The Journal of Experimental Education

ISSN: 0022-0973 (Print) 1940-0683 (Online) Journal homepage: www.tandfonline.com/journals/vjxe20

Predicting Kindergarteners' Achievement and Motivation From Observational Measures of Teaching Effectiveness

Panayota Mantzicopoulos, Helen Patrick, Anna Strati & Jesse S. Watson

To cite this article: Panayota Mantzicopoulos, Helen Patrick, Anna Strati & Jesse S. Watson (2018) Predicting Kindergarteners' Achievement and Motivation From Observational Measures of Teaching Effectiveness, The Journal of Experimental Education, 86:2, 214-232, DOI: 10.1080/00220973.2016.1277338

To link to this article: https://doi.org/10.1080/00220973.2016.1277338

Published online: 23 Feb 2017.

THE JOURNAL OF EXPERIMENTAL EDUCATION
2018, VOL. 86, NO. 2, 214–232
https://doi.org/10.1080/00220973.2016.1277338

Routledge
Taylor & Francis Group

MOTIVATION AND SOCIAL PROCESSES

Predicting Kindergarteners' Achievement and Motivation From Observational Measures of Teaching Effectiveness

Panayota Mantzicopoulos[a], Helen Patrick[a], Anna Strati[b], and Jesse S. Watson[a]

[a]Purdue University, West Lafayette, IN, USA; [b]Aurora University, Aurora, IL, USA

ABSTRACT

We investigated the premise that observation measures of instruction are indicators of effective teaching, using the definition of effectiveness articulated by departments of education: teaching that boosts student achievement. We argued that student motivation is equally as important as achievement in the evaluation of teaching effectiveness (TE); therefore, we examined students' ($N = 145$) achievement and motivation outcomes. We scored 40 lessons (from 10 kindergarten teachers) with two TE observation measures: the content-independent Classroom Assessment Scoring System (CLASS) and the content-specific Reformed Teaching Observation Protocol (RTOP). We found that the two measures' scores were related differently to student outcomes. Instructionally supportive practices (CLASS and RTOP total) predicted achievement and motivation. Emotional support (CLASS) was positively related to motivation but not to achievement. Classroom organization (CLASS) was negatively related to both motivation and achievement. The CLASS total score did not predict student outcomes; its use masked differences across domains of teaching practices.

KEYWORDS

Classroom observation; elementary schools; motivation; science education; teacher practices

CITING THE CRITICAL importance of retaining only effective teachers in the nation's schools, federal accountability policies stipulate that states engage in efforts to evaluate teachers' effectiveness (USDOE, 2009, 2011). As defined in federal documents, effective teachers are those "whose students achieve acceptable rates (e.g., at least one grade level in an academic year) of student growth" (USDOE, 2009, p. 12). In other words, considering that growth is determined from performance on standardized achievement tests, teaching effectiveness (TE) is equated with instruction that produces strong achievement test scores.

To comply with federal mandates, states have engaged in the development and implementation of teacher evaluation systems that call for evidence from measures of student achievement and from observation-based assessments of instruction. Yet, in practice, TE determinations are based largely on observations of teachers' performance in the classroom (Garrett & Steinberg, 2015; Herlihy et al., 2014). For teachers in the early grades, as well as teachers whose students are typically not included in the annual state assessment systems, observers' ratings may account for more than 75% of the TE score (Whitehurst, Chingos, & Lindquist, 2014). Thus, in the current accountability climate, observational measures of instruction—once of interest only to researchers—have become standard tools in the teacher evaluation process. Of note, although teachers typically express dissatisfaction with the use of standardized tests to judge their effectiveness, they tend to view observational assessments as appropriate for this purpose (Kimball & Milanowski, 2009; Sullivan, 2012).

CONTACT Panayota Mantzicopoulos mantzi@purdue.edu Department of Educational Studies, Beering Hall, Purdue University, 100 N. University Street, West Lafayette, IN 47907, USA.

There is a dearth of evidence, particularly in the early grades, showing that measures of TE *do* predict students' achievement. Such evidence is necessary, given the high stakes (e.g., teachers' promotion, termination) attached to TE scores. A goal of this study, therefore, was to examine how well TE scores, derived from observation of teachers' instructional practices, predict kindergarten students' achievement.

A second issue warranting attention is policy makers' consideration of TE solely in terms of achievement. We argue that because motivation is equally as important as achievement for academic success (Wigfield & Cambria, 2010), it is also crucial to evaluate TE with reference to students' motivation. Our second goal, therefore, was to examine how well observational TE scores predict students' motivation.

Our third objective was to consider whether student outcomes are differentially predicted by the type of observational instrument (content-independent or content-specific) used to rate instructional practices. The observation-based measures that are recommended by state departments of education for assessing TE are overwhelmingly content-independent (Center on Great Teachers and Leaders, 2013). That is, they are considered appropriate for evaluating instruction in any subject area, even though their adequacy—relative to content-specific measures (i.e., developed specifically for each subject area)—has received little empirical scrutiny and none in the early grades. Accordingly, focusing on a core content area—science—we examined scores from a content-independent observational measure and a second that is specific to science instruction.

TE and student motivation

Rationale for a focus on motivation

As we noted, current accountability policies specifically target student achievement growth (USDOE, 2009, 2012), an unequivocally critical outcome. Equally critical, however, are other outcomes that, despite not being tested in state assessment programs, are also consequential to student success. In particular, motivation, with its far-reaching implications for students' performance, engagement, and educational choices (Jacobs & Simpkins, 2005; Wentzel & Miele, 2016; Wigfield & Cambria, 2010), is key among these outcomes. Understanding the effects of instruction (as assessed by observational measures of TE) on student motivation is imperative, especially because some practices that boost achievement in the short term also thwart motivation for continued learning in the interim and longer term (Stipek, Feiler, Daniels, & Milburn, 1995). We argue that teaching in a way that promotes achievement while reducing motivation does not constitute an effective approach that should be sought after or rewarded.

A body of literature spanning early childhood through adolescence confirms that students' motivational beliefs are crucial for learning and achievement (Wigfield, Eccles, Schiefele, Roeser, & Davis-Kean, 2006). For young students, research framed within expectancy-value theory has documented that academic motivation comprises sets of beliefs that are differentiated into competence (or self-concept or self-efficacy perceptions) and task value (or enjoyment, liking, or interest; Wigfield & Cambria, 2010; Wigfield & Eccles, 2002; Wigfield, et al., 2006). These motivational conceptions have been consistently identified in different subjects (i.e., reading, math, and science; Chapman & Tunmer, 1995; Eccles, Wigfield, Harold, & Blumenfeld, 1993; Mantzicopoulos, Patrick, & Samarpungavan, 2008; Wigfield et al., 1997).

Subject-specific motivational beliefs develop hand-in-hand with students' learning; both are influenced significantly by teachers' practices and the opportunities they provide for students to learn specific content, including math (Schweinle, Meyer, & Turner, 2006), reading (Nolen, 2001), and science (Mantzicopoulos, Patrick, & Samarapungavan, 2013). Therefore, because what teachers do affects both students' achievement and motivation, TE scores should predict them both. In addition, we expect that TE scores should be associated with students' perceptions of the learning context: their conceptions about the support they receive for learning and the learning opportunities they are provided with (Brophy, 1988). We include two measures that reflect this set of beliefs.

Evidence on the association of TE measures with student motivation

There is some, albeit limited, correlational evidence that TE ratings predict scores on typically non-tested student outcomes, including motivation. An example is the Measures of Effective Teaching (MET) project (Kane & Staiger, 2012). Although the focus of that project was primarily on the association of TE scores with students' achievement, researchers also considered two indices of students' reported engagement: their (a) positive emotional attachment to school (two items; e.g., "This class is a happy place for me to be") and (b) effort expenditure, a behavioral indicator of motivation (three items; e.g., "My teacher pushes us to think hard about things we read," p. 49). The rationale for including these two noncognitive constructs rested on claims that "parents care about other outcomes, such as whether their children are developing a positive emotional attachment to school and are engaged in their learning" (Kane & Staiger, 2012, p. 12).

MET teachers who were rated highly on the TE observation measures tended to have students who reported "higher levels of effort and greater enjoyment in class" (Kane & Staiger, 2012, p. 11). However, teacher effects, as measured by the observation instruments, were rather small, perhaps because effort and school attachment were represented by few items, some of which were quite general.

Although we do not disagree with the MET researchers that parents care about many different outcomes that affect their children's development in school, there are additional, compelling, theoretically and empirically driven reasons for the study of motivation in the context of TE. A wealth of empirical evidence indicates that students' motivation is critical because it influences the attention they pay and the effort they apply to what is being taught: important contributors to further learning. Further, motivation influences students' use of learning strategies (e.g., monitoring progress, help seeking), their behavior at school, and their concurrent and future school- and career-related choices (Patrick, Mantzicopoulos, & Sears, 2012).

The literature is unequivocal about the pivotal role of instruction on the motivation of students who, as a result of their engagement with the content during classroom activities, develop motivational beliefs about themselves as capable learners (competence beliefs) and about the content area as worthwhile, enjoyable, and interesting (valuing beliefs; Patrick & Mantzicopoulos, 2015; Urdan & Turner, 2005). Therefore, we would expect that TE measures would capture evidence that relates to these beliefs as well as to students' content knowledge.

Along these lines, a recent study examined the links between mathematics teachers' effectiveness scores and their upper elementary students' math self-efficacy and "happiness in class" (Blazar & Kraft, 2015, p. 38). Students' math self-efficacy, a construct similar to competence beliefs (Wigfield et al., 2006), comprised 10 items (e.g., "I have been able to figure out the most difficult work in this math class"), whereas the 5-item happiness-in-class scale reflected the constructs of interest and valuing of math (e.g., "The things that I have done in math this year are interesting," "I enjoy math class this year," and "… I am learning to love math."). The results supported the assertion that the quality of mathematics instruction is related to student motivation, and the report concluded that teachers have just as much influence on students' content-specific learning as on their motivation. Adding to this literature, we focus on both achievement and motivational outcomes, as well as on children's constructions of teacher support and the provision of opportunities for learning the content, in the first year of school (i.e., kindergarten). We examine the associations between this set of outcomes with TE scores derived from measures that, as we discuss next, view instructional effectiveness from different vantage points.

Observing instruction from different vantage points: Content-independent vs. content-specific assessments

There are numerous observational measures available for states and school districts to choose among to evaluate TE (Center on Great Teachers & Leaders, 2013). Although the measures may share a number of common characteristics (e.g., many prevailing measures include scales that assess the emotional climate of the classroom as well as teachers' supportive instructional practices), they also differ in important ways. One central difference is whether the measure is designed to evaluate instruction in all

school subjects (i.e., the measure is content area– or subject-independent) or whether it is developed for a specific subject area (i.e., content area– or subject-specific).

Despite the option to include content-specific assessments, school districts across the nation routinely assess TE with content-independent, or curriculum-free, observational measures (Center on Great Teachers & Leaders, 2013). Prominent examples of these are the Classroom Assessment Scoring System (CLASS; Pianta, LaParo, & Hamre, 2008), the Framework for Teaching (FFT; Danielson, 2013), and the Marzano Teacher Evaluation Model (Marzano, Carbaugh, Rutherford, & Toth, 2014). These and other content-independent measures are based on the premise that TE involves generic skills that are fundamental to effective instruction in all subject areas (Danielson, 2007; Pianta et al., 2008); thus, they do not include specific provisions for assessing instruction within different subjects.

An important consideration that may offset the desirability of content-independent measures is that effective instruction in any subject area requires more than generic teaching skills. Discipline-specific subject-matter knowledge, along with pedagogical content knowledge, are critical and influence student learning (Shulman, 1987). The assessment of instructional practices, therefore, should be sensitive to those strategies and skills that are central to and appropriate for supporting content area learning. This is a need that content-specific assessments are designed to meet (e.g., Protocol for Language Arts Teaching Observations, Grossman, Loeb, Cohen, & Wyckoff, 2013; Mathematical Quality of Instruction, Hill, Charalambous, & Kraft, 2012; Reformed Teaching Observation Protocol [RTOP], Piburn et al., 2000; and UTeach Observation Protocol [UTOP], Walkington et al., 2012). These measures define and document effective instruction in particular curricular areas on the basis of discipline-specific standards and practices.

Justification for the pervasive use of content-independent assessments of TE requires evidence that links the observation scores directly to subject-specific student learning and motivation and compares those scores to content-specific TE measures. Very little research has done this, however. The MET project (Bill & Melinda Gates Foundation, 2014), which was conducted in grades 4 through 8 in the domains of English language arts and math, included evidence that content-specific observational measures predicted students' achievement better than the content-independent measures did. However, the results were not clear-cut (Grossman, Cohen, Ronfeldt, & Brown, 2014; Kane & Staiger, 2012). More recently, Blazar (2015), using data from fourth- and fifth-grade math classrooms, concluded that math achievement was consistently predicted by a math-specific observational assessment of TE. That study also included ratings of teachers' classroom emotional climate and organization based on a content-independent measure. However, students' math scores were not related to scores on the content-independent TE measure.

No studies have been conducted in the science domain to address differences between TE measures that vary with respect to their subject specificity (i.e., content-specific or -independent). This is a grave omission, given that science is central to the nation's economic future (National Research Council, 2011). We address this paucity of research in the present investigation. We focus specifically on kindergarten for two reasons. First, early science experiences have both immediate and long-term implications for children's science learning and motivation (Trundle & Saçkes, 2015). Children need to establish a solid grasp of the disciplinary practices, crosscutting concepts, and core ideas of science early in their education, which then serves as a foundation for understanding increasingly complex ideas (NGSS Lead States, 2013). Despite this need, science instruction in kindergarten is understudied. And, second, because the effectiveness of various TE measures appears to differ across grade levels (Mihaly & McCaffrey, 2014), findings are likely to be clearest when grades are examined individually.

Rationale for the choice of TE measures

We identified two observational TE measures appropriate for the purposes of this study: The content-independent CLASS K-3 (Pianta et al., 2008) and the science-specific RTOP (Piburn et al., 2000). Both measures purport to be suitable for evaluating instruction in the early grades of school. The CLASS K-3 is widely considered sensitive to the intended grade levels. The RTOP, like other comparable science specific measures (e.g., UTOP; Walkington et al., 2012), addresses instruction across all grade

levels. We selected it for this investigation because its constructivist orientation aligns well with recommended practices for teaching science to young children (e.g., National Association for the Education of Young Children, 2014, National Science Teachers Association, 2014).

Unlike the RTOP, the CLASS is used extensively in TE evaluations. For instance, the Office of Head Start (2015) uses evidence from the CLASS to measure the quality of Head Start programs (Hamre, 2017). In addition, some states recommend the CLASS—there are versions specific to various grade levels—for evaluating TE beyond preschool (Center on Great Teachers & Leaders, 2013; Hamre, 2017).

The CLASS, which was developed originally for research purposes, is "based on developmental theory and research suggesting that interactions between students and adults are the primary mechanism on student development and learning" (Pianta et al., 2008, p. 1); teacher–student relationships are posited to transcend the presence of materials and content of lessons. Similar to other measures (e.g., FFT; Danielson, 2013; Marzano Teacher Evaluation Model; Marzano et al., 2014), the CLASS documents the teacher's relationships with students and the classroom's general affective climate (i.e., emotional support) and classroom orderliness, formats, and routines (i.e., classroom organization). Also like other measures, the CLASS documents instruction through items that are common to all subjects (e.g., the teacher's use of complex vocabulary, questioning strategies, feedback to students), regardless of the curriculum being taught.

The CLASS views TE from a lens that contrasts sharply with the position that science-specific content and pedagogical content knowledge are critical for effective science teaching (Berry, Friedrichsen, & Loughran, 2015; Magnusson, Krajcik, & Borko, 1999). The RTOP was constructed on this premise. Therefore, it is critical that the TE assessment literature addresses how well TE scores derived from measures that view instruction from very different frames of reference predict student outcomes.

We offer three reasons why a science-specific observational measure of TE, such as the RTOP, may predict students' science achievement and motivation better than a content-independent measure does. First, the RTOP—but not content-independent measures—assesses the extent to which lessons involve fundamental concepts of the discipline. Second, the RTOP considers the accuracy and depth of teacher content-specific knowledge: facets clearly related to student achievement but not evaluated in the predominant content-independent measures. And third, teachers' pedagogical content knowledge, such as how to elicit and respond to students' lesson-specific preconceptions or how to promote conceptual understanding by guiding students to represent phenomena in a variety of ways, contribute to RTOP scores, but not those from content-independent measures.

Summary of study

In the present study, we address the relative validity of different observational measures of TE: An issue with enormous consequences for students and teachers but one that has received very little attention, particularly with regard to teaching science. We examine two types of observational measures: one that is independent of content (i.e., CLASS K-3) and one that is specific to science (i.e., RTOP). We compare how well teachers' effectiveness scores on both instruments predict students' end-of-year outcomes (i.e., science knowledge, science motivation, and perceptions of the learning context), while taking student science knowledge and motivation at the start of school into account. In addition, we control for socioeconomic status, because its effects on the children's learning and socioemotional outcomes have been well documented (e.g., Bradley & Corwyn, 2002).

Method

Participants

Participants were 10 kindergarten teachers and their 172 students with informed consent in four ethnically diverse elementary schools within the same midwestern, suburban public school district. The participating students represented 84.2% of the kindergarten population in the schools. At the end of the year, data on variables examined in the study were available on 145 students; 27 students lacked

complete data (fall and spring) as a result of enrolling in school after the baseline (early fall) testing had been completed, being assigned to a separate special education room, or having moved away after the fall testing.

The schools were comparable across achievement and sociodemographic variables. According to data provided by the state's Department of Education, all four schools had large numbers of underperforming and low-income students; the average academic performance of students in all four schools was below the state's average and more than 50% of students received free or reduced-price lunch. In the participating classrooms, free lunch data were available for 170 kindergarteners, 123 of whom received free or reduced-price lunch. All teachers were White women with 1 to 26 years of teaching experience ($M_{years} = 11.5$). The science curriculum taught in all classrooms was consistent with the state's standards for kindergarten science.

TE measures and procedure

In each classroom we video-recorded four entire science lessons (two in each of the fall and spring semesters for each teacher). The average lesson length was 36 minutes, 35 seconds. Most lessons ($n = 29$) involved life science topics (animals, plants, and their life cycles) intended to meet the state's kindergarten science standards. Eleven lessons covered measuring, mixing colors and making solutions, sinking and floating, or how objects move. As we describe later in this section, each lesson was rated at two different times, once with the CLASS and once with the RTOP.

CLASS

The CLASS K-3 (Pianta et al., 2008) is specific to the early grades and has a long history of use in national research studies (e.g., Hamre et al., 2013; Hamre, Pianta, Mashburn, & Downer, 2007). It includes 10 dimensions that are grouped into three broad domains: emotional support, classroom organization, and instructional support. Emotional support assesses the classroom's emotional climate and comprises four dimensions: positive climate (e.g., positive communication, respect), negative climate (e.g., punitive control, sarcasm), teacher sensitivity (e.g., teacher is aware of students' needs, teacher addresses problems in a proactive manner), and regard for student perspectives (e.g., teacher provides support for student autonomy, teacher is flexible and adapts instruction in response to students' ideas). Classroom organization comprises three dimensions: behavior management (e.g., student behavior, behavior expectations), productivity (e.g., use of learning time, routines), and instructional learning formats (e.g., variety of materials, clarity of objectives). Finally, instructional support reflects the teacher's efforts to involve students in the learning activities and comprises three dimensions: concept development (e.g., reasoning, connections and application to the real world), quality of feedback (e.g., scaffolding, providing information), and language modeling (e.g., advanced language, open-ended vs. closed-ended questions).

The CLASS observation procedure involves cycles of observing a lesson for 10 to 20 minutes, then stopping and rating the segment before proceeding to the next cycle. In each cycle, raters score each of the measure's 10 dimensions based on 42 behavioral markers. Scores are assigned on scale from 1 to 7 and are grouped into three broad ranges: low (1–2), middle (3–5), and high (6–7). Lesson scores from the 10 dimensions across all observation cycles are averaged and aggregated into the three classroom domains: emotional support, classroom organization, and instructional support.

The three CLASS domain scores are supported by factor analyses, reported in the measure's technical manual (Pianta et al., 2008), and are used in practice to make judgments about instructional quality. We also created a composite CLASS score (CLASS total) for two reasons: first, school districts necessarily condense data—each teacher's evaluation consists ultimately of a single effectiveness score—and, second, the CLASS composite score has been recently used in research as an index of overall classroom quality (e.g., Ponitz, Rimm-Kaufman, Grimm, & Curby, 2009; Williford, Maier, Downer, Pianta, & Howes, 2013) or of responsive teaching (Hamre, Hatfield, Pianta, & Jamil, 2014).

RTOP

The RTOP's (Piburn et al., 2000; Sawada et al., 2002) development was guided by research in science education and the National Science Education Standards (1996) and is also consistent with the Next Generation Science Standards (NGSS Lead States, 2013). This measure has some similarities with the CLASS instructional support scale. For instance, the RTOP, just like the CLASS's instructional support domain, documents the extent to which the teacher asks open-ended questions and offers opportunities for dialogic exchanges that support the articulation and sharing of ideas during instruction. The RTOP, however, is specifically focused on instructional practices that are in line with the disciplinary knowledge, skills, norms, language, and practice of science, as teachers "encourage and model the skills of scientific inquiry as well as the curiosity, openness to new ideas and data, and skepticism that characterize science" (Piburn et al., 2000, p. 5).

The RTOP comprises 25 items rated on a 5-point Likert scale (0 = *not observed* to 4 = *very descriptive*). The items are grouped into three broad scales: lesson design and implementation, content, and classroom culture. Lesson design and implementation comprises five items that assess the extent to which instruction takes into account students' prior knowledge, engages students as members of the learning community, provides opportunities for student exploration, and values a variety of approaches. The content domain includes 10 items that evaluate the extent to which the lesson promotes both propositional and procedural knowledge (5 items each) by assessing the teacher's "solid grasp" of the content, as well as the extent to which the lesson builds conceptual understanding of fundamental science concepts. Finally, within the classroom culture scale, 5 items assess communicative interactions and the teachers' support of student initiative; an additional 5 items measure the relational climate of the classroom.

Predictive validity data, based on introductory, college-level science, indicate that the RTOP is related positively ($r = .88$) to student achievement growth in science (represented by a gain score). Evidence about the construct validity of the RTOP is based on data from middle school, high school, and university science classrooms (Sawada et al., 2002). Factor analyses indicate that the three RTOP subscales are "uni-factorial" and represent "a single construct of inquiry" (Sawada et al., 2002, p. 24). Thus, researchers using this measure have computed a total score, derived from summing scores across the 25 items. Because the measure has not been used in kindergarten, we were interested in examining the potentially unique contributions of each of the RTOP scales to achievement and motivation. Therefore, in addition to calculating the total score, we created scores for each of the three scales by averaging the ratings across the respective items.

Rating lessons

Lessons were scored with both the CLASS and RTOP on different occasions and by different raters. There were four CLASS raters and four RTOP raters; one rater used both measures. Raters did not score the same lesson with both instruments. Lessons were not rated chronologically but were counterbalanced so that approximately half the lessons were rated first on the CLASS and the other half were scored first with the RTOP. For each measure, scoring was balanced across raters so that each teacher was scored by all raters.

Prior to beginning scoring the lessons, project members became proficient in using the instruments. We conducted extensive reliability training using kindergarten science lessons that were not part of the current study but were of comparable length and lesson format. We then established interrater agreement between every pair of raters. We report these details next.

CLASS reliability

Four raters were certified to use CLASS. Certification involved (a) attending a 2-day reliability training provided by a certified CLASS trainer and (b) passing an online certification test that requires examinees to view and score five different videos and achieve at least 80% agreement with the scores of CLASS master coders. Once certified, raters viewed and scored 5 video-recorded kindergarten science lessons individually and compared scores with the other raters. Interrater agreement between all rater pairs ranged from 81.0% to 88.8% ($M = 84.3\%$).

RTOP reliability

There is no external certification process for the RTOP; therefore, we developed a training process similar to that for the CLASS. After extensively reviewing the observation protocol as a team, the four RTOP raters engaged in cycles of watching and scoring video-recorded lessons individually and then met to discuss scoring, resolve discrepancies, and develop consensus. After this training period, each rater watched and scored five new lessons independently. Interrater agreement ranged from 80% to 100% ($M = 92.8\%$).

Child outcome measures and procedure

All child measures were specific to learning science. They were administered by project members, individually, in a quiet location during regular school hours. All outcome measures were administered in the spring. We also measured students' science achievement and self-perceived science competence during the first month of school; those scores, along with socioeconomic status (free or reduced-cost lunch status), were used as controls in all analyses.

Science achievement. We measured children's knowledge of science with the science knowledge subscale of the Woodcock-Johnson Tests of Achievement III (WJ-III; Woodcock, McGrew, & Mather, 2001). A standardized assessment, the WJ-III is designed to assess general knowledge in biological and physical sciences. Items appropriate for young children reflect science-related vocabulary and general knowledge and prompt for children's knowledge of different body parts (e.g., eye, nose, knee), animals (e.g., dog, bird, frog, kangaroo), and processes (e.g., pollution, hibernation). Psychometric information for the full academic knowledge cluster subtest (it comprises not only science knowledge but also social studies and humanities) includes 1-year test-retest reliability for 2- to 7-year-olds (.84) and split-half reliability for 4- to 6-year-olds (.92). Correlations of the WJ-III with other achievement measures are reported in the test's technical manual (McGrew & Woodcock, 2001) as evidence of the test's validity.

Perceived science competence and *liking science* were measured with subscales of the Puppet Interview Scales of Competence in and Enjoyment of Science (Mantzicopoulos et al., 2008). The perceived science competence subscale ($\alpha = .85$) had 19 items, such as "I know a lot about science." The liking science subscale had 7 items ($\alpha = .76$); an example is "I have fun learning science." The administration procedure used for this assessment involves two puppets, one of which makes a positive statement (e.g., "I know a lot about living things") and a second that follows with a corresponding negative statement ("I don't know a lot about living things yet"). The interviewer then asks the child to tell which of the two puppets is more like him or her (positive statements receive a score of 1, whereas negative statements receive a score of 0). The presentation of the questions is counterbalanced so that puppets alternate making positive and negative statements. Both scales (perceived science competence and liking science) have small but significant associations with measures of achievement.

Opportunities for learning science was measured with the Children's Perceptions of What I Learn in Kindergarten (Mantzicopoulos et al., 2013), which assesses children's perceptions of whether they learn about science topics in kindergarten ($\alpha = .91$; 13 items). Items are scored dichotomously (1 = *yes*; 0 = *no*). An example is "In school we learn about how living things grow."

Perceived teacher support for learning science (Mantzicopoulos et al., 2013) was assessed with 5 items ($\alpha = .82$) that are scored on a 3-point Likert format (1 = *no*, 2 = *sometimes*, 3 = *a lot of the time*). Item examples are "My teacher helps me understand things about science" and "My teacher tells me that I can be a scientist."

Data analysis procedures

Consistent with the approach followed in the MET study (Kane & Staiger, 2012), we first created separate CLASS and RTOP composite scores for each teacher by averaging her scores across the four lessons. This was necessary because single-occasion observation scores are unreliable and produce

imprecise estimates of TE (Kane & Staiger, 2012; Whitehurst et al., 2014). Next, we conducted preliminary analyses to provide descriptive data on the measures and examine their intercorrelations. We then used hierarchical linear modeling procedures (HLM; Raudenbush & Bryk, 2002) to estimate fully unconditional models for each measure and calculated the measures' intraclass correlations (indicating the variance between classrooms).

The HLM analyses examined whether the CLASS and RTOP scores differentially predicted students' end-of-year science knowledge and motivation. Because the students were nested within classrooms, we built and ran a series of identical two-level models, using restricted maximum likelihood as the method of estimation. At the first level we included students' science knowledge and perceived science competence at the beginning of the year, as well as their free lunch status (socioeconomic status) in all analyses, to control for the effects of these variables on achievement, motivation, and perceptions of teacher support and opportunities for learning science. Teachers' CLASS or RTOP scores, averaged across the four lessons, were included at the second, classroom, level. We built final models by testing one variable at a time, checking the reliability estimates of variance components and fixing coefficients as needed. All coefficients reported are standardized.

Results

Preliminary analysis

TE measures

Descriptive statistics and correlations for the mean ratings of the 10 participating teachers across the CLASS and RTOP scores are shown in Table 1. Average scores on the CLASS instructional support ($M = 3.40$) domain were lower than on either the emotional support ($M = 5.20$) or the classroom organization ($M = 5.57$) domains. The trend for instructional support ratings to be in the lower mid- to low range of effectiveness is a consistent finding, reported both in the CLASS K-3 manual (Pianta et al., 2008) as well as in other studies with the CLASS K-3 (e.g., Curby, Grimm, & Pianta, 2010; Plank & Condliffe, 2013) and with different versions of the CLASS (Kane & Staiger, 2012; Office of Head Start, 2015).

Recognizing that with a small sample, correlations in the small or moderate range may be less stable (e.g., Schönbrodt & Perugini, 2013), we examined the bivariate associations among all CLASS domains, all RTOP domains, and the domains of two measures. As shown in Table 1, emotional support was related significantly to instructional support ($r = .74$). Classroom organization, however, was not correlated with instructional support ($r = .12$) and was moderately, yet not significantly, correlated with emotional support ($r = .54$). The CLASS total score (i.e., the average of the three domain scores) correlated positively with emotional support ($r = .95$), classroom organization ($r = .62$), and instructional support ($r = .83$); all three correlation estimates were statistically significant.

Table 1. Descriptive statistics and correlations among CLASS and RTOP scores ($N = 10$).

TE domain	1	2	3	4	5	6	7	8
1. CLASS: Emotional support								
2. CLASS: Classroom organization	0.54							
3. CLASS: Instructional support	0.74**	0.12						
4. CLASS total	0.95***	0.62**	0.83**					
5. RTOP: Lesson design and implementation	0.66*	−0.13	0.83**	0.63*				
6. RTOP: Content	0.67*	0.00	0.79**	0.65*	0.96**			
7. RTOP: Classroom culture	0.72*	0.01	0.80**	0.68*	0.97**	0.99**		
8. RTOP total	0.69*	−0.05	0.82**	0.66*	0.99**	0.99**	1.00**	
M	5.20	5.57	3.40	4.73	1.84	2.07	2.02	1.98
SD	0.55	0.44	0.62	0.43	0.63	0.46	0.59	0.55
Minimum score	4.44	4.86	2.40	4.15	0.95	1.38	1.17	1.17
Maximum score	5.78	6.29	4.51	5.39	2.65	2.65	2.83	2.71

Note. $^{*}p < .05$. $^{**}p < 0.01$.

The correlations between the three RTOP scales were uniformly high (i.e., > .96). Even with a small sample of 10 teachers, these intercorrelations are highly stable (i.e., the confidence interval for $r = .96$ ranges from .84 to .99). These estimates suggest that, just as when used with teachers of older students (e.g., Sawada et al., 2002), the measure also has a unidimensional structure in the early grades. Therefore, we did not maintain separate RTOP domain scores for analyses but used only a composite RTOP score, computed by averaging lesson scores across the three RTOP domains. The mean RTOP total score ($M = 1.98$) for our sample of teachers placed them in the mid-level category, suggesting that, on average, the instructional strategies assessed by this measure were observed for some of the time. For comparison purposes, we note that mean scores on the RTOP total, reported for samples of middle and high school students (e.g., Ms = 2.0 and 1.67, respectively), were also at or below the mid-level category (Sawada et al., 2007).

We next examined correlations between ratings of teachers' practices on the RTOP and the CLASS. Scores on the RTOP total correlated moderately but significantly with CLASS's emotional support ($r = .69$) and the CLASS total ($r = .66$). In addition, the RTOP total was correlated strongly with CLASS's instructional support ($r = .82$). The RTOP total did not correlate with CLASS classroom organization scores ($r = -.05$).

Student outcome measures

Descriptive statistics and correlations for the measures of students' knowledge, motivation, and perceptions of the learning context are shown in Table 2. Students' science knowledge score at the end of the year was related significantly to their scores at the beginning of the year ($r = .67$) as well as to socioeconomic status ($r = .36$), liking science ($r = .24$), and perceived opportunities for learning science ($r = .18$). The end-of-year science outcome measures were also correlated moderately and significantly with each other (estimates ranged from $r = .45$ to $r = .73$). There was a small but significant negative association between socioeconomic status and end-of-year perceived science competence ($r = -.19$).

Between-classroom variance

Fully unconditional models for the achievement, motivation, and perceived learning context outcomes indicated substantial variance both within classrooms (level 1) and between classrooms (level 2), thus warranting multilevel analyses. The intraclass correlation coefficients for the outcome measures were 6% for science knowledge, 41% for perceived science competence, 23% for science liking, 54% for perceived opportunities to learn science, and 38% for teacher support for learning science.

Associations of CLASS domain scores with children's outcomes

Our findings on the associations of classroom emotional support, classroom organization, and instructional support with children's end-of-year science knowledge and motivation—while controlling for children's beginning science knowledge, perceived competence, and socioeconomic status—are based on two separate HLM analyses. In the first analysis, we entered emotional support and classroom

Table 2. Descriptive statistics and correlations among student measures.

Variable	*M*	SD	*1*	*2*	*3*	*4*	*5*	*6*	*7*
1. Science knowledge: Beginning of year	11.02	2.17						.	
2. Perceived science competence: Beginning of year[1]	0.53	0.26	−0.08						
3. Science knowledge: End of year	12.74	1.71	0.67**	−0.13			.		
4. Perceived science competence: End of year[1]	0.62	0.27	−0.09	0.16	0.14				
5. Liking science: End of year[1]	0.76	0.27	0.02	0.04	0.24**	0.64**			
6. Opportunities for learning science: End of year[1]	0.75	0.29	−0.10	−0.01	0.18*	0.66**	0.45**		
7. Teacher support for learning science: End of year[2]	2.35	0.70	−0.13	−0.01	0.11	0.61**	0.47**	0.73**	
8. Socioeconomic status[3]	0.29	0.46	0.41**	−0.10	0.36**	−0.19*	−0.01	−0.10	−0.08

Note. [1]Scored 0–1; [2]scored 0–2; [3]0 = receiving free or reduced-cost lunch, 1 = self-pay.
*$p < .05$. **$p < .01$.

Table 3a. Associations of CLASS emotional support and classroom organization domain scores with children's end-of-year science knowledge, motivation, and perceptions of the context for learning science.

	End-of-year outcomes									
	Science knowledge		Perceived science competence		Liking science		Opportunities for learning science		Teacher Support for Learning Science	
Fixed effects	*β*	*SE*	*β*	*SE*	*β*	*SE*	*β*	*SE*	*β*	*SE*
CLASS emotional support	0.06	0.07	0.38*	0.16	0.43*	0.15	0.45*	0.19	0.41*	0.17
CLASS classroom organization	−0.15**	0.05	−0.58**	0.14	−0.30*	0.13	−0.59**	0.19	−0.44*	0.17
Beginning of year:										
Science knowledge	0.62***	0.04	−0.01	0.06	0.01	0.08	−0.03	0.07	−0.03	0.07
Perceived science competence	−0.03	0.06	0.22**	0.07	0.08	0.08	0.06	0.04	0.01	0.10
Socioeconomic status	0.22	0.18	−0.19	0.11	0.10	0.15	0.01	0.09	0.05	0.15

Note. Socioeconomic status is coded 0 = receiving free or reduced-cost lunch, 1 = self-pay.
*$p < .05$. **$p < .01$. ***$p < .001$.

organization in the model, whereas in the second analysis we entered classroom organization and instructional support. It was necessary to evaluate these two separate models because the high correlation between emotional support and instructional support masked the contributions of each when used together in a single model.

Prior to running this final set of analyses, however, we examined the evidence resulting from entering all three CLASS domains (emotional support, classroom organization, and instructional support) in a single model. In this model, (a) neither emotional support nor instructional support were consistently associated with children's motivation; (b) emotional support had a statistically significant, negative association with end-of-year science knowledge, an artifact of the high correlation of this variable with instructional support; and (c) classroom organization had significant negative associations with the outcomes of interest. Of note, these negative associations were maintained when classroom organization was entered alone in the model (after taking into account the effects of the three control variables).

To disambiguate the contributions of emotional support and instructional support to children's end-of-year outcomes, we opted against creating a composite score for these two variables. Instead, we examined their unique contributions in the context of classroom organization, as shown in Tables 3a and 3b.

Teachers' emotional support score consistently predicted children's end-of-year motivation, as well as their perceptions of teacher support and opportunities for learning science (Table 3a). That is, teachers with higher ratings on emotionally supportive strategies tended to have students who, at the end of the year, reported significantly higher levels of perceived science competence ($\beta = 0.38$), science liking

Table 3b. Associations of CLASS instructional support and classroom organization domain scores with children's end-of-year science knowledge, motivation, and perceptions of the context for learning science.

	End-of-year outcomes									
	Science knowledge		Perceived science competence		Liking science		Opportunities for learning science		Teacher Support for Learning Science	
Fixed effects	*β*	*SE*	*β*	*SE*	*β*	*SE*	*β*	*SE*	*β*	*SE*
CLASS instructional support	0.16**	0.06	0.28*	0.12	0.34**	0.09	0.24*	0.10	0.35**	0.10
CLASS classroom organization	−0.10	0.07	−0.35*	0.13	−0.03	0.12	−0.37*	0.15	−0.34*	0.16
Beginning of year:										
Science knowledge	0.63***	0.06	0.02	0.06	0.01	0.08	−0.04	0.07	−0.06	0.05
Perceived science competence	−0.04	0.06	0.22*	0.07	0.08	0.08	0.07*	0.04	0.03	0.09
Socioeconomic status	0.21	0.19	−0.22*	0.12	0.12	0.15	0.02	0.08	−0.01	0.10

Note. Socioeconomic status is coded 0 = receiving free or reduced-cost lunch, 1 = self-pay.
*$p < .05$. **$p < .01$. ***$p < .001$.

($\beta = 0.43$), more opportunities for learning science in their classrooms ($\beta = 0.45$), and greater levels of teacher support for learning science ($\beta = 0.41$). Contrary to expectation, emotional support was not related significantly to children's end-of-year science knowledge ($\beta = 0.06$). Also contrary to our expectations, classroom organization was associated negatively with all measures. Specifically, as teachers' classroom organization score increased, their students tended to score significantly lower on science knowledge ($\beta = -0.15$) and report lower levels of science competence ($\beta = -0.58$) and liking science ($\beta = -0.30$), fewer opportunities for learning science ($\beta = -0.59$), and lower levels of teacher support for learning science ($\beta = -0.44$).

The findings for classroom organization were replicated in the analyses examining this CLASS domain in the context of teachers' instructional support strategies (Table 3b). In this model, however, instructional support was positively and significantly related to students' science knowledge ($\beta = 0.16$) and motivation (βs = 0.28 and 0.34 for perceived competence and liking, respectively), as well as to perceived opportunities to learn science ($\beta = 0.24$) and teacher support for learning science ($\beta = 0.35$).

As shown in Tables 3a and 3b, children's end-of-year science knowledge was significantly predicted by their beginning science knowledge (βs = 0.62 and 0.63, respectively). In addition, beginning-of the-year perceived science competence significantly predicted end-of-year perceived science competence ($\beta s = 0.22$). These effects were in the expected direction; children with higher scores on science knowledge at the start of kindergarten and more positive science-related competence beliefs scored higher at the end of the year on each measure. Finally, in the analysis reported in Table 3b, socioeconomic status was negatively associated with end-of-year perceived science competence ($\beta = -0.22$). This suggests that children who were more economically advantaged tended to report lower levels of science competence.

Associations of CLASS total score with children's outcomes

Table 4 displays the results of the HLM analyses examining the associations between the CLASS total score (average of the three domains) and students' end-of-year outcomes, controlling for children's socioeconomic status, prior science knowledge, and early science competence beliefs. We identified one statistically significant association between the CLASS total rating and science liking ($\beta = 0.33$).

As in the previous set of analyses, there were significant relations between beginning-of-the-year and end-of-year science knowledge ($\beta = 0.63$); beginning-of-the-year and end-of-year perceived science competence ($\beta = 0.22$); and socioeconomic status and perceived science competence ($\beta = -0.22$). The first two effects were in the expected direction; regardless of their teacher's rating on the CLASS total score, children who started kindergarten with more science knowledge and more positive science-related competence beliefs performed significantly better at the end of the year on each measure. The effects of socioeconomic status indicated that, regardless of their teacher's overall CLASS

Table 4. Associations of the CLASS total score with children's end-of-year science knowledge, motivation, and perceptions of the context for learning science.

	End-of-year outcomes									
	Science knowledge		Perceived science competence		Liking science		Opportunities for learning science		Teacher Support for Learning Science	
Fixed effects	β	*SE*	β	*SE*	β	*SE*	β	*SE*	β	*SE*
CLASS total score	0.03	0.08	0.10	0.20	0.33**	0.10	0.01	0.20	0.12	0.18
Beginning of year:										
Science knowledge	0.63***	0.04	0.03	0.05	0.05	0.09	−0.04	0.07	−0.01	0.08
Perceived science competence	−0.03	0.05	0.22*	0.07	0.08	0.07	0.08*	0.04	0.00	0.09
Socioeconomic status	0.22	0.19	−0.22*	0.13	0.09	0.17	0.03	0.08	0.06	0.17

Note. Socioeconomic status is coded 0 = receiving free or reduced-cost lunch, 1 = self-pay.
*$p < .05$, **$p < .01$, ***$p < .001$.

Table 5. Associations of the RTOP total score with children's end-of-year science knowledge, motivation, and perceptions of the context for learning science.

	End-of-year outcomes									
	Science knowledge		Perceived science competence		Liking science		Opportunities for learning science		Teacher Support for Learning Science	
Fixed effects	*β*	*SE*	*β*	*SE*	*β*	*SE*	*β*	*SE*	*β*	*SE*
RTOP total score	0.14*	0.06	0.24	0.17	0.28*	0.12	0.21	0.15	0.30*	0.15
Beginning of year:										
Science knowledge	0.65***	0.04	0.03	0.06	0.02	0.08	−0.04	0.07	−0.07	0.04
Perceived science competence	−0.04	0.05	0.22*	0.07	0.10	0.08	0.07*	0.04	0.04	0.10
Socioeconomic status	0.22	0.18	−0.22*	0.13	0.15	0.15	0.03	0.08	0.01	0.10

Note. Socioeconomic status is coded 0 = receiving free or reduced-cost lunch, 1 = self-pay.
*$p < .05$. **$p < .01$. ***$p < .001$.

rating, children who were more economically advantaged reported significantly lower levels of perceived science competence.

Associations of RTOP total score with children's outcomes

The results of analyses examining associations between the RTOP total score and students' end-of-year outcomes are shown in Table 5. After controlling for students' socioeconomic status and beginning-of-kindergarten measures, the RTOP total significantly predicted students' science knowledge ($\beta = 0.14$), science liking ($\beta = 0.28$), and their reports of teacher support for learning science ($\beta = 0.30$). The associations of the RTOP total score with students' perceived opportunities for learning science and perceptions of science competence did not reach statistical significance, despite being in the expected direction.

As with the previous sets of analyses, children's end-of-year science knowledge was predicted by their science knowledge at the start of the year ($\beta = 0.65$), whereas end-of-year science competence beliefs were predicted by science competence beliefs at the start of the year ($\beta = 0.22$) and socioeconomic status ($\beta = -0.22$).

Discussion

Our study contributes to the literature on the evaluation of teachers' instructional practices in at least two ways. First, we extended the focus of effective teaching beyond solely student achievement by considering students' motivation and perceptions of the learning context as additional outcomes of effective instruction. Second, we included both a content-independent measure of TE—the norm in practice—and a measure specific to science instruction. We examined how well each predicted student outcomes in science specifically, and this work represents the first study, to our knowledge, to address this question. In particular, we evaluated associations between the observational TE measures and students' achievement and motivation in the early elementary years, complementing studies that focused on middle and high school grades (e.g., MET project, Kane & Staiger, 2012).

Our findings indicate that TE scores derived from observation measures that view instruction from different lenses provide differential information about the relationship between domains of teaching practices and student outcomes. Teachers' instructionally supportive practices, as reflected in the CLASS instructional support and the RTOP ratings, were positively associated with students' science knowledge, motivation, and perceptions of the learning context. Even though CLASS's emotional support domain was positively related to students' motivation and their perceived learning environment, there was no evidence that the practices measured in this domain contribute to young students' learning, at least in science.

Contrary to our expectation, classroom organization was negatively related to student outcomes, regardless of whether it was examined concurrently with emotional or instructional support. This result suggests that highly structured classroom environments, as reflected in the CLASS, may not support students' engagement and interest in science, at least in kindergarten.

Extending the conceptualization of TE to include student motivation

In the introduction to the current study, we argued that promoting achievement without also fostering motivation is insufficient for teachers to be considered effective, because students' content-specific motivation is influenced by their early experiences with particular subjects. Specifically, practices that promote early, appropriate, and sustained engagement with science nurture children's positive motivational beliefs, including their interest and enjoyment in science, perceptions of competence, and beliefs that learning science is relevant for them (Mantzicopoulos et al., 2013). Science instruction that is engaging and motivating is particularly critical in the early grades because it represents a path to addressing documented achievement gaps (Quinn & Cooc, 2015). Therefore, the extent to which instruction is effective should be gauged from children's motivation and engagement with the discipline, not only from their knowledge.

The end-of-year science motivation of the kindergarteners in our study was not related to their initial knowledge of science. By the year's end, however, there was significant variability—between approximately one-quarter and one-third of overall variability—in children's motivation among classrooms, which suggests that differences in teachers' practices were related to differences in motivation. Specifically, there was evidence that subject-specific motivation is enhanced when teachers create an inviting and positive classroom climate, a finding that parallels evidence in a recent study of fourth- and fifth-grade mathematics instruction in 310 classrooms (Blazar & Kraft, 2015).

At the same time, the positive effects of classroom climate—which include caring relationships, respectful communication, and the teacher's sensitivity to student perspectives—are not sufficient for children's learning. This finding also parallels evidence reported when CLASS's emotional support score was used to predict fourth and fifth graders' math achievement (Blazar, 2015). Together, these findings uphold the premise that students' science knowledge—as well as their motivation and beliefs about science-specific opportunities and supports in their classroom—are dependent on teachers' instructional actions. These actions target students' conceptual development by extending their thinking and scaffold the development of advanced, discipline-specific language through engaged discussion about the content. This set of practices is reflected to some extent in the CLASS instructional support domain and, to a greater extent, in the full RTOP measure. Given that relations between achievement and motivation are reciprocal and accumulate to develop increasingly stable trajectories (Marsh, Trautwein, Lüdtke, Küller, & Baumert, 2005), it is reasonable to expect that, over time, these between-classroom differences in science motivation would manifest as classroom differences in achievement.

Despite the RTOP's science-specific focus, however, and even though ratings on this measure predicted students' science knowledge, its scores did not consistently predict all motivation and perceived learning context outcomes. The RTOP coefficients across all outcomes were in the expected direction, but not all reached statistical significance, due to relatively large standard errors, compared to the CLASS instructional support estimates. Of note, even though the developers of the RTOP claim that it is appropriate for the early grades of school, the measure had not previously been used in kindergarten. Because the measure is used to evaluate teachers' strategies following professional development (e.g., Lakshmanan, Heath, Perlmutter, & Elder, 2010), our findings highlight the need to demonstrate the links between teachers' RTOP scores and students' outcomes. We venture that measures of effective science instruction merit further attention for at least two reasons. First, high-quality science education is "a cultural imperative, essential to the nation's future" (National Research Council, 2011, p. 26). And second, instruction aimed at closing science gaps along gender, race, and socioeconomic lines should begin while students are young, because these gaps appear early—as early as third grade according to some (Quinn & Cooc, 2015). The chances that these objectives will be met are arguably greatest if

quality science instruction is established from the outset of children's schooling, a situation that can only be assured when science-specific outcomes are considered.

Contrary to our expectation, the CLASS classroom organization score, whether examined in the context of emotional or instructional support, was related negatively to all outcomes—perceived science competence, liking science, and both perceived opportunities and support for learning science. This consistent pattern of results seems to run counter to the accepted premise that well-organized classrooms promote learning and engagement. However, other researchers have reported similar results on the association between the CLASS's classroom organization subscale and student engagement (e.g., Ponitz et al., 2009). Blazar and Kraft (2015), in their study of mathematics teaching practices and their effects on students' achievement and motivation, reported that classroom organization was positively associated with students' behavior but negatively associated with student reports of their liking and interest in math. We thus reason that our finding highlights an important issue that warrants further investigation. Perhaps practices that are scored highly on CLASS's classroom organization domain are perceived by children as highly controlling and autonomy-thwarting: perceptions that undermine student motivation (Reeve, 2002; Soenens, Sierens, Vansteenkiste, Dochy, & Goossens, 2012). High levels of classroom order and behavioral control may constrain inquiry-based and student-centered practices that are central to desirable instruction (i.e., an emphasis on student initiative rather than teacher control or a focus on the articulation and construction of knowledge vs. a reliance on providing the correct answer).

Although classroom chaos is clearly problematic and not at all conducive to either learning or motivation, perhaps once classroom organization surpasses a particular level it then begins to discourage children's curiosity and propensity to learn. In fact, research that examined reading outcomes in first-grade classrooms provides some support for this assertion; students' reading scores were lower in classrooms where organization remained high, or increased, over time (Cameron, Connor, Morrison, & Jewkes, 2008). Furthermore, in related work with the CLASS, researchers concluded that the effects of classroom organization may be moderated by student skill and content area. For mathematics, high-achieving students "may actually function well in poorly managed classes," as they are able to work independently without explicit support from the teacher (Curby, Rimm-Kaufman & Ponitz, 2009, p. 921). Taken together, the evidence calls attention to the constructs reflected in the CLASS classroom organization domain and their effects on student outcomes.

Comparing a content-independent and a science-specific observational measure

Despite clear consensus that teaching demands both disciplinary and pedagogical content knowledge, in addition to general instructional skills, the most widely used TE observational measures are designed to evaluate instruction irrespective of the content (e.g., Danielson, 2007; Pianta et al., 2008). The little research that has considered specific content areas in the elementary school grades has involved only English/language arts and mathematics (Kane, Kerr, & Pianta, 2014). Confidence in the evaluation of teachers' quality of instruction in other core content areas, such as science, is likely greatest when their teaching is evaluated relative to discipline-relevant outcomes specifically, rather than inferred from generic metrics that are associated with reading or math scores.

Our findings in the domain of science suggest that not all TE domains are sensitive to the knowledge and practices that are contributors to positive science-related outcomes (Bartos & Lederman, 2014; Yore et al., 2008). Although emotionally supportive environments create a context for positive student perceptions of their classroom and their own competence, it is the instructionally supportive practices that contribute to students' subject-specific knowledge. Therefore, reducing TE into an overall composite score is not likely to provide clear information about those practices that are critical contributors to student success. This is illustrated in the present study, when the scores from the three CLASS domains were averaged into a total score that predicted only one of the four motivation-related measures—children's liking of science—and did not predict student science knowledge. This limited sensitivity of the total score is noteworthy, given that school administrators' assessments must result in a single, overall evaluation score for each teacher.

Even though a logical approach to managing the three CLASS subscales may be to create a composite, our findings contraindicate any reliance on the CLASS total score because it masks differences across teaching practices that have important achievement and motivational implications for children. This evidence is consistent with the results of the RTOP, a measure that focuses entirely on science-specific instruction. However, as we noted in the preceding section, scales that, like the RTOP, claim to be appropriate across all grade levels may need extensive work in order to increase their sensitivity to practices that enhance learning and motivation outcomes for young children. The CLASS purports to have this advantage, as it is based on extensive work in pre–K to third grade. However, this measure's instructional support domain may also profit from the inclusion of content-specific pedagogical strategies beyond the generically appropriate practices that it assesses. This is much needed to strengthen the measure's ability to predict student success across a variety attributes and skills and to support its use in teacher education and, perhaps eventually, in high-stakes decision making.

In addition to accountability, a major goal of evaluating TE is to shape teachers' behaviors and persuade them, through incentives, to use particular instructional practices and not others (USDOE, 2012). Thus, the practices addressed in any observational measure are likely to become the most valued and the focus of professional development efforts. Practices that are encouraged based on generic, content-independent measures will only support students' learning and motivation if those particular measure(s) predict important outcomes across the array of content areas. However, if generic instructional practices are central to optimally teaching some content areas but not others, less optimal or even undermining practices may be inadvertently encouraged.

Limitations and directions for future research

The generalizability of this study is limited by the small sample size and the focus on only one grade level: kindergarten. The findings, however, are encouraging and warrant further research with larger groups of classrooms and across the early grades of school. Attention to effects by grade level is needed because effective instruction in different content areas may vary at different grade levels. At least for content-independent measures, observation-based scores are related to student achievement differently, depending on grade level (Mihaly & McCaffrey, 2014). Therefore, aggregating across grades may lead to masking grade-specific effects. Accordingly, further research is needed that considers science instruction and instruction in various content areas at different, specific grades.

The fact that the RTOP was developed for use in both science and math classes, rather than only science classes, may be considered a limitation. However, we were not able to locate any observational measure specific only to science instruction. Moreover, there is a paucity of evidence about the reliability and validity of the RTOP below the middle school grades. Given that effective instruction is important for science education and that observational measures are becoming ubiquitous for measuring effective instruction, it is crucial to know that the measures used to evaluate science teaching do actually produce the desired science outcomes.

In addition, even though in the current study scores from the science knowledge subscale of the WJ-III were associated with teachers' instructional strategies, the resulting coefficients were small, albeit statistically significant. To interpret the findings it is important to consider that this scale was not constructed with attention to key science themes. Moreover, its items draw on children's general science knowledge and vocabulary skills without probing for conceptual understanding. Measures of children's knowledge across specific areas that are addressed in the science curriculum would provide greater insight into the links between instructional practices and student learning.

Nonetheless, standardized assessments that, like the one used in this study, are not proximal to the instructional context are typically used for accountability purposes to measure the broad impact of science curricula (Hickey, Taasoobshirazi, & Cross, 2012). However, if a goal of accountability is to improve instruction, it is important to have access to measures of children's science learning that reflect

the curriculum and are consistent with grade-specific standards, particularly in the early grades of school.

Moving beyond science education specifically, there is an urgent need to examine the validity of all measures that are used to evaluate teachers' effectiveness. This includes investigating whether measures' scores are sensitive to instruction that leads to achievement and motivation in an array of content areas and across all grade levels. At present, high-stakes outcomes are tied to teachers' effectiveness scores in the absence of grade level– and content-specific empirical support.

References

Bartos, S., & Lederman, N. G. (2014). Teachers' knowledge structures for nature of science and scientific inquiry: Conceptions and classroom practice. *Journal of Research in Science Teaching*, *51*, 1150–1184.

Berry, A., Friedrichsen, P., & Loughran, J. (Eds.) (2015). *Re-examining pedagogical content knowledge in science education*. New York, NY: Routledge.

Bill & Melinda Gates Foundation. (2013). *Measures of effective teaching*. Retrieved from http://www.metproject.org/

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

Blazar, D., & Kraft, M. A. (2015). *Teacher and teaching effects on students' academic behaviors and mindsets* (Working Paper 41). Retrieved from https://www.mathematica-mpr.com/our-publications-and-findings/publications/teacher-and-teaching-effects-on-students-academic-behaviors-and-mindsets

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, *53*, 371–379.

Brophy, J. (1988). Research linking teacher behavior to student achievement: Potential implications for instruction of Chapter 1 students. *Educational Psychologist*, *23*, 235–286.

Cameron, C. E., Connor, C. M., Morrison, F. J., & Jewkes, A. M. (2008). Effects of classroom organization on letter-word reading in first grade. *Journal of School Psychology*, *46*, 173–192.

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

Chapman, J. W., & Tunmer, W. E. (1995). Development of young children's reading self-concepts: An examination of emerging subcomponents and their relationship with reading achievement. *Journal of Educational Psychology*, *87*, 154–167

Curby, T. W., Grimm, K. J., & Pianta, R. C. (2010). Stability and change in early childhood classroom interactions during the first two hours of a day. *Early Childhood Research Quarterly, 25,* 373–384.

Curby, T. W., Rimm-Kaufman, S. E. & Ponitz, C. C. (2009). Teacher-child interactions and children's achievement trajectories across kindergarten and first grade. *Journal of Educational Psychology*, *101*, 912–925.

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 ed.). Retrieved from https://www.danielsongroup.org/framework/

Eccles, J., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development*, *64*, 830–847.

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.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationships between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher*, *53*, 293–303.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: the relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education*, *119*(3), 445–470.

Hamre, B. K. (2017). *Using classroom observation to gauge teacher effectiveness: Classroom Assessment Scoring System (CLASS).* Retrieved from http://cepr.harvard.edu/files/cepr/files/ncte-conference-class-hamre.pdf

Hamre, B. K., Hatfield, B., Pianta, R., & Jamil, F. (2014). Evidence for general and domain-specific elements of teacher-child interactions: Associations with preschool children's development. *Child Development*, *85*, 1257–1274.

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., & Hamagami, A. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *The Elementary School Journal*, *113*, 461–487.

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007). *Building a science of classrooms: Application of the CLASS framework in over 4,000 U.S. early childhood and elementary classrooms*. New York: Foundation for Child Development. Retrieved from http://fcd-us.org/sites/default/files/BuildingAScienceOfClassroomsPiantaHamre.pdf

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record*, *116*, 1–28.

Hickey, D. T., Taasoobshirazi, G., & Cross, D. (2012). Assessment as learning: Enhancing discourse, understanding, and achievement in innovative science curricula. *Journal of Research in Science Teaching*, *49*, 1240–1270.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64.

Jacobs, J. E., & Simpkins, S. D. (2005). Mapping leaks in the math, science, and technology pipeline. *New Directions for Child and Adolescent Development*, *110*, 3–6.

Kane, T. J., Kerr, K. A., & Pianta, R. C. (2014). *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project*. San Francisco, CA: Jossey-Bass.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Kimball, S. M. & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly*, *45*, 34–70.

Lakshmanan, A., Heath, B. P., Perlmutter, A., & Elder, M. (2010). The impact of science content and professional learning communities on science teaching efficacy and standards-based instruction. *Journal of Research in Science Teaching*, *48*, 534–551.

Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, sources, and development of pedagogical content knowledge for science teaching. In J. Gess-Newsome & N. G. Lederman (Eds.), *Examining pedagogical content knowledge: The construct and its implications for and science education* (pp. 95–132). Norwell, MA: Kluwer.

Mantzicopoulos, P., Patrick, H., & Samarapungavan, A. (2008). Young children's motivational beliefs about learning science. *Early Childhood Research Quarterly*, *23*, 378–394.

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.

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, *76*, 397–416.

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 http://www.marzanocenter.com/Teacher-Evaluation-2014-Model.pdf

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

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.

National Association for the Education of Young Children. (2014). *Issues in early childhood education: Science. Position statement of the National Science Teachers Association, endorsed by the NAEYC Governing Board April 2014*. Retrieved from http://www.naeyc.org/positionstatements

National Research Council. (2011). *Report to congress*. Retrieved from http://www.nationalacademies.org/annualreport/Report_to_Congress_2011.pdf

National Science Teachers Association. (2014). *NSTA position statement: Early childhood science education*. Retrieved from http://www.naeyc.org/files/naeyc/Early%20Childhood%20FINAL%20FINAL%201-30-14%20%281%29.pdf

NGSS Lead States. (2013). *Next generation science standards: For states, by states. Volume 1*. Washington, DC: National Academies Press.

Nolen, S. B. (2001). Constructing literacy in the kindergarten: Task structure, collaboration, and motivation. *Cognition and Instruction*, *18*, 95–142.

Office of Head Start. (2015). *A national overview of grantee CLASS® scores in 2015*. Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/data/class-reports/docs/national-class-2015-data.pdf

Patrick, H., & Mantzicopoulos, P. (2015). The role of meaning systems in the development of motivation. In C. M. Rubie-Davies, J. M. Stephens, & P. Watson (Eds.), *The Routledge international handbook of social psychology of the classroom* (pp. 67–79). New York, NY: Routledge.

Patrick, H., Mantzicopoulos, P., & Sears, D. (2012). Effective classrooms. In K. Harris, S. Graham & T. Urdan (Eds.), *APA educational psychology handbook. Volume 2: Individual differences and cultural and contextual factors* (pp. 443–469). Washington, DC: American Psychological Association.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system manual K-3*. Baltimore, MD: Brookes Publishing.

Piburn, M., Sawada, D., Turley, J., Falconer, K., Benford, R., Bloom, I., & Judson, E. (2000). *Reformed Teaching Observation Protocol (RTOP: Reference manual)*. ACEPT Technical Report No. IN003. Tempe, AZ: Arizona Collaborative for Excellence in the Preparation of Teachers. Retrieved from http://files.eric.ed.gov/fulltext/ED447205.pdf

Plank, S. B., & Condliffe, B. F. (2013). Pressures of the season: An examination of classroom quality and high-stakes accountability. *American Educational Research Journal*, *50*, 1152–1182.

Ponitz, C. C., Rimm-Kaufman, S. E., Grimm, K. J., & Curby, T. W. (2009). Kindergarten classroom quality, behavioral engagement, and reading achievement. *School Psychology Review*, *38*, 102–120.

Quinn, D. M., & Cooc, N. (2015). Science achievement gaps by gender and race/ethnicity in elementary and middle school: Trends and predictors. *Educational Researcher*, *44*, 336–346.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, 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). Rochester, NY: University of Rochester Press.
Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The Reformed Teaching Observation Protocol. *School Science and Mathematics*, *102*, 205–253.
Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize?. *Journal of Personality Research*, *47*, 609–612.
Schweinle, A., Meyer, D. K., & Turner, J. C. (2006). Striking the right balance: Students' motivation and affect in elementary mathematics. *The Journal of Educational Research*, *99*, 271–293.
Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*, 1–22.
Soenens, B., Sierens, E., Vansteenkiste, M., Dochy, F., & Goossens, L. (2012). Psychologically controlling teaching: Examining outcomes, antecedents, and mediators. *Journal of Educational Psychology*, *104*, 108–120.
Stipek, D., Feiler, R., Daniels, & Milburn, S. (1995). Effects of different instructional approaches on young children's achievement and motivation. *Child Development*, *66*, 209–223.
Sullivan, J. P. (2012). A collaborative effort: Peer review and the history of teacher evaluations in Montgomery County, Maryland. *Harvard Educational Review*, *82*, 142–152.
Trundle, K. C., & Saçkes, M. (Eds.) (2015). *Research in early childhood science education*. NY: Springer.
Urdan, T., & Turner, J. (2005). Competence motivation in the classroom. In A. Elliot & C. Dweck (Eds.), *Handbook of competence and motivation* (pp. 297–317). New York: Guilford Press.
U.S. Department of Education. (2009). *Race to the Top program: Executive summary*. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf
U.S. Department of Education. (2011). *Fact sheet: Bringing flexibility and focus to education law*. Retrieved from http://www.whitehouse.gov/sites/default/files/fact_sheet_bringing_flexibility_and_focus_to_education_law_0.pdf
U.S. Department of Education. (2012). *ESEA flexibility: Flexibility to improve student academic achievement and increase the quality of instruction*. Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html
Walkington, C., Arora, P., Ihorn, S., Gordon, J., Walker, M., Abraham, L., & Marder, M. (2012). *Development of the UTeach Observation Protocol: A classroom observation instrument to evaluate mathematics and science teachers from the UTeach preparation program*. Retrieved from http://cwalkington.com/UTOP_Paper_2011.pdf
Wentzel, K. R., & Miele, D. R. (Eds.). (2016). *Handbook of motivation at school* (2nd ed.): New York, NY: Routledge.
Whitehurst, G. J., Chingos, M. M., & Lindquist, K. (2014). *Evaluating teachers with classroom observations. Lessons learned in four districts*. Retrieved from http://www.brookings.edu/~/media/research/files/reports/2014/05/13-teacher-evaluation/evaluating-teachers-with-classroom-observations.pdf
Wigfield, A., & Cambria, J. (2010). Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review*, *30*, 1–35.
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., Schiefele, U., Roeser, R. W., & Davis-Kean, P. (2006). Development of achievement motivation. In W. Damon, R. M. Lerner (Series Eds.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (6th ed., pp. 933–1002). New York: Wiley.
Wigfield, A, Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A. J. A., Freedman-Doan, C., & Blumenfeld, P. C. (1997). Change in children's competence beliefs and subjective task values across the elementary school years: A 3-year study. *Journal of Educational Psychology*, *89*, 451–469.
Williford, A. P., Maier, M. F., Downer, J. T., Pianta, R. C., & Howes, C. (2013). Understanding how children's engagement and teachers' interactions combine to predict school readiness. *Journal of Applied Developmental Psychology*, *34*, 299–309.
Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of achievement*. Itasca, IL: Riverside Publishing.
Yore, L. D., Henriques, L., Crawford, B., Smith, L., Gomez-Zwiep, S., & Tillotson, J. (2008). Selecting and using inquiry approaches to teach science: The influence of context in elementary, middle, and secondary schools. In E. Abrams, S. A. Southerland, & P. Silva (Eds.), *Inquiry in the classroom: Realities and opportunities* (pp. 41–87). Charlotte, NC: Information Age.