); these skills and behaviors are generally not realistic expectations for kindergarteners.

A truncated range of teacher scores is not limited to kindergarten, however. The FFT scores of almost all teachers—93% of those in the MET project (Ho & Kane, 2013)—fall between the score points of 2 and 3 (Sartain, Stoelinga, & Brown, 2011). This narrow range across teachers presents a challenge in meaningfully differentiating among their effectiveness. Moreover, this narrow use of the score scale likely limits the distributions of FFT scores. This restriction of the score distribution, real or artificial, can likely attenuate the relations between FFT scores and outcomes. In-depth validity studies, such as think-aloud protocols, could investigate how raters cognitively process these items within a given domain to understand their use of the score range. Such evidence may inform changes to the scale structure or perhaps rater training and calibration.

Limited Range of Student Measures

In the early grades, student achievement is typically measured in terms of their mastery of specific grade-level performance standards or criteria identified by the state. These standards presumably reflect skills that can realistically be mastered by almost all students in the corresponding grade. Although many students may have mastered some of the later grades' standards, this achievement is generally not measured. Therefore, by the end of the year, the distribution of student achievement relative to the standards is likely small, presenting little variance for teacher ratings to explain. In our findings, the FFT scores were not associated with student SBAch. They were, however, for the norm-referenced Woodcock-Johnson achievement measures, which do not have upper limits for scores; this type of assessment is not used in kindergarten, however.

Restricted range may also be a factor in the lack of association between FFT scores and students' motivation. Although some children have developed maladaptive motivational patterns by the end of kindergarten (Hirvonen et al., 2012; Patrick et al., 2008), the greatest majority are generally motivated to learn and interested in expanding their knowledge (Wigfield & Eccles, 2002). Thus, these restrictions in distributions for various reasons may have limited the strength of the relations we observed.

Unwarranted Assumptions About Links Between Instruction and Achievement

Another reason for our, and others', findings that FFT scores—and OMI scores in general—are poor predictors of student achievement involve overly simplistic assumptions about associations between instruction and learning. Although OMI scores include teacher practices indicated by research as related to achievement, they do not measure all aspects of high-quality instruction. However, because assessments serve as indictors of what is valued, teachers are likely to focus on using practices included in protocols and overlooking those not included. This may result in the omission of potentially effective practices. For example, although data from a national study of mathematics instruction in kindergarten indicate that frequent use of drill and practice is an effective way to increase students' mathematics achievement (Bottia, Moller, Mickelson, & Stearns, 2014; Guarino, Dieterle, Bargagliotti, & Mason, 2013), it would result in low scores on Instruction components. Furthermore, some aspects of instruction may require very low incidence to foster achievement, with no added benefit from more frequent use (Brophy, 1988); this situation is not reflected in typical rating systems whereby higher scores require more prominent use of practices.

Issues With the Stability of Teachers' Practices

A final issue that may have affected the predictive validity of FFT scores is the reliability of teacher scores. Given the well-documented large variability (i.e., low stability or reliability) of teachers' practices (Konstantopoulos, 2014; Lavigne & Good, 2014), observing eight lessons per teacher for each content area and achieving high interrater agreement may not have been sufficient for reliable observation scores.

Differences in Teachers' FFT Scores

Domain Differences

Teachers, on average, were rated higher for creating a positive, respectful, orderly, and well-managed learning environment, as measured by the Classroom Environment domain, than they were in engaging students in learning by giving clear directions and explanations, using questions and discussions, and assessing learning, measured by the Instruction domain. Our results are consistent with those of other studies that used the FFT, where teachers also scored higher on Classroom Environment than Instruction (Tyler et al., 2010). They also mirror others' who used comparable observation measures and found that preschool (Curby, Grimm, & Pianta, 2010), early elementary (Mantzicopoulos, Patrick, Strati, & Watson, 2018), and high school (Praetorius, Pauli, Reusser, Rakoczy, & Klieme, 2014) teachers received higher scores for creating a supportive environment than for instructional quality.

There are plausible explanations for why teachers may score higher for their classroom environment than for their instruction. Behaviors that facilitate supportive classroom environments are appropriate for every lesson, whereas practices included in the Instruction domain, such as assessing student work or engaging in discussions, are not always relevant to a specific lesson. When one (or more) Instruction component is not relevant to a lesson's objectives, that domain's score is necessarily depressed. Furthermore, relational skills needed for a high Classroom Environment score (e.g., maintaining respectful talk and active listening, or communicating high student expectations) may be less difficult to develop than comparably rated skills in the Instruction domain (e.g., matching the cognitive challenge of higher order questions to students' ability, or stepping out of the central mediating role once a discussion has begun).

Content Area Differences

Teachers' Classroom Environment and Instruction scores were higher for reading than for mathematics, which appears to reflect the privileging of reading over all other content areas in the early grades (Rouge et al., 2008). FFT scores explained almost twice the total variance in standardized reading, compared with mathematics, achievement. However, at 2.5% and 1.3%, for reading and math, respectively, the variance explained was negligible. Regardless, our evidence suggests that the FFT is not content independent in kindergarten, and teachers' evaluations are likely affected by the content taught while being observed. Whether reading and mathematics instruction are scored comparably in elementary grades beyond kindergarten is an important question for future research.

Limitations and Directions for Future Research

Some of the limitations of this study involve the differences between the FFT being used for research or practice (i.e., teacher evaluation). For the teachers in our study, the lessons and associated FFT scores were not part of an evaluation system and there were no stakes attached. It is possible that teachers' scores would be different if there were. Interestingly, there is evidence that "evaluators systematically assign higher summative ratings to teachers relative to formative ratings that are decoupled from high-stakes consequences" (Steinberg & Kraft, 2017, p. 384). The variance of scores was similar in both situations, however (Steinberg & Kraft, 2017). Our study also differed from typical practice in that we achieved high interrater agreement, which also, ironically, limits the generalizability of our findings.

Another limitation is that three of the four student outcomes involved teacher ratings. As a reviewer noted, it may be that less effective teachers provide less accurate ratings of their students. Given evidence that teachers rate girls' proficiency in mathematics lower than that of boys with comparable achievement and engagement (Cimpian, Lubienski, Timmer, Makowski, & Miller, 2016), teachers' ratings may be biased against student sex or other demographics.

Our measures of student motivation also involved teacher ratings, rather than responses from children themselves. However, the items referred to specific behavioral indicators of motivation, such as effort, persistence, enthusiasm, and frustration, rather than conjectures of students' internal perceptions. Construct validity comes from the significant correlations of teacher-rated student maladaptive motivation from different teachers between kindergarten and fourth grade (Hirvonen et al., 2012). We believe it is important that motivation be considered a central indicator of teacher effectiveness, in addition to achievement, and, consequently, that researchers include it in their studies. Student-reported motivation will be a necessary part of this research.

Our findings suggest that the FFT is not appropriate for evaluating kindergarten teachers. They also raise questions about how well FFT scores predict achievement and motivation in first, second and even third grade, given Kimball et al.'s (2004) findings that scores were not associated with mathematics or reading achievement in third grade. Other questions involve whether FFT scores interact with other factors, such as ELL status or class size, to predict student outcomes differentially. Research is needed to address these questions. Furthermore, research should examine grade levels separately, rather than assume that FFT scores are grade-level invariant, and predict student outcomes similarly. That said, comparisons are needed to examine the assumption that the FFT scores are invariant across content areas and grade levels.

It is likely that the FFT is not the only OMI with low predictive validity estimates. Kindergarten teachers' Classroom Assessment Scoring System scores also did not predict their students' year-end achievement (Mantzicopoulos et al., 2018). The predictive validity of scores from other observation measures that are prominent in evaluation systems (e.g., Marzano, Carbaugh, Rutherford, & Toth, 2014) has received little consideration, perhaps because they were not included in the MET project. Research investigating OMI is needed urgently. Although it may be "unrealistic . . . to validate every new measure of teaching" (Kane et al., 2013, p. 39), we believe it is crucial to examine at least the most prominent OMI, to ensure that they measure what is intended; that is, that they differentiate among teachers in terms of their students' growth in achievement. Without compelling, independently reviewed evidence that OMI do so, and particularly given the empirical evidence that they do not, we question the wisdom of districts allocating considerable resources to classroom observations for evaluating teachers. Using observations for formative assessment and to guide professional development is a different issue. However, although feedback and support guided by observations can be valuable for teachers, we question the usefulness of conducting costly observations for formative purposes using a protocol where scores either do not predict student outcomes or explain extremely little.

Summary

In summary, our results with the FFT in kindergarten do not support the premise for using observation measures to evaluate teachers: that the instruments measure instruction that leads to gains in student achievement. This is of concern. It is also concerning that there is little to no evidence of predictive validity for FFT scores, and scores from other OMI, in other grade levels and content areas. At present, teachers in the early grades are especially vulnerable to evaluation with inadequate observation measures, because their scores cannot be combined with test scores or student surveys. However, it is possible that recent state-level legislation to prohibit or postpone the use of value-added indices (Close et al., 2018; Croft et al., 2018) may result in, for some states, OMI scores forming the basis of all teachers' evaluations. It is also possible, though, that states may exercise the flexibility afforded them by ESSA and radically pare down teacher evaluation systems, or even dispense with them altogether. We hope that decisions about teacher accountability, and about education in general, are based on sound, independently reviewed empirical evidence rather than the rhetoric of special-interest groups. Our study is an example of such evidence that should be considered prior to decision making that affects teachers, their students, and the communities they serve.

Note

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140664 to Helen Patrick, Panayota Mantzicopoulos, and Brian F. French. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We greatly appreciate the involvement of the kindergarten teachers and their students. We also thank Inok Ahn, Hyeree Cho, Jimena Cosso, Qian Li, ChangChia James Liu, Yaheng Lu, Sam Watson, and Elisha Yuan for their assistance.

References

Alexander, K. L., & Entwisle, D. R. (1988). Achievement in the first two years of school: Patterns and processes. *Monographs of the Society for Research in Child Development*, *53*(2, Serial No. 218), 1–157. doi:10.2307/1166081

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, *84*, 261–271.

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

Bottia, M. C., Moller, S., Mickelson, R. A., & Stearns, E. (2014). Foundations of mathematics achievement: Instructional practices and diverse kindergarten students. *Elementary School Journal*, *115*, 124–150.

Bill & Melinda Gates Foundation. (2010, June). *Working with teachers to develop fair and reliable measures of effective teaching*. Retrieved from https://docs.gatesfoundation.org/documents/met-framing-paper.pdf

Brophy, J. (1973). Stability of teacher effectiveness. *American Educational Research Journal*, *10*, 245–252.

Brophy, J. (1988). Research on teacher effects: Uses and abuses. *Elementary School Journal*, *89*, 3–21.

Brophy, J. (2006). Observational research on generic aspects of classroom teaching. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 755–780). Mahwah, NJ: Lawrence Erlbaum.

Brophy, J., Coulter, C. L., Crawford, J., Evertson, C. M., & King, C. E. (1975). Classroom observation scales: Stability across time and context and relationships with student learning gains. *Journal of Educational Psychology*, *67*, 873–881.

Center on Great Teachers and Leaders. (2013). *Measures of teacher performance* [Databases on state teacher and principal evaluation policies]. Retrieved from http://resource.tqsource.org/stateevaldb/Compare50States.aspx

Chait, R. (2010, March 10). *Removing chronically ineffective teachers: Barriers and opportunities*. Retrieved from https://www.americanprogress.org/issues/education-k-12/reports/2010/03/10/7525/removing-chronically-ineffective-teachers/

Cimpian, J. R., Lubienski, S. T., Timmer, J. D., Makowski, M. B., & Miller, E. K. (2016). Have gender gaps in math closed? Achievement, teacher perceptions, and learning behaviors across two ECLS-K cohorts. *AERA Open*, *2*(4). doi:10.1177/2332858416673617

Close, K., Amrein-Beardsley, A., & Collins, C. (2018, June 5). *State-level assessments and teacher evaluation systems after the passage of the Every Student Succeeds Act: Some steps in the right direction*. National Education Policy Center. Retrieved from https://nepc.colorado.edu/publication/state-assessment

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, *45*, 378–387.

Croft, M., Guffy, G., & Vitale, D. (2018, July). *The shrinking use of growth: Teacher evaluation legislation since ESSA* (ACT Research & Policy Issue Brief). Retrieved from https://www.act.org/content/dam/act/unsecured/documents/teacher-evaluation-legislation-since-essa.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.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C. (2013). *The Framework for Teaching evaluation instrument: 2013 edition*. Princeton, NJ: The Danielson Group.

Dee, T., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from impact. *Journal of Policy Analysis and Management*, *34*, 267–297.

Dweck, C. S. (2002). The development of ability conceptions. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 57–88). San Diego, CA: Academic Press.

Dwyer, C. A. (1998). Psychometrics of Praxis III: Classroom performance assessments. *Journal of Personnel Evaluation in Education*, *12*, 163–187.

Emmer, E., Evertson, C., & Brophy, J. (1979). Stability of teacher effects in junior high classrooms. *American Educational Research Journal*, *16*, 71–75.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121–138.

Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis*, *37*, 224–242.

Good, T. L. (1979). Teacher effectiveness in the elementary school. *Journal of Teacher Education*, *30*(2), 52–64.

Good, T., & Grouws, D. (1977). Teaching effects: A process-product study in fourth-grade mathematics classrooms. *Journal of Teacher Education*, *28*(3), 49–54.

Good, T. L., & Lavigne, A. L. (2015). Issues of teacher performance stability are not new: Limitations and possibilities. *Education Policy Analysis Archives*, *23*(2). doi:10.14507/epaa.v23.1916

Gottfried, A. E., Fleming, J. S., & Gottfried, A. W. (2001). Continuity of academic intrinsic motivation from childhood through late adolescence: A longitudinal study. *Journal of Educational Psychology*, *93*, 3–13.

Griffith, D., & McDougald, V. (2016). *Undue process: Why bad teachers in twenty-five diverse districts rarely get fired*. Washington, DC: Thomas B. Fordham Institute.

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

Guarino, C., Dieterle, S. G., Bargagliotti, A. D., & Mason, W. M. (2013). What can we learn about effective early mathematics teaching? A framework for estimating causal effects using longitudinal survey data. *Journal of Research on Educational Effectiveness*, *6*, 164–198.

Halpin, P. F., & Kieffer, M. J. (2015). Describing profiles of instructional practice: A new approach to analyzing classroom observation data. *Educational Researcher*, *44*, 263–277.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge.

Hickey, D. T., Zuiker, S. J., Taasoobshirazi, G., Schafer, N. J., & Michael, M A. (2006). Three is the magic number: A design-based framework for balancing formative and summative functions of assessment. *Studies in Educational Evaluation*, *32*, 180–201.

Hirvonen, R., Tolvanen, A., Aunola, K., & Nurmi, J. (2012). The developmental dynamics of task-avoidant behavior and math performance in kindergarten and elementary school. *Learning and Individual Differences*, *22*, 715–723.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Retrieved from https://k12education.gatesfoundation.org/download/?Num=2520&filename=MET_Reliability-of-Classroom-Observations_Research-Paper.pdf

Indiana Department of Education. (n.d.-a). *RISE evaluation and development system: Evaluator and teacher handbook—Version 2.0*. Retrieved from https://www.doe.in.gov/sites/default/files/evaluations/rise-handbook-2-0-final.pdf

Indiana Department of Education. (n.d.-b). *RISE w Danielson references*. Retrieved from https://www.doe.in.gov/sites/default/files/evaluations/rise-w-danielson-references.pdf

Indiana Department of Education. (2014). *Kindergarten Indiana academic standards 2014: English/language arts*. Retrieved from https://www.doe.in.gov/sites/default/files/standards/enla/grade-k-ias-2014-update-2017.pdf

Indiana Department of Education. (2019). *DOE Compass* [Database to search school and corporation reports]. Retrieved from https://compass.doe.in.gov/dashboard/overview.aspx

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher perspectives on evaluation reform: Chicago's REACH students. *Educational Researcher*, *44*, 105–116.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching*. Seattle, WA: Bill & Melinda Gates Foundation.

Kaplan, A., & Patrick, H. (2016). Learning environments and motivation. In K. R. Wentzel & D. B. Miele (Eds.), *Handbook of motivation at school* (2nd ed., pp. 251–274). New York, NY: Routledge.

Karabenick, S. A., & Urdan, T. C. (Eds.). (2014). *Advances in motivation and achievement. Volume 18: Motivational interventions*. Bingley, England: Emerald Group.

Kimball, S. M. (2002). Analysis of feedback, enabling conditions and fairness perceptions of teachers in three school districts with new standards-based evaluation systems. *Journal of Personnel Evaluation in Education*, *16*, 241–268.

Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education*, *79*(4), 54–78.

Konstantopoulos, S. (2014). Teacher effects, value-added models, and accountability. *Teachers College Record*, *116*. Article ID 17290.

Lash, A., Tran, L., & Huang, M. (2016, May). *Examining the validity of ratings from a classroom observation instrument for use in a district's teacher evaluation system* (REL 2016-135). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West. Retrieved from https://files.eric.ed.gov/fulltext/ED565904.pdf

Lavigne, A. L., & Good, T. L. (2014). *Teacher and student evaluation: Moving beyond the failure of school reform*. New York, NY: Routledge.

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

Mantzicopoulos, P., French, B. F., & Patrick, H. (2019). The quality of mathematics instruction in kindergarten: Associations with students' achievement and motivation. *Elementary School Journal*, *119*, 651–676.

Mantzicopoulos, P., Patrick, H., & Samarapungavan, A. (2013). Science literacy in school and home contexts: Kindergarteners' science achievement and motivation. *Cognition and Instruction*, *31*, 62–119.

Mantzicopoulos, P., Patrick, H., Strati, A., & Watson, J. S. (2018). Predicting kindergarteners' achievement and motivation from observational measures of teaching effectiveness. *Journal of Experimental Education*, *86*, 214–232.

Martinez, J. F., Schweig, J., & Goldschmidt, P. (2016). Approaches for combining multiple measures of teacher performance: Reliability, validity, and implications for evaluation policy. *Educational Evaluation and Policy Analysis*, *38*, 738–756.

Marzano, R., J., Carbaugh, B., Rutherford, A., & Toth, M. D. (2014). *Marzano Center Teacher Observation protocol for the 2014 Marzano teacher evaluation model*. Retrieved from https://www.learningsciences.com/wp/wp-content/uploads/2017/06/2014-Protocol.pdf

McGrew, K. S., & Woodcock, R. W. (2001). *Technical manual: Woodcock-Johnson III*. Itasca, IL: Riverside.

Meyer, L. A., Linn, R. L., & Hastings, C. N. (1991). Teacher stability from morning to afternoon and from year to year. *American Educational Research Journal*, *28*, 825–847.

Mihaly, K., & McCaffrey, D. F. (2014). Grade-level variation in observational measures of teacher effectiveness. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project* (pp. 9–49). San Francisco, CA: Jossey-Bass.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*, *79*(4), 33–53.

Muñoz, M. A., & Dossett, D. H. (2016). Multiple measures of teaching effectiveness: Classroom observations and student surveys as predictors of student learning. *Planning and Changing*, *47*, 123–140.

National Council of Teachers of Mathematics. (2013, October). *Mathematics in early childhood learning: A position of the National Council of Teachers of Mathematics*. Retrieved from https://www.nctm.org/uploadedFiles/Standards_and_Positions/Position_Statements/Early%20Childhood%20Mathematics%20(2013) .pdf

National Research Council. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academies Press.

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academies Press.

National Research Council. (2009). *Mathematics learning in early childhood: Paths towards excellence and equity*. Washington, DC: National Academies Press.

The New Teacher Project. (2010). *Teacher evaluation 2.0*. Retrieved from https://tntp.org/assets/documents/Teacher-Evaluation-Oct10F.pdf

Patrick, H., French, B. F., & Mantzicopoulos, P. (2019). *The reliability of Framework for Teaching scores in kindergarten*. Manuscript submitted for publication.

Patrick, H., Mantzicopoulos, P. Y., & French, B. F. (2019). *Using classroom observations to measure teachers' effectiveness in promoting student achievement*. Manuscript submitted for publication.

Patrick, H., Mantzicopoulos, P., Samarapungavan, A., & French, B. F. (2008). Patterns of young children's motivation for science and teacher-child relationships. *Journal of Experimental Education*, *76*, 121–144.

Perry, N., Turner, J. C., & Meyer, D. K. (2006). Student engagement in the classroom. In P. Alexander & P. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 327–348). Mahwah, NJ: Lawrence Erlbaum.

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis*, *36*, 399–416.

Praetorius, A., 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.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Reeve, J. (2002). Self-determination theory applied to educational settings. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 183–203). New York, NY: University of Rochester Press.

Rosenshine, B. (1970). The stability of teacher effects upon student achievement. *Review of Educational Research*, *40*, 847–662.

Rouge, E. C., Hansen, J., Muller, P., & Chien, R. (2008). *Evaluation of Indiana Reading First program*. Bloomington, IN: Center for Evaluation & Education Policy.

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011, November). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation*. Chicago, IL: UChicago Consortium on Chicago School.

Schunk, D. H., Meece, J. L., & Pintrich, P. R. (2014). *Motivation in education: Theory, research, and applications* (4th ed.). Upper Saddle River, NJ: Pearson.

Schwinger, M., Wirthwein, L., Lemmer, G., & Steinmayr, R. (2014). Academic self-handicapping and achievement: A meta-analysis. *Journal of Educational Psychology*, *106*, 744–761.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.

Snijders, T., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and applied multilevel analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Steinberg, M. P., & Kraft, M. A. (2017). The sensitivity of teacher performance ratings to the design of teacher evaluation systems. *Educational Researcher*, *46*, 378–396.

Stipek, D. (2002). Good instruction is motivating. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 309–332). San Diego, CA: Academic Press.

Teachscape. (n.d.). *Framework for teaching proficiency system*. Retrieved from https://growththroughlearningillinois.org/Support/TeachscapeFAQ.aspx

Toch, T., & Rothman, R. (2008, January). *Rush to judgment: Teacher evaluation in public education*. Washington, DC: Education Sector. Retrieved from https://www.issuelab.org/resources/1076/1076.pdf

Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2010). Using student performance data to identify effective classroom practices. *American Economic Review*, *100*, 256–260.

U.S. Department of Education. (2007). *State and local implementation of the No Child Left Behind Act. Volume II—Teacher quality under NCLB: Interim report*. Washington, DC: Author.

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

U.S. Department of Education. (2011, September 23). *Fact sheet: Bringing flexibility and focus to education law*. Retrieved from https://obamawhitehouse.archives.gov/the-press-office/2011/09/23/fact-sheet-bringing-flexibility-and-focus-education-law

U.S. Department of Education. (2016). *Every Student Succeeds Act*. Retrieved from https://www.ed.gov/essa?src=rn

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009, June 8). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Washington, DC: New Teacher Project. Retrieved from https://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf

Wentzel, K. R., & Brophy, J. E. (2014). *Motivation students to learn* (4th ed.). New York, NY: Routledge.

White, M. C. (2018). Rater performance standards for classroom observation instruments. *Educational Researcher*, *47*, 492–501.

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. (2014, May). *Evaluating teachers with classroom observations. Lessons learned in four districts*. Retrieved from https://www.brookings.edu/wp-content/uploads/2016/06/Evaluating-Teachers-with-Classroom-Observations.pdf

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*, 68–81.

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 91–120). San Diego, CA: Academic Press.

Wigfield, A., Eccles, J. S., Fredricks, J. A., Simpkins, S., Roeser, R. W., & Schiefele, U. (2015). Development of achievement motivation and engagement. In M. E. Lamb & R. M. Lerner (Eds.), *Handbook of child psychology and developmental science, Vol. 3: Socioemotional processes* (7th ed., pp. 657–700). Hoboken, NJ: John Wiley.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement*. Itasca, IL: Riverside.

Manuscript received November 1, 2018
Final revision received November 3, 2019
Accepted November 4, 2019