Educational Evaluation and Policy Analysis
March 2017, Vol. 39, No. 1, pp. 146–170
DOI: 10.3102/0162373716670260

Teacher and Teaching Effects on Students' Attitudes and Behaviors

David Blazar
Harvard University
Matthew A. Kraft
Brown University

Research has focused predominantly on how teachers affect students' achievement on standardized tests despite evidence that a broad range of attitudes and behaviors are equally important to their long-term success. We find that upper-elementary teachers have large effects on self-reported measures of students' self-efficacy in math, and happiness and behavior in class. Students' attitudes and behaviors are predicted by teaching practices most proximal to these measures, including teachers' emotional support and classroom organization. However, teachers who are effective at improving test scores often are not equally effective at improving students' attitudes and behaviors. These findings lend empirical evidence to well-established theory on the multidimensional nature of teaching and the need to identify strategies for improving the full range of teachers' skills.

Keywords: *teacher effectiveness, instruction, noncognitive outcomes, self-efficacy, happiness, behavior*

Introduction

Empirical research on the education production function traditionally has examined how teachers and their background characteristics contribute to students' performance on standardized tests (Hanushek & Rivkin, 2010; Todd & Wolpin, 2003). However, a substantial body of evidence indicates that student learning is multidimensional, with many factors beyond their core academic knowledge as important contributors to both short- and long-term success.[1] For example, psychologists find that emotion and personality influence the quality of one's thinking (Baron, 1982) and how much a child learns in school (Duckworth, Quinn, & Tsukayama, 2012). Longitudinal studies document the strong predictive power of measures of childhood self-control, emotional stability, persistence, and motivation on health and labor market outcomes in adulthood (Borghans, Duckworth, Heckman, & Ter Weel, 2008; Chetty et al., 2011; Moffitt et al., 2011). In fact, these sorts of attitudes and behaviors are stronger predictors of some long-term outcomes than test scores (Chetty et al., 2011).

Consistent with these findings, decades worth of theory also have characterized teaching as multidimensional. High-quality teachers are thought and expected not only to raise test scores but also to provide emotionally supportive environments that contribute to students' social and emotional development, manage classroom behaviors, deliver accurate content, and support critical thinking (Cohen, 2011; Lampert, 2001; Pianta & Hamre, 2009). In recent years, two research traditions have emerged to test this theory using empirical evidence. The first tradition has focused on observations of classrooms as a means of identifying unique domains of teaching practice (Blazar, Braslow, Charalambous, & Hill, 2015; Hamre et al., 2013). Several of these

domains, including teachers' interactions with students, classroom organization, and emphasis on critical thinking within specific content areas, aim to support students' development in areas beyond their core academic skill. The second research tradition has focused on estimating teachers' contribution to student outcomes, often referred to as "teacher effects" (Chetty, Friedman, & Rockoff, 2014; Hanushek & Rivkin, 2010). These studies have found that, as with test scores, teachers vary considerably in their ability to affect students' social and emotional development and a variety of observed school behaviors (Backes & Hansen, 2015; Gershenson, 2016; Jackson, 2012; Jennings & DiPrete, 2010; Koedel, 2008; Kraft & Grace, 2016; Ladd & Sorensen, 2015; Ruzek, Domina, Conley, Duncan, & Karabenick, 2015). Furthermore, weak to moderate correlations between teacher effects on different student outcomes suggest that test scores alone cannot identify teachers' overall skill in the classroom.

Our study is among the first to integrate these two research traditions, which largely have developed in isolation. Working at the intersection of these traditions, we aim both to minimize threats to internal validity and to open up the "black box" of teacher effects by examining whether certain dimensions of teaching practice predict students' attitudes and behaviors. We refer to these relationships between teaching practice and student outcomes as "teaching effects." Specifically, we ask the following three research questions:

Research Question 1: To what extent do teachers affect students' attitudes and behaviors in class?

Research Question 2: To what extent do specific teaching practices affect students' attitudes and behaviors in class?

Research Question 3: Are teachers who are effective at raising test-score outcomes equally effective at developing positive attitudes and behaviors in class?

To answer our research questions, we draw on a rich dataset from the National Center for Teacher Effectiveness (NCTE) of upper-elementary classrooms that collected teacher–student links, observations of teaching practice scored on two established instruments, students' math performance on both high- and low-stakes tests, and a student survey that captured their attitudes and behaviors in class. We used this survey to construct our three primary outcomes: students' self-reported self-efficacy in math, happiness in class, and behavior in class. All three measures are important outcomes of interest to researchers, policymakers, and parents (Borghans et al., 2008; Chetty et al., 2011; Farrington et al., 2012). They also align with theories linking teachers and teaching practice to outcomes beyond students' core academic skills (Bandura, Barbaranelli, Caprara, & Pastorelli, 1996; Pianta & Hamre, 2009), allowing us to test these theories explicitly.

We find that upper-elementary teachers have substantive impacts on students' self-reported attitudes and behaviors in addition to their math performance. We estimate that the variation in teacher effects on students' self-efficacy in math and behavior in class is of similar magnitude to the variation in teacher effects on math test scores. The variation of teacher effects on students' happiness in class is even larger. Furthermore, these outcomes are predicted by teaching practices most proximal to these measures, thus aligning with theory and providing important face and construct validity to these measures. Specifically, teachers' emotional support for students is related both to their self-efficacy in math and happiness in class. Teachers' classroom organization predicts students' reports of their own behavior in class. Errors in teachers' presentation of mathematical content are negatively related to students' self-efficacy in math and happiness in class, as well as students' math performance. Finally, we find that teachers are not equally effective at improving all outcomes. Compared with a correlation of .64 between teacher effects on our two math achievement tests, the strongest correlation between teacher effects on students' math achievement and effects on their attitudes or behaviors is .19.

Together, these findings add further evidence for the multidimensional nature of teaching and, thus, the need for researchers, policymakers, and practitioners to identify strategies for improving these skills. In our conclusion, we discuss several ways that policymakers and practitioners may start to do so, including through the design and implementation of teacher evaluation systems,

professional development, recruitment, and strategic teacher assignments.

Review of Related Research

Theories of teaching and learning have long emphasized the important role teachers play in supporting students' development in areas beyond their core academic skill. For example, in their conceptualization of high-quality teaching, Pianta and Hamre (2009) described a set of emotional supports and organizational techniques that are equally important to learners as teachers' instructional methods. They posit that, by providing "emotional support and a predictable, consistent, and safe environment" (p. 113), teachers can help students become more self-reliant, motivated to learn, and willing to take risks. Furthermore, by modeling strong organizational and management structures, teachers can help build students' own ability to self-regulate. Content-specific views of teaching also highlight the importance of teacher behaviors that develop students' attitudes and behaviors in ways that may not directly affect test scores. In mathematics, researchers and professional organizations have advocated for teaching practices that emphasize critical thinking and problem solving around authentic tasks (Lampert, 2001; National Council of Teachers of Mathematics [NCTM], 1989, 2014). Others have pointed to teachers' important role of developing students' self-efficacy and decreasing their anxiety in math (Bandura et al., 1996; Usher & Pajares, 2008; Wigfield & Meece, 1988).

In recent years, development and use of observation instruments that capture the quality of teachers' instruction have provided a unique opportunity to examine these theories empirically. One instrument in particular, the Classroom Assessment Scoring System (CLASS), is organized around "meaningful patterns of [teacher] behavior . . . tied to underlying developmental processes [in students]" (Pianta & Hamre, 2009, p. 112). Factor analyses of data collected by this instrument have identified several unique aspects of teachers' instruction: teachers' social and emotional interactions with students, their ability to organize and manage the classroom environment, and their instructional supports in the delivery of content (Hafen et al., 2015; Hamre et al., 2013). A number of studies from developers of the CLASS instrument and their colleagues have described relationships between these dimensions and closely related student attitudes and behaviors. For example, teachers' interactions with students predict students' social competence, engagement, and risk taking; teachers' classroom organization predict students' engagement and behavior in class (Burchinal et al., 2008; Downer, Rimm-Kaufman, & Pianta, 2007; Hamre, Hatfield, Pianta, & Jamil, 2014; Hamre & Pianta, 2001; Luckner & Pianta, 2011; Mashburn et al., 2008; Pianta, La Paro, Payne, Cox, & Bradley, 2002). With only a few exceptions (see Downer et al., 2007; Hamre & Pianta, 2001; Luckner & Pianta, 2011), though, these studies have focused on pre-kindergarten settings.

Additional content-specific observation instruments highlight several other teaching competencies with links to students' attitudes and behaviors. For example, in this study, we draw on the Mathematical Quality of Instruction (MQI) to capture math-specific dimensions of teachers' classroom practice. Factor analyses of data captured both by this instrument and the CLASS identified two teaching skills in addition to those described above: the cognitive demand of math activities that teachers provide to students and the precision with which they deliver this content (Blazar et al., 2015). Validity evidence for the MQI has focused on the relationship between these teaching practices and students' math test scores (Blazar, 2015; Kane & Staiger, 2012), which makes sense given the theoretical link between teachers' content knowledge, delivery of this content, and students' own understanding (Hill et al., 2008). However, professional organizations and researchers also describe theoretical links between the sorts of teaching practices captured on the MQI and student outcomes beyond test scores (Bandura et al., 1996; Lampert, 2001; NCTM, 1989, 2014; Usher & Pajares, 2008; Wigfield & Meece, 1988) that, to our knowledge, have not been tested.

In a separate line of research, several recent studies have borrowed from the literature on teachers' "value-added" to student test scores to document the magnitude of teacher effects on a range of other outcomes. These studies attempt to isolate the unique effect of teachers on nontested outcomes from factors outside of teachers' control

(e.g., students' prior achievement, race, gender, socioeconomic status) and to limit any bias due to nonrandom sorting. Jennings and DiPrete (2010) estimated the role that teachers play in developing kindergarten and first-grade students' social and behavioral outcomes. They found within-school teacher effects on social and behavioral outcomes that were even larger (0.21 standard deviation [*SD*]) than effects on students' academic achievement (between 0.12 *SD* and 0.15 *SD*, depending on grade level and subject area). In a study of 35 middle school math teachers, Ruzek et al. (2015) found small but meaningful teacher effects on students' motivation between 0.03 *SD* and 0.08 *SD* among seventh graders. Kraft and Grace (2016) found teacher effects on students' self-reported measures of grit, growth mind-set, and effort in class ranging between 0.14 and 0.17 *SD*. Additional studies identified teacher effects on students' observed school behaviors, including absences, suspensions, grades, grade progression, and graduation (Backes & Hansen, 2015; Gershenson, 2016; Jackson, 2012; Koedel, 2008; Ladd & Sorensen, 2015).

To date, evidence is mixed on the extent to which teachers who improve test scores also improve other outcomes. Four of the studies described above found weak relationships between teacher effects on students' academic performance and effects on other outcome measures. Compared with a correlation of .42 between teacher effects on math versus reading achievement, Jennings and DiPrete (2010) found correlations of .15 between teacher effects on students' social and behavioral outcomes and effects on either math or reading achievement. Kraft and Grace (2016) found that correlations between teacher effects on achievement outcomes and effects on multiple social-emotional competencies were sometimes nonexistent and never greater than .23. Similarly, Gershenson (2016) and Jackson (2012) found weak or null relationships between teacher effects on students' academic performance and effects on observed schools behaviors. However, correlations from two other studies were larger. Ruzek et al. (2015) estimated a correlation of .50 between teacher effects on achievement versus effects on students' motivation in math class. Mihaly, McCaffrey, Staiger, and Lockwood (2013) found a correlation of .57 between middle school teacher effects on students' self-reported effort versus effects on math test scores.

Our analyses extend this body of research by estimating teacher effects on additional attitudes and behaviors captured by students in upper-elementary grades. Our data offer the unique combination of a moderately sized sample of teachers and students with lagged survey measures. We also utilize similar econometric approaches to test the relationship between teaching practice and these same attitudes and behaviors. These analyses allow us to examine the face validity of our teacher effect estimates and the extent to which they align with existing theory.

Data and Sample

Beginning in the 2010 to 2011 school year, the NCTE engaged in a 3-year data collection process. Data came from participating fourth- and fifth-grade teachers ($N = 310$) in four anonymous, medium to large school districts on the East coast of the United States who agreed to have their classes videotaped, complete a teacher questionnaire, and help collect a set of student outcomes. Teachers were clustered within 52 schools, with an average of six teachers per school. Although NCTE focused on teachers' math instruction, participants were generalists who taught all subject areas. This is important, as it allowed us to isolate the contribution of individual teachers to students' attitudes and behaviors, which is considerably more challenging when students are taught by multiple teachers. It also suggests that the observation measures, which assessed teachers' instruction during math lessons, are likely to capture aspects of their classroom practice that are common across content areas.

In Table 1, we present descriptive statistics on participating teachers and their students. We do so for the full NCTE sample, as well as for a subsample of teachers whose students were in the project in both the current and prior years. This latter sample allowed us to capture prior measures of students' attitudes and behaviors, a strategy that we use to increase internal validity and that we discuss in more detail below.[2] When we compare these samples, we find that teachers look relatively similar with no statistically significant differences on any observable characteristic. Reflecting national patterns, the

TABLE 1

Participant Demographics

	Full sample	Attitudes and behaviors sample	*p* value on difference
Teachers			
Male	0.16	0.16	.949
African American	0.22	0.22	.972
Asian	0.03	0.00	.087
Hispanic	0.03	0.03	.904
White	0.65	0.66	.829
Mathematics coursework (1 to 4 Likert-type scale)	2.58	2.55	.697
Mathematical Content Knowledge (standardized scale)	0.01	0.03	.859
Alternative certification	0.08	0.08	.884
Teaching experience (years)	10.29	10.61	.677
Value added on high-stakes math test (standardized scale)	0.01	0.00	.505
Observations	310	111	
Students			
Male	0.50	0.49	.371
African American	0.40	0.40	.421
Asian	0.08	0.07	.640
Hispanic	0.23	0.20	.003
White	0.24	0.28	<.001
FRPL	0.64	0.59	.000
SPED	0.11	0.09	.008
LEP	0.20	0.14	<.001
Prior score on high-stakes math test (standardized scale)	0.10	0.18	<.001
Prior score on high-stakes ELA test (standardized scale)	0.09	0.20	<.001
Observations	10,575	1,529	

Note. FRPL = free- or reduced-price lunch; SPED = special education; LEP = limited English proficiency; ELA = English Language Arts.

vast majority of elementary teachers in our sample are White females who earned their teaching credential through traditional certification programs. (See Hill, Blazar, & Lynch, 2015, for a discussion of how these teacher characteristics were measured.)

Students in our samples look similar to those in many urban districts in the United States. In these national data, roughly 68% are eligible for free or reduced-price lunch (FRPL), 14% are classified as in need of special education (SPED) services, and 16% are identified as limited English proficient (LEP); roughly 31% are African American, 39% are Hispanic, and 28% are White (Council of the Great City Schools, 2013). We do observe some statistically significant differences between student characteristics in the full sample versus our analytic subsample. For example, the percentage of LEP students was 20% in the full sample compared with 14% in the sample of students who ever were part of analyses drawing on our survey measures. Although variation in samples could result in dissimilar estimates across models, the overall character of our findings is unlikely to be driven by these modest differences.

Students' Attitudes and Behaviors

As part of the expansive data collection effort, researchers administered a student survey with items ($n = 18$) that were adapted from other large-scale surveys including the Tripod, the Measures of Effective Teaching (MET) project, the National Assessment of Educational Progress (NAEP), and the Trends in International Mathematics and Science Study (TIMSS) (see the Appendix for a full list of items). Items were selected based on a review of the research literature and identification of constructs thought most likely to be influenced by upper-elementary teachers. Students rated all items on a 5-point Likert-type scale where 1 = *totally untrue* and 5 = *totally true*.

We identified a parsimonious set of three outcome measures based on a combination of theory and exploratory factor analyses (see the Appendix).[3] The first outcome, which we call Self-Efficacy in Math (10 items), is a variation on well-known constructs related to students' effort, initiative, and perception that they can complete tasks. The second related outcome measure is Happiness in Class (five items), which was collected in the second and third years of the study. Exploratory factor analyses suggest that these items cluster together with those from Self-Efficacy in Math to form a single construct. However, post hoc review of these items against the psychology literature from which they were derived suggests that they can be divided into a separate domain. As above, this measure is a school-specific version of well-known scales that capture students' affect and enjoyment (Diener, 2000). Both Self-Efficacy in Math and Happiness in Class have relatively high internal consistency reliabilities (.76 and .82, respectively) that are similar to those of self-reported attitudes and behaviors explored in other studies (Duckworth, Peterson, Matthews, & Kelly, 2007; John & Srivastava, 1999; Tsukayama, Duckworth, & Kim, 2013). Furthermore, self-reported measures of similar constructs have been linked to long-term outcomes, including academic engagement and earnings in adulthood, even conditioning on cognitive ability (King, McInerney, Ganotice, & Villarosa, 2015; Lyubomirsky, King, & Diener, 2005).

The third and final construct consists of three items that were meant to hold together and which we call Behavior in Class (internal consistency reliability is .74). Higher scores reflect better, less disruptive behavior. Teacher reports of students' classroom behavior have been found to relate to antisocial behaviors in adolescence, criminal behavior in adulthood, and earnings (Chetty et al., 2011; Moffitt et al., 2011; Segal, 2013; Tremblay et al., 1992). Our analysis differs from these other studies in the self-reported nature of the behavior outcome. That said, measurement studies suggest that we can draw valid conclusions from our student-reported data. For example, other work has found that student reports of their own behavior correlates with teacher and parent reports at similar magnitudes to how teacher and parent reports of student behavior correlate with each other (Achenbach, McConaughy, & Howell, 1987; Goodman, 2001). Furthermore, relationships between teacher-reported behavior and elementary students' math achievement in other studies (between $r = .22$ and .28; Miles & Stipek, 2006; Tremblay et al., 1992) are very similar to the correlations we find between students' self-reported Behavior in Class and our two math test scores ($r = .24$ and .26; see Table 2). Together, this evidence provides both convergent and consequential validity evidence for this outcome measure.

For all three of these outcomes, we created final scales by reverse coding items with negative valence and averaging raw student responses across all available items.[4] We standardized these final scores within years, given that, for some measures, the set of survey items varied across years.

Student Demographic and Test Score Information

Student demographic and achievement data came from district administrative records. Demographic data include gender, race/ethnicity, FRPL eligibility, LEP status, and SPED status. These records also included current- and prior-year test scores in math and English Language Arts (ELA) on state assessments, which we standardized within districts by grade, subject, and year using the entire sample of students.

The project also administered a low-stakes mathematics assessment to all students in the study. Internal consistency reliability is .82 or higher for each form across grade levels and school years (Hickman, Fu, & Hill, 2012). We

used this assessment in addition to high-stakes tests given that teacher effects on two outcomes that aim to capture similar underlying constructs (i.e., math achievement) provide a unique point of comparison when examining the relationship between teacher effects on student outcomes that are less closely related (i.e., math achievement vs. attitudes and behaviors). Indeed, students' high- and low-stakes math test scores are correlated more strongly (r = .70) than any other two outcomes (see Table 1).[5]

Mathematics Lessons

Teachers' mathematics lessons were captured over a 3-year period with an average of three lessons per teacher per year.[6] Trained raters scored these lessons on two established observation instruments, the CLASS and the MQI. Analyses of these same data show that items cluster into four main factors (Blazar et al., 2015). The two dimensions from the CLASS instrument capture general teaching practices: Emotional Support focuses on teachers' interactions with students and the emotional environment in the classroom and is thought to increase students' social and emotional development; and Classroom Organization focuses on behavior management and productivity of the lesson and is thought to improve students' self-regulatory behaviors (Pianta & Hamre, 2009).[7] The two dimensions from the MQI capture mathematics-specific practices: Ambitious Mathematics Instruction focuses on the complexity of the tasks that teachers provide to their students and their interactions around the content, thus corresponding to the set of professional standards described by NCTM (1989, 2014) and many elements contained within the Common Core State Standards for Mathematics (National Governors Association Center for Best Practices, 2010); Mathematical Errors identifies any mathematical errors or imprecisions the teacher introduces into the lesson. For this last dimension, higher scores indicate that teachers made more errors in their instruction and, therefore, worse performance. Both dimensions from the MQI are linked to teachers' mathematical knowledge for teaching and, in turn, to students' math achievement (Blazar, 2015; Hill et al., 2008; Hill, Schilling, & Ball, 2004). Correlations between dimensions range from roughly 0 (between Emotional Support and Mathematical Errors) to .46 (between Emotional Support and Classroom Organization; see Table 3).

We estimated reliability for these metrics by calculating the amount of variance in teacher scores that is attributable to the teacher (the intraclass correlation [ICC]), adjusted for the modal number of lessons. These estimates are .53, .63, .74, and .56 for Emotional Support, Classroom Organization, Ambitious Mathematics Instruction, and Mathematical Errors, respectively (see Table 3). Although some of these estimates are lower than conventionally acceptable levels (.7), they are consistent with those generated from similar studies (Kane & Staiger, 2012). We standardized scores within the full sample of teachers to have a mean of zero and a standard deviation of one.

Empirical Strategy

Estimating Teacher Effects on Students' Attitudes and Behaviors

Like others who aim to examine the contribution of individual teachers to student outcomes, we began by specifying an education production function model of each outcome for student i in district d, school s, grade g, class c with teacher j at time t:

$$\begin{aligned} \text{OUTCOME}_{idsgjct} &= \alpha f\left(A_{it-1}\right) + \\ &\pi X_{it} + \varphi \bar{X}_{it}^{c} + \tau_{dgt} + \left(\mu_{j} + \delta_{jc} + \varepsilon_{idsgjct}\right). \end{aligned} \quad (1)$$

$\text{OUTCOME}_{idsgjct}$ is used interchangeably for both math test scores and students' attitudes and behaviors, which we modeled in separate equations as a cubic function of students' prior achievement, A_{it-1}, in both math and ELA on the high-stakes district tests[8]; demographic characteristics, X_{it}, including gender, race, FRPL eligibility, SPED status, and LEP status; these same test-score variables and demographic characteristics averaged to the class level, $\bar{X}_{it}^{c}$; and district-by-grade-by-year fixed effects, τ_{dgt}, that account for scaling of high-stakes test at this level. The residual portion of the model can be decomposed into a teacher effect, μ_j, which is our main parameter of interest and captures the contribution of teachers to student outcomes above and beyond factors already controlled for in the model; a class effect, δ_{jc}, which is estimated by

TABLE 2

Descriptive Statistics for Students' Academic Performance, Attitudes, and Behaviors

	Univariate statistics			Pairwise correlations				
	M	*SD*	Internal consistency reliability	High-Stakes math test	Low-Stakes math test	Self-Efficacy in Math	Happiness in Class	Behavior in Class
High-Stakes math test	0.10	0.91	—	1.00				
Low-Stakes math test	0.61	1.1	.82	.70***	1.00			
Self-Efficacy in Math	4.17	0.58	.76	.25***	.22***	1.00		
Happiness in Class	4.10	0.85	.82	.15***	.10***	.62***	1.00	
Behavior in Class	4.10	0.93	.74	.24***	.26***	.35***	.27***	1.00

Note. For High-Stakes math test, reliability varies by district. Self-Efficacy in Math, Happiness in Class, and Behavior in Class are measured on a 1 to 5 Likert-type Scale. Statistics were generated from all available data.
***$p < .001$.

TABLE 3

Descriptive Statistics for CLASS and MQI Dimensions

	Univariate statistics			Pairwise correlations			
	M	*SD*	Adjusted intraclass correlation	Emotional Support	Classroom Organization	Ambitious Mathematics Instruction	Mathematical Errors
Emotional Support	4.28	0.48	.53	1.00			
Classroom Organization	6.41	0.39	.63	.46***	1.00		
Ambitious Mathematics Instruction	1.27	0.11	.74	.22***	.23***	1.00	
Mathematical Errors	1.12	0.09	.56	.01	.09	−.27***	1.00

Note. Intraclass correlations were adjusted for the modal number of lessons. CLASS items (from Emotional Support and Classroom Organization) were scored on a scale from 1 to 7. MQI items (from Ambitious Instruction and Errors) were scored on a scale from 1 to 3. Statistics were generated from all available data. CLASS = Classroom Assessment Scoring System; MQI = Mathematical Quality of Instruction.
***$p < .001$.

observing teachers over multiple school years; and a student-specific error term, $\varepsilon_{idsgjct}$.[9]

The key identifying assumption of this model is that teacher effect estimates are not biased by nonrandom sorting of students to teachers. Recent experimental (Kane, McCaffrey, Miller, & Staiger, 2013) and quasi-experimental (Chetty et al., 2014) analyses provide strong empirical support for this claim when student achievement is the outcome of interest. However, much less is known about bias and sorting mechanisms when other outcomes are used. For example, it is quite possible that students were sorted to teachers based on their classroom behavior in ways that were unrelated to their prior achievement. To address this possibility, we made two modifications to Equation 1. First, we included school fixed effects, ω_s, to account for sorting of

students and teachers across schools. This means that estimates rely only on between-school variation, which has been common practice in the literature estimating teacher effects on student achievement. In their review of this literature, Hanushek and Rivkin (2010) proposed ignoring the between-school component because it is "surprisingly small" and because including this component leads to "potential sorting, testing, and other interpretative problems" (p. 268). Other recent studies estimating teacher effects on student outcomes beyond test scores have used this same approach (Backes & Hansen, 2015; Gershenson, 2016; Jackson, 2012; Jennings & DiPrete, 2010; Ladd & Sorensen, 2015; Ruzek et al., 2015). Another important benefit of using school fixed effects is that this approach minimizes the possibility of reference bias in our self-reported measures (Duckworth & Yeager, 2015; West et al., 2016). Differences in school-wide norms around behavior and effort may change the implicit standard of comparison (i.e., reference group) that students use to judge their own behavior and effort. Restricting comparisons with other teachers and students within the same school minimizes this concern. As a second modification for models that predict each of our three student survey measures, we included OUTCOME_{it-1} on the right-hand side of the equation in addition to prior achievement—that is, when predicting students' Behavior in Class, we controlled for students' self-reported Behavior in Class in the prior year.[10] This strategy helps account for within-school sorting on factors other than prior achievement.

Using our modified version of Equation 1, we estimated the variance of μ_j, which is the stable component of teacher effects. We report the standard deviation of these estimates across outcomes. This parameter captures the magnitude of the variability of teacher effects. With the exception of teacher effects on students' Happiness in Class, where survey items were not available in the first year of the study, we included δ_{jc} to separate out the time-varying portion of teacher effects, combined with peer effects and any other class-level shocks. The fact that we are able to separate class effects from teacher effects is an important extension of prior studies examining the contribution of teachers to student outcomes beyond test scores, many of which only observed teachers at one point in time.

Following Chetty et al. (2011), we estimated the magnitude of the variance of teacher effects using a direct, model-based estimate derived via restricted maximum likelihood estimation. This approach produces a consistent estimator for the true variance of teacher effects (Raudenbush & Bryk, 2002). Calculating the variation across individual teacher effect estimates using Ordinary Least Squares regression would bias our variance estimates upward because it would conflate true variation with estimation error, particularly in instances where only a handful of students are attached to each teacher. Alternatively, estimating the variation in post hoc predicted "shrunken" empirical Bayes estimates would bias our variance estimates downward relative to the size of the measurement error (Jacob & Lefgren, 2005).

Estimating Teaching Effects on Students' Attitudes and Behaviors

We examined the contribution of teachers' classroom practices to our set of student outcomes by estimating a variation of Equation 1:

$$\begin{aligned}\text{OUTCOME}_{idsgjct} = {} & \beta\widehat{\text{OBSERVATION}}_{lj,-t} + \\ & \alpha f\left(A_{it-1}\right) + \gamma\text{OUTCOME}_{it-1} + \pi X_{it} + \\ & \varphi\bar{X}^{c}_{it} + \omega_s + \tau_{dgt} + \left(\mu_j + \delta_{jc} + \varepsilon_{idsgjct}\right).\end{aligned} \tag{2}$$

This multilevel model includes the same set of control variables as above to account for the non-random sorting of students to teachers and for factors beyond teachers' control that might influence each of our outcomes. We further included a vector of their teacher j's observation scores, $\widehat{\text{OBSERVATION}}_{lj,-t}$. These scores were averaged across lessons l in years other than t (denoted by $-t$). We used predicted shrunken observation score estimates that account for the fact that teachers contributed different numbers of lessons to the project, and fewer lessons could lead to measurement error in these scores (Hill, Charalambous, & Kraft, 2012).[11] The coefficients on these variables are our main parameters of interest and can be interpreted as the change in standard deviation units for each outcome associated with exposure to teaching practice one standard deviation above the mean.

One concern when relating observation scores to student survey outcomes is that they

may capture the same behaviors. For example, teachers may receive credit on the Classroom Organization domain when their students demonstrate orderly behavior. In this case, we would have the same observed behavior on both the left and right side of our equation relating instructional quality to student outcomes, which would inflate our teaching effect estimates. A related concern is that the specific students in the classroom may influence teachers' instructional quality (Hill et al., 2015; Steinberg & Garrett, 2016; Whitehurst, Chingos, & Lindquist, 2014). Although the direction of bias is not as clear here—as either lesser or higher quality teachers could be sorted to harder to educate classrooms—this possibility also could lead to incorrect estimates. We avoid these sources of bias by only including lessons captured in years other than those in which student outcomes were measured, denoted by $-t$ in the subscript of $\widehat{\text{OBSERVATION}}_{ij,-t}$. To the extent that instructional quality varies across years, using out-of-year observation scores creates a lower-bound estimate of the true relationship between instructional quality and student outcomes. We consider this an important trade-off to minimize potential bias.

An additional concern for identification is the endogeneity of observed classroom quality. In other words, specific teaching practices are not randomly assigned to teachers. Our preferred analytic approach attempted to account for potential sources of bias by conditioning estimates of the relationship between one dimension of teaching practice and student outcomes on the three other dimensions. An important caveat here is that we only observed teachers' instruction during math lessons and, thus, may not capture important pedagogical practices teachers used with these students when teaching other subjects. Including dimensions from the CLASS instrument, which are meant to capture instructional quality across subject areas (Pianta & Hamre, 2009), helps account for some of this concern. However, given that we were not able to isolate one dimension of teaching quality from all others, we consider this approach as providing suggestive rather than conclusive evidence on the underlying causal relationship between teaching practice and students' attitudes and behaviors.

Estimating the Relationship Between Teacher Effects Across Multiple Student Outcomes

In our third and final set of analyses, we examined whether teachers who are effective at raising math test scores are equally effective at developing students' attitudes and behaviors. To do so, we drew on our modified version of Equation 1 to estimate $\hat{\mu}_j$ for each outcome and teacher j. Following Chetty et al. (2014), we use post hoc predicted "shrunken" empirical Bayes estimates of $\hat{\mu}_j$ derived from Equation 1. Then, we generated a correlation matrix of these teacher effect estimates.

Despite attempts to increase the precision of these estimates through empirical Bayes estimation, estimates of individual teacher effects are measured with error that will attenuate these correlations (Spearman, 1904). Thus, if we were to find weak to moderate correlations between different measures of teacher effectiveness, this could identify multidimensionality or could result from measurement challenges, including the reliability of individual constructs (Chin & Goldhaber, 2015). For example, prior research suggests that different tests of students' academic performance can lead to different teacher rankings, even when those tests measure similar underlying constructs (Lockwood et al., 2007; Papay, 2011). To address this concern, we focus our discussion on relative rankings in correlations between teacher effect estimates rather than their absolute magnitudes. Specifically, we examine how correlations between teacher effects on two closely related outcomes (e.g., two math achievement tests) compare with correlations between teacher effects on outcomes that aim to capture different underlying constructs. In light of research highlighted above, we did not expect the correlation between teacher effects on the two math tests to be 1 (or, for that matter, close to 1). However, we hypothesized that these relationships should be stronger than the relationship between teacher effects on students' math performance and effects on their attitudes and behaviors.

Results

Do Teachers Affect Students' Attitudes and Behaviors?

We begin by presenting results of the magnitude of teacher effects in Table 4. Here, we

TABLE 4

Teacher Effects on Students' Academic Performance, Attitudes, and Behaviors

	Observations		Standard deviation of teacher-level variance
	Teachers	Students	
High-Stakes math test	310	10,575	0.18
Low-Stakes math test	310	10,575	0.17
Self-Efficacy in Math	108	1,433	0.14
Happiness in Class	51	548	0.31
Behavior in Class	111	1,529	0.15

Note. Cells contain estimates from separate multilevel regression models. All effects are statistically significant at the .05 level.

observe sizable teacher effects on students' attitudes and behaviors that are similar to teacher effects on students' academic performance. Starting first with teacher effects on students' academic performance, we find that a 1 *SD* difference in teacher effectiveness is equivalent to a 0.17 *SD* or 0.18 *SD* difference in students' math achievement. In other words, relative to an average teacher, teachers at the 84th percentile of the distribution of effectiveness move the medium student up to roughly the 57th percentile of math achievement. Notably, these findings are similar to those from other studies that also estimate within-school teacher effects in large administrative datasets (Hanushek & Rivkin, 2010). This suggests that our use of school fixed effects with a more limited number of teachers observed within a given school does not appear to overly restrict our identifying variation. In Appendix A (available in the online version of the journal), where we present the magnitude of teacher effects from alternative model specifications, we show that results are robust to models that exclude school fixed effects or replace school fixed effects with observable school characteristics. Estimated teacher effects on students' self-reported Self-Efficacy in Math and Behavior in Class are 0.14 *SD* and 0.15 *SD*, respectively. The largest teacher effects we observe are on students' Happiness in Class, of 0.31 *SD*. Given that we do not have multiple years of data to separate out class effects for this measure, we interpret this estimate as the upward bound of true teacher effects on Happiness in Class. Rescaling this estimate by the ratio of teacher effects with and without class effects for Self-Efficacy in Math (0.14 / 0.19 = 0.74; see Appendix A available in the online version of the journal) produces an estimate of stable teacher effects on Happiness in Class of 0.23 *SD*, still larger than effects for other outcomes.

Do Specific Teaching Practices Affect Students' Attitudes and Behaviors?

Next, we examine whether certain characteristics of teachers' instructional practice help explain the sizable teacher effects described above. We present unconditional estimates in Table 5 Panel A, where the relationship between one dimension of teaching practice and student outcomes is estimated without controlling for the other three dimensions. Thus, cells contain estimates from separate regression models. In Table 5 Panel B, we present conditional estimates, where all four dimensions of teaching quality are included in the same regression model. We present all estimates as standardized effect sizes, which allows us to make comparisons across models and outcome measures. Unconditional and conditional estimates generally are quite similar. Therefore, we focus our discussion on our preferred conditional estimates. We remind readers that we use out-of-year observation scores to avoid several sources of bias.

We find that students' attitudes and behaviors are predicted by both general and content-specific teaching practices in ways that generally align with theory. For example, teachers' Emotional Support is positively associated with the two closely related student constructs, Self-Efficacy in Math and Happiness in Class. Specifically, a 1 *SD* increase in teachers' Emotional Support is

TABLE 5

Teaching Effects on Students' Academic Performance, Attitudes, and Behaviors

	High-Stakes math test	Low-Stakes math test	Self-Efficacy in Math	Happiness in Class	Behavior in Class
Panel A: Unconditional estimates					
Emotional Support	0.012 (0.013)	0.018 (0.014)	0.142*** (0.031)	0.279*** (0.082)	0.039 (0.027)
Classroom Organization	−0.017 (0.014)	−0.010 (0.014)	0.065† (0.038)	0.001 (0.090)	0.081* (0.033)
Ambitious Mathematics Instruction	0.017 (0.015)	0.021 (0.015)	0.077* (0.036)	0.082 (0.068)	0.004 (0.032)
Mathematical Errors	−0.027* (0.013)	−0.009 (0.014)	−0.107*** (0.030)	−0.164* (0.076)	−0.027 (0.027)
Panel B: Conditional estimates					
Emotional Support	0.015 (0.014)	0.020 (0.015)	0.135*** (0.034)	0.368*** (0.090)	0.030 (0.030)
Classroom Organization	−0.022 (0.014)	−0.018 (0.015)	−0.020 (0.042)	−0.227* (0.096)	0.077* (0.036)
Ambitious Mathematics Instruction	0.014 (0.015)	0.019 (0.016)	−0.006 (0.040)	0.079 (0.068)	−0.034 (0.036)
Mathematical Errors	−0.024† (0.013)	−0.005 (0.014)	−0.094** (0.033)	−0.181* (0.081)	−0.009 (0.029)
Teacher Observations	196	196	90	47	93
Student Observations	8,660	8,660	1,275	517	1,362

Note. In Panel A, cells contain estimates and associated standard errors from separate regression models. In Panel B, all four teaching practice measures are included in the same model. All models control for student and class characteristics, school fixed effects, and district-by-grade-by-year fixed effects, and include teacher random effects. Models predicting all outcomes except for Happiness in Class also include class random effects.

†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

associated with a 0.14 *SD* increase in students' Self-Efficacy in Math and a 0.37 *SD* increase in students' Happiness in Class. These findings make sense given that Emotional Support captures teacher behaviors such as their sensitivity to students, regard for students' perspective, and the extent to which they create a positive climate in the classroom. As a point of comparison, these estimates are substantively larger than those between principal ratings of teachers' ability to improve test scores and their actual ability to do so, which fall in the range of 0.02 *SD* and 0.08 *SD* (Jacob & Lefgren, 2008; Rockoff & Speroni, 2010; Rockoff, Staiger, Kane, & Taylor, 2012).

We also find that Classroom Organization, which captures teachers' behavior management skills and productivity in delivering content, is positively related to students' reports of their own Behavior in Class (0.08 *SD*). This suggests that teachers who create an orderly classroom likely provide a model for students' own ability to self-regulate. Despite this positive relationship, we find that Classroom Organization is negatively associated with Happiness in Class (−0.23 *SD*), suggesting that classrooms that are overly focused on routines and management are negatively related to students' enjoyment in class. At the same time, this is one instance where our estimate is sensitive to whether other teaching characteristics are included in the model. When we estimate the relationship between teachers' Classroom Organization and students' Happiness in Class without controlling for the three other dimensions of teaching quality, this estimate approaches 0 and is no longer statistically significant.[12] We return to a discussion of the potential trade-offs between Classroom Organization and students' Happiness in Class in our conclusion.

Finally, we find that the degree to which teachers commit Mathematical Errors is negatively related to students' Self-Efficacy in Math (−0.09 *SD*) and Happiness in Class (−0.18 *SD*). These findings illuminate how a teacher's ability to present mathematics with clarity and without serious mistakes is related to their students' perceptions that they can complete math tasks and their enjoyment in class.

Comparatively, when predicting scores on both math tests, we only find one marginally significant relationship—between Mathematical Errors and the High-Stakes math test (−0.02 *SD*). For two other dimensions of teaching quality, Emotional Support and Ambitious Mathematics Instruction, estimates are signed the way we would expect and with similar magnitudes, though they are not statistically significant. Given the consistency of estimates across the two math tests and our restricted sample size, it is possible that nonsignificant results are due to limited statistical power.[13] At the same time, even if true relationships exist between these teaching practices and students' math test scores, they likely are weaker than those between teaching practices and students' attitudes and behaviors. For example, we find that the 95% confidence intervals (CIs) relating Classroom Emotional Support to Self-Efficacy in Math [0.068, 0.202] and Happiness in Class [0.162, 0.544] do not overlap with the 95% CIs for any of the point estimates predicting math test scores. We interpret these results as indication that, still, very little is known about how specific classroom teaching practices are related to student achievement in math.[14]

In Appendix B (available in the online version of the journal), we show that results are robust to a variety of different specifications, including (a) adjusting observation scores for characteristics of students in the classroom, (b) controlling for teacher background characteristics (i.e., teaching experience, math content knowledge, certification pathway, education), and (c) using raw out-of-year observation scores (rather than shrunken scores). This suggests that our approach likely accounts for many potential sources of bias in our teaching effect estimates.

Are Teachers Equally Effective at Raising Different Student Outcomes?

In Table 6, we present correlations between teacher effects on each of our student outcomes. The fact that teacher effects are measured with error makes it difficult to estimate the precise magnitude of these correlations. Instead, we describe relative differences in correlations, focusing on the extent to which teacher effects within outcome type—that is, correlations between teacher effects on the two math achievement tests, or correlations between teacher effects on two measures of students' attitudes and behaviors—are similar or

TABLE 6

Correlations Between Teacher Effects on Students' Academic Performance, Attitudes, and Behaviors

	High-Stakes math test	Low-Stakes math test	Self-Efficacy in Math	Happiness in Class	Behavior in Class
High-Stakes math test	1.00				
	—				
Low-Stakes math test	.64***	1.00			
	(0.04)	—			
Self-Efficacy in Math	.16†	.19*	1.00		
	(0.10)	(0.10)	—		
Happiness in Class	−.09	−.21	.26†	1.00	
	(0.14)	(0.14)	(0.14)	—	
Behavior in Class	.10	.12	.49***	.21†	1.00
	(0.10)	(0.10)	(0.08)	(0.14)	—

Note. Standard errors in parentheses. See Table 4 for sample sizes used to calculate teacher effect estimates. The sample for each correlation is the minimum number of teachers between the two measures.
†$p < .10$. *$p < .05$. ***$p < .001$.

different from correlations between teacher effects across outcome type. We illustrate these differences in Figure 1, where Panel A presents scatterplots of these relationships between teacher effects within outcome type, and Panel B does the same for teacher effects across outcome type. Recognizing that not all of our survey outcomes are meant to capture the same underlying construct, we also describe relative differences in correlations between teacher effects on these different measures. In Appendix C (available in the online version of the journal), we find that an extremely conservative adjustment that scales correlations by the inverse of the square root of the product of the reliabilities leads to a similar overall pattern of results.

Examining the correlations of teacher effect estimates reveals that individual teachers vary considerably in their ability to affect different student outcomes. As hypothesized, we find the strongest correlations between teacher effects within outcome type. Similar to Corcoran and Jennings (2012), we estimate a correlation of .64 between teacher effects on our high- and low-stakes math achievement tests. We also observe a strong correlation of .49 between teacher effects on two of the student survey measures, students' Behavior in Class and Self-Efficacy in Math. Comparatively, the correlations between teacher effects across outcome type are much weaker. Examining the scatterplots in Figure 1, we observe much more dispersion around the best-fit line in Panel B than in Panel A. The strongest relationship we observe across outcome types is between teacher effects on the Low-Stakes math test and effects on Self-Efficacy in Math ($r = .19$). The lower bound of the 95% CI around the correlation between teacher effects on the two achievement measures [0.56, 0.72] does not overlap with the 95% CI of the correlation between teacher effects on the Low-Stakes math test and effects on Self-Efficacy in Math [−0.01, 0.39], indicating that these two correlations are substantively and statistically significantly different from each other. Using this same approach, we also can distinguish the correlation describing the relationship between teacher effects on the two math tests from all other correlations relating teacher effects on test scores to effects on students' attitudes and behaviors. We caution against placing too much emphasis on the negative correlations between teacher effects on test scores and effects on Happiness in Class ($r = −.09$ and $−.21$ for the high- and low-stakes tests, respectively). Given limited precision of this relationship, we cannot reject the null hypothesis of no relationship or rule out weak, positive, or negative correlations among these measures.

Although it is useful to make comparisons between the strength of the relationships between teacher effects on different measures of students' attitudes and behaviors, measurement error limits our ability to do so precisely. At face value, we

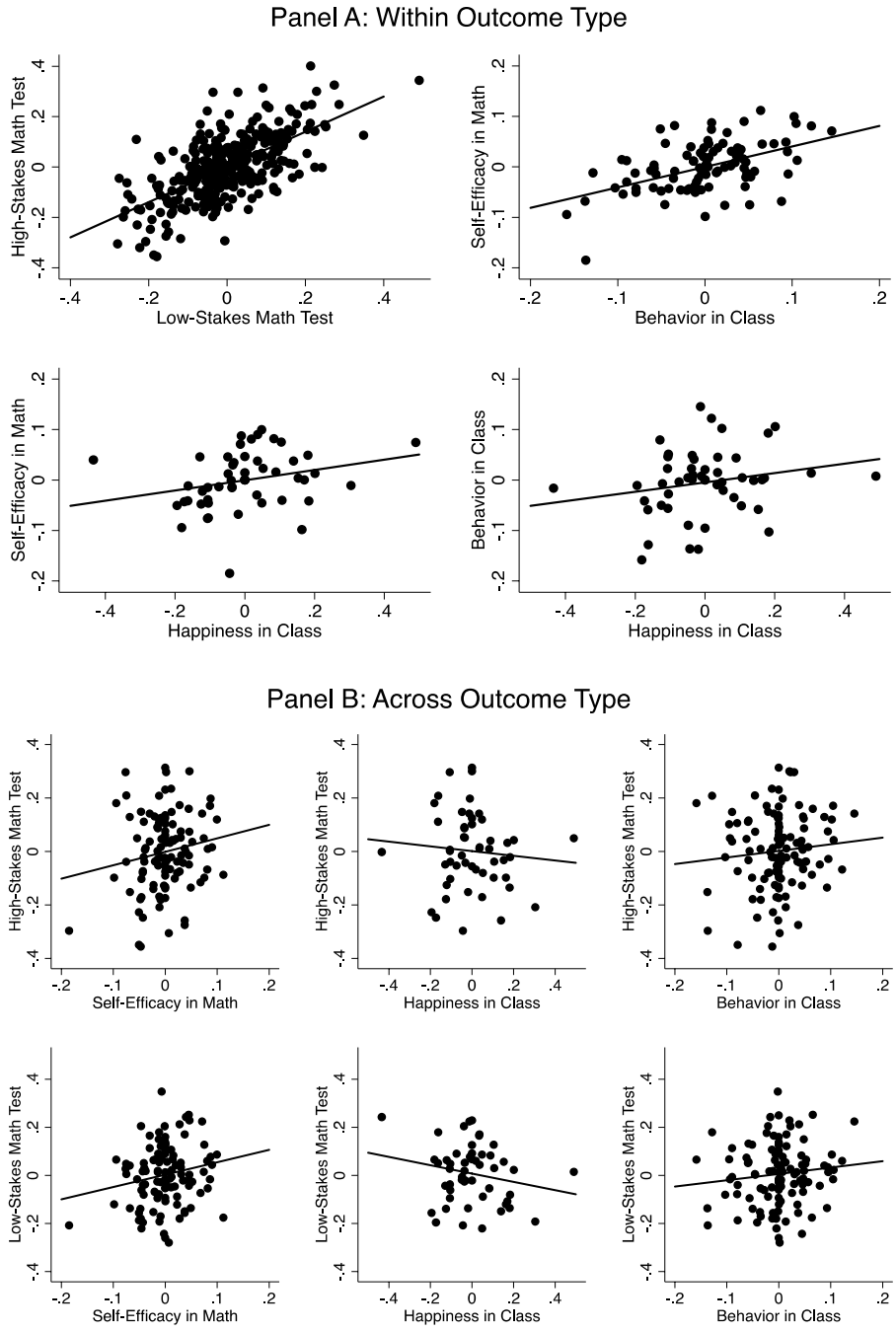

FIGURE 1. *Scatterplots of teacher effects across outcomes.*
Note. Solid lines represent the best-fit regression line.

find correlations between teacher effects on Happiness in Class and effects on the two other survey measures (r = .26 for Self-Efficacy in Math and .21 for Behavior in Class) that are weaker than the correlation between teacher effects on Self-Efficacy in Math and effects on Behavior in Class described above (r = .49). One possible interpretation of these findings is that teachers who improve students' Happiness in Class are not equally effective at raising other

attitudes and behaviors. For example, teachers might make students happy in class in unconstructive ways that do not also benefit their self-efficacy or behavior. At the same time, these correlations between teacher effects on Happiness in Class and the other two survey measures have large CIs, likely due to imprecision in our estimate of teacher effects on Happiness in Class. Thus, we are not able to distinguish either correlation from the correlation between teacher effects on Behavior in Class and effects on Self-Efficacy in Math.

Discussion and Conclusion

Relationship Between Our Findings and Prior Research

The teacher effectiveness literature has profoundly shaped education policy over the last decade, and has served as the catalyst for sweeping reforms around teacher recruitment, evaluation, development, and retention. However, by and large, this literature has focused on teachers' contribution to students' test scores. Even research studies such as the Measures of Effective Teaching project and new teacher evaluation systems that focus on "multiple measures" of teacher effectiveness (Center on Great Teachers and Leaders, 2013; Kane et al., 2013) generally attempt to validate other measures, such as observations of teaching practice, by examining their relationship to estimates of teacher effects on students' academic performance.

Our study extends an emerging body of research examining the effect of teachers on student outcomes beyond test scores. In many ways, our findings align with conclusions drawn from previous studies that also identify teacher effects on students' attitudes and behaviors (Jennings & DiPrete, 2010; Kraft & Grace, 2016; Ruzek et al., 2015), as well as weak relationships between different measures of teacher effectiveness (Gershenson, 2016; Jackson, 2012; Kane & Staiger, 2012). To our knowledge, this study is the first to identify teacher effects on measures of students' Self-Efficacy in Math and Happiness in Class, as well as on a self-reported measure of students' Behavior in Class. These findings suggest that teachers can and do help develop attitudes and behaviors among their students that are important for success in life. By interpreting teacher effects alongside teaching effects, we also provide strong face and construct validity for our teacher effect estimates. We find that improvements in upper-elementary students' attitudes and behaviors are predicted by general teaching practices in ways that align with hypotheses laid out by instrument developers (Pianta & Hamre, 2009). Findings linking errors in teachers' presentation of math content to students' Self-Efficacy in Math, in addition to their math performance, also are consistent with theory (Bandura et al., 1996). Finally, the broad data collection effort from NCTE allows us to examine relative differences in relationships between measures of teacher effectiveness, thus avoiding some concerns about how best to interpret correlations that differ substantively across studies (Chin & Goldhaber, 2015). We find that correlations between teacher effects on student outcomes that aim to capture different underlying constructs (e.g., math test scores and Behavior in Class) are weaker than correlations between teacher effects on two outcomes that are much more closely related (e.g., math achievement).

Implications for Policy

These findings can inform policy in several key ways. First, our findings may contribute to the growing interest in incorporating measures of students' attitudes and behaviors—and teachers' ability to improve these outcomes—into accountability policy (see Duckworth, 2016; Miller, 2015; Zernike, 2016, for discussion of these efforts in the press). After passage of the Every Student Succeeds Act (ESSA, 2015), states now are required to select a nonacademic indicator with which to assess students' success in school. Including measures of students' attitudes and behaviors in accountability or evaluation systems, even with very small associated weights, could serve as a strong signal that schools and educators should value and attend to developing these skills in the classroom.

At the same time, like other researchers (Duckworth & Yeager, 2015), we caution against a rush to incorporate these measures into high-stakes decisions. The science of measuring students' attitudes and behaviors is relatively new compared with the long history of developing valid and reliable assessments of cognitive

aptitude and content knowledge. Most existing measures, including those used in this study, were developed for research purposes rather than large-scale testing with repeated administrations. Open questions remain about whether reference bias substantially distorts comparisons across schools. Similar to previous studies, we include school fixed effects in all of our models, which helps reduce this and other potential sources of bias. However, as a result, our estimates are restricted to within-school comparisons of teachers and cannot be applied to inform the type of across-school comparisons that districts typically seek to make. There also are outstanding questions regarding the susceptibility of these measures to survey coaching when high-stakes incentives are attached. Such incentives likely would render teacher or self-assessments of students' attitudes and behaviors inappropriate. Some researchers have started to explore other ways to capture students' attitudes and behaviors, including objective performance-based tasks and administrative proxies such as attendance, suspensions, and participation in extracurricular activities (Hitt, Trivitt, & Cheng, 2016; Jackson, 2012; Whitehurst, 2016). This line of research shows promise but still is in its early phases. Furthermore, although our modeling strategy aims to reduce bias due to nonrandom sorting of students to teachers, additional evidence is needed to assess the validity of this approach. Without first addressing these concerns, we believe that adding untested measures into accountability systems could lead to superficial and, ultimately, counterproductive efforts to support the positive development of students' attitudes and behaviors.

An alternative approach to incorporating teacher effects on students' attitudes and behaviors into teacher evaluation may be through observations of teaching practice. Our findings suggest that specific domains captured on classroom observation instruments (i.e., Emotional Support and Classroom Organization from the CLASS and Mathematical Errors from the MQI) may serve as indirect measures of the degree to which teachers affect students' attitudes and behaviors. One benefit of this approach is that districts commonly collect related measures as part of teacher evaluation systems (Center on Great Teachers and Leaders, 2013), and such measures are not restricted to teachers who work in tested grades and subjects.

Similar to Whitehurst (2016), we also see alternative uses of teacher effects on students' attitudes and behaviors that fall within and would enhance existing school practices. In particular, measures of teachers' effectiveness at improving students' attitudes and behaviors could be used to identify areas for professional growth and connect teachers with targeted professional development. This suggestion is not new and, in fact, builds on the vision and purpose of teacher evaluation described by many other researchers (Darling-Hammond, 2013; Hill & Grossman, 2013; Papay, 2012). However, to leverage these measures for instructional improvement, we add an important caveat: Performance evaluations—whether formative or summative—should avoid placing teachers into a single performance category whenever possible. Although many researchers and policymakers argue for creating a single-weighted composite of different measures of teachers' effectiveness (Center on Great Teachers and Leaders, 2013; Kane et al., 2013), doing so likely oversimplifies the complex nature of teaching. For example, a teacher who excels at developing students' math content knowledge but struggles to promote joy in learning or students' own Self-Efficacy in Math is a very different teacher than one who is middling across all three measures. Looking at these two teachers' composite scores would suggest they are similarly effective. A single overall evaluation score lends itself to a systematized process for making binary decisions such as whether to grant teachers tenure, but such decisions would be better informed by recognizing and considering the full complexity of classroom practice.

We also see opportunities to maximize students' exposure to the range of teaching skills we examine through strategic teacher assignments. Creating a teacher workforce skilled in most or all areas of teaching practice is, in our view, the ultimate goal. However, this goal likely will require substantial changes to teacher preparation programs and curriculum materials, as well as new policies around teacher recruitment, evaluation, and development. In middle and high schools, content-area specialization or departmentalization often is used to ensure that students have access to teachers with skills in distinct content

areas. Some, including the National Association of Elementary School Principals, also see this as a viable strategy at the elementary level (Chan & Jarman, 2004). Similar approaches may be taken to expose students to a collection of teachers who together can develop a range of academic skills, attitudes, and behaviors. For example, when configuring grade-level teams, principals may pair a math teacher who excels in her ability to improve students' behavior with an ELA or reading teacher who excels in his ability to improve students' happiness and engagement. Viewing teachers as complements to each other may help maximize outcomes within existing resource constraints.

Finally, we consider the implications of our findings for the teaching profession more broadly. Although our findings lend empirical support to research on the multidimensional nature of teaching (Cohen, 2011; Lampert, 2001; Pianta & Hamre, 2009), we also identify tensions inherent in this sort of complexity and potential trade-offs between some teaching practices. In our primary analyses, we find that high-quality instruction around Classroom Organization is positively related to students' self-reported Behavior in Class but negatively related to their Happiness in Class. Our results here are not conclusive, as the negative relationship between Classroom Organization and students' Happiness in Class is sensitive to model specification. However, if there indeed is a negative causal relationship, it raises questions about the relative benefits of fostering orderly classroom environments for learning versus supporting student engagement by promoting positive experiences with schooling. Our own experience as educators and researchers suggests this need not be a fixed trade-off. Future research should examine ways in which teachers can develop classroom environments that engender both constructive classroom behavior and students' happiness and engagement. As our study draws on a small sample of students who had current and prior-year scores for Happiness in Class, we also encourage new studies with greater statistical power that may be able to uncover additional complexities (e.g., nonlinear relationships) in these sorts of data.

Our findings also demonstrate a need to integrate general and more content-specific perspectives on teaching, a historical challenge in both research and practice (Grossman & McDonald, 2008; Hamre et al., 2013). We find that both math-specific and general teaching practices predict a range of student attitudes and behaviors. Yet, particularly at the elementary level, teachers' math training often is overlooked. Prospective elementary teachers often gain licensure without taking college-level math classes; in many states, they do not need to pass the math subsection of their licensure exam to earn a passing grade overall (Epstein & Miller, 2011). Striking the right balance between general and content-specific teaching practices is not a trivial task, but it likely is a necessary one.

For decades, efforts to improve the quality of the teacher workforce have focused on teachers' abilities to raise students' academic achievement. Our work further illustrates the potential and importance of expanding this focus to include teachers' abilities to promote students' attitudes and behaviors that are equally important for students' long-term success.

Appendix

Factor Loadings for Items From the Student Survey

	Year 1		Year 2		Year 3	
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2
Eigenvalue	2.13	0.78	4.84	1.33	5.44	1.26
Proportion of variance explained	0.92	0.34	0.79	0.22	0.82	0.19
Self-Efficacy in Math						
I have pushed myself hard to completely understand math in this class.	0.32	0.18	0.43	0.00	0.44	−0.03
If I need help with math, I make sure that someone gives me the help I need.	0.34	0.25	0.42	0.09	0.49	0.01

(continued)

Appendix (continued)

	Year 1		Year 2		Year 3	
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2
If a math problem is hard to solve, I often give up before I solve it.	−0.46	0.01	−0.38	0.28	−0.42	0.25
Doing homework problems helps me get better at doing math.	0.30	0.31	0.54	0.24	0.52	0.18
In this class, math is too hard.	−0.39	−0.03	−0.38	0.22	−0.42	0.16
Even when math is hard, I know I can learn it.	0.47	0.35	0.56	0.05	0.64	0.02
I can do almost all the math in this class if I don't give up.	0.45	0.35	0.51	0.05	0.60	0.05
I'm certain I can master the math skills taught in this class.			0.53	0.01	0.56	0.03
When doing work for this math class, focus on learning not time work takes.			0.58	0.09	0.62	0.06
I have been able to figure out the most difficult work in this math class.			0.51	0.10	0.57	0.04
Happiness in Class						
This math class is a happy place for me to be.			0.67	0.18	0.68	0.20
Being in this math class makes me feel sad or angry.			−0.50	0.15	−0.54	0.16
The things we have done in math this year are interesting.			0.56	0.24	0.57	0.27
Because of this teacher, I am learning to love math.			0.67	0.26	0.67	0.28
I enjoy math class this year.			0.71	0.21	0.75	0.26
Behavior in Class						
My behavior in this class is good.	0.60	−0.18	0.47	−0.42	0.48	−0.37
My behavior in this class sometimes annoys the teacher.	−0.58	0.40	−0.35	0.59	−0.37	0.61
My behavior is a problem for the teacher in this class.	−0.59	0.39	−0.38	0.60	−0.36	0.57

Note. Estimates drawn from all available data. Loadings of roughly 0.4 or higher are highlighted to identify patterns. Items shaded in light gray cluster onto the first factor. Items shaded in darker gray cluster onto the second factor.

Acknowledgment

The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C090023 to the President and Fellows of Harvard College to support the National Center for Teacher Effectiveness. Additional support came from the William T. Grant Foundation, the Albert Shanker Institute, and Mathematica Policy Research's summer fellowship.

Notes

1. Although student outcomes beyond test scores often are referred to as "noncognitive" skills, our preference, like others (Duckworth & Yeager, 2015; Farrington et al., 2012), is to refer to each competency by name. For brevity, we refer to them as "attitudes

and behaviors," which closely characterizes the measures we focus on in this article.

2. Analyses include additional subsamples of teachers and students. In analyses that predict students' survey response, we included between 51 and 111 teachers and between 548 and 1,529 students. This range is due to the fact that some survey items were not available in the first year of the study. Furthermore, in analyses relating domains of teaching practice to student outcomes, we further restricted our sample to teachers who themselves were part of the study for more than 1 year, which allowed us to use out-of-year observation scores that were not confounded with the specific set of students in the classroom. This reduced our analysis samples to between 47 and 93 teachers and between 517 and 1,362 students when predicting students' attitudes and behaviors, and 196 teachers and 8,660 students when predicting math test scores. Descriptive statistics and formal comparisons of other samples show similar patterns as those presented in Table 1.

3. We conducted exploratory factor analyses separately by year, given that additional items were added in the second and third years to help increase reliability. In the second and third years, we find evidence for two factors. Each of these has an eigenvalue above one, a conventionally used threshold for selecting factors (Kline, 1994). Even though the second factor consists of three items that also have loadings on the first factor greater than 0.40—often taken as the minimum acceptable factor loading (Field, 2013; Kline, 1994)—this second factor explains roughly 20% more of the variation across teachers and, therefore, has strong support for a substantively separate construct (Field, 2013; Tabachnick & Fidell, 2001). In the first year of the study, evidence points to a single factor. The eigenvalue on this second factor is less strong (0.78), the two items that load onto it also load onto the first factor, and the total proportion of variance explained by both factors is greater than 1 suggesting that we specified too many factors. A likely reason for this different solution is that we had roughly half as many items in the first year compared to later years.

4. Depending on the outcome, between 4% and 8% of students were missing a subset of items from survey scales. In these instances, we created final scores by averaging across all available information.

5. Coding of items from both the low- and high-stakes tests also identify a large degree of overlap in terms of content coverage and cognitive demand (Lynch, Chin, & Blazar, 2015). All tests focused most on numbers and operations (40% to 60%), followed by geometry (roughly 15%), and algebra (15% to 20%). By asking students to provide explanations of their thinking and to solve nonroutine problems such as identifying patterns, the low-stakes test also was similar to the high-stakes tests in two districts; in the other two districts, items often asked students to execute basic procedures.

6. As described by Blazar (2015), capture occurred with a three-camera, digital recording device and lasted between 45 and 60 minutes. Teachers were allowed to choose the dates for capture in advance and directed to select typical lessons and exclude days on which students were taking a test. Although it is possible that these lessons were unique from a teachers' general instruction, teachers did not have any incentive to select lessons strategically as no rewards or sanctions were involved with data collection or analyses. In addition, analyses from the Measures of Effective Teaching (MET) project indicate that teachers are ranked almost identically when they choose lessons themselves compared with when lessons are chosen for them (Ho & Kane, 2013).

7. Developers of the Classroom Assessment Scoring System (CLASS) instrument identify a third dimension, Classroom Instructional Support. Factor analyses of data used in this study showed that items from this dimension formed a single construct with items from Emotional Support (Blazar, Braslow, Charalambous, & Hill, 2015). Given theoretical overlap between Classroom Instructional Support and dimensions from the Mathematical Quality of Instruction (MQI) instrument, we excluded these items from our work and focused only on Emotional Support.

8. We controlled for prior-year scores only on the high-stakes assessments and not on the low-stakes assessment for three reasons. First, including prior low-stakes test scores would reduce our full sample by more than 2,200 students. This is because the assessment was not given to students in District 4 in the first year of the study (N = 1,826 students). Furthermore, an additional 413 students were missing fall test scores given that they were not present in class on the day it was administered. Second, prior-year scores on the high- and low-stakes test are correlated at .71, suggesting that including both would not help to explain substantively more variation in our outcomes. Third, sorting of students to teachers is most likely to occur based on student performance on the high-stakes assessments as it was readily observable to schools; achievement on the low-stakes test was not.

9. An alternative approach would be to specify teacher effects as fixed, rather than random, which relaxes the assumption that teacher assignment is uncorrelated with factors that also predict student outcomes (Guarino, Maxfield, Reckase, Thompson, & Wooldridge, 2015). Ultimately, we prefer the random effects specification for three reasons. First, it allows us to separate out teacher effects from class effects by including a random effect for both in our model. Second, this approach allows us to control for a variety of variables that are dropped from the model

when teacher fixed effects also are included. Given that all teachers in our sample remained in the same school from 1 year to the next, school fixed effects are collinear with teacher fixed effects. In instances where teachers had data for only 1 year, class characteristics and district-by-grade-by-year fixed effects also are collinear with teacher fixed effects. Finally, and most importantly, we find that fixed and random effects specifications that condition on students' prior achievement and demographic characteristics return almost identical teacher effect estimates. When comparing teacher fixed effects with the "shrunken" empirical Bayes estimates that we use throughout the article, we find correlations between .79 and .99. As expected, the variance of the teacher fixed effects is larger than the variance of teacher random effects, differing by the shrinkage factor. When we instead calculate teacher random effects without shrinkage by averaging student residuals to the teacher level (i.e., "teacher average residuals"; see Guarino et al., 2015, for a discussion of this approach), they are almost identical to the teacher fixed effects estimates. Correlations are .99 or above across outcome measures, and unstandardized regression coefficients that retain the original scale of each measure range from 0.91 *SD* to 0.99 *SD*.

10. Adding prior survey responses to the education production function is not entirely analogous to doing so with prior achievement. Although achievement outcomes have roughly the same reference group across administrations, the surveys do not. This is because survey items often asked about students' experiences "in this class." All three Behavior in Class items and all five Happiness in Class items included this or similar language, as did five of the 10 items from Self-Efficacy in Math. That said, moderate year-to-year correlations of .39, .38, and .53 for Self-Efficacy in Math, Happiness in Class, and Behavior in Class, respectively, suggest that these items do serve as important controls. Comparatively, year-to-year correlations for the high- and low-stakes tests are .75 and .77.

11. To estimate these scores, we specified the following hierarchical linear model separately for each school year:

$$\text{OBSERVATION}_{lj,-t} = \gamma_j + \varepsilon_{ljt}$$

The outcome is the observation score for lesson l from teacher j in years other than t; γ_j is a random effect for each teacher, and ε_{ljt} is the residual. For each domain of teaching practice and school year, we utilized standardized estimates of the teacher-level residual as each teacher's observation score in that year. Thus, scores vary across time. In the main text, we refer to the teacher-level residual as $\widehat{\text{OBSERVATION}}_{lj,-t}$ rather than $\hat{\gamma}_j$ for ease of interpretation for readers.

12. One explanation for these findings is that the relationship between teachers' Classroom Organization and students' Happiness in Class is nonlinear. For example, it is possible that students' happiness increases as the class becomes more organized, but then begins to decrease in classrooms with an intensive focus on order and discipline. To explore this possibility, we first examined the scatterplot of the relationship between teachers' Classroom Organization and teachers' ability to improve students' Happiness in Class. Next, we reestimated Equation 2 including a quadratic, cubic, and quartic specification of teachers' Classroom Organization scores. In both sets of analyses, we found no evidence for a nonlinear relationship. Given our small sample size and limited statistical power, though, we suggest that this may be a focus of future research.

13. In similar analyses in a subset of the National Center for Teacher Effectiveness (NCTE) data, Blazar (2015) did find a statistically significant relationship between Ambitious Mathematics Instruction and the Low-Stakes math test of 0.11 *SD*. The 95% confidence interval (CI) around that point estimate overlaps with the 95% CI relating Ambitious Mathematics Instruction to the Low-Stakes math test in this analysis. Estimates of the relationship between the other three domains of teaching practice and Low-Stakes math test scores were of smaller magnitude and not statistically significant. Differences between the two studies likely emerge from the fact that we drew on a larger sample with an additional district and year of data, as well as slight modifications to our identification strategy.

14. When we adjusted *p* values for estimates presented in Table 5 to account for multiple hypothesis testing using both the Šidák and Bonferroni algorithms (Dunn, 1961; Šidák, 1967), relationships between Emotional Support and both Self-Efficacy in Math and Happiness in Class, as well as between Mathematical Errors and Self-Efficacy in Math remained statistically significant.

References

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*, 213–232.

Backes, B., & Hansen, M. (2015, October). *Teach for America impact estimates on nontested student outcomes* (Working Paper 146). Washington, DC: National Center for Analysis of Longitudinal in Education Research. Retrieved from http://www.caldercenter.org/sites/default/files/WP%20146.pdf

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Multifaceted impact of

self-efficacy beliefs on academic functioning. *Child Development*, *67*, 1206–1222.

Baron, J. (1982). Personality and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 308–351). New York, NY: Cambridge University Press.

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

Blazar, D., Braslow, D., Charalambous, C. Y., & Hill, H. C. (2015). *Attending to general and content-specific dimensions of teaching: Exploring factors across two observation instruments* (Working paper). Cambridge, MA: National Center for Teacher Effectiveness. Retrieved from http://scholar.harvard.edu/files/david_blazar/files/blazar_et_al_attending_to_general_and_content_specific_dimensions_of_teaching.pdf

Borghans, L., Duckworth, A. L., Heckman, J. J., & Ter Weel, B. (2008). The economics and psychology of personality traits. *Journal of Human Resources*, *43*, 972–1059.

Burchinal, M., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher-child interactions and instruction. *Applied Developmental Science*, *12*, 140–153.

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

Chan, T. C., & Jarman, D. (2004). Departmentalize elementary schools. *Principal*, *84*(1), 70–72.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics*, *126*, 1593–1660.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*, 2593–2632.

Chin, M., & Goldhaber, D. (2015). *Exploring explanations for the "weak" relationship between value added and observation-based measures of teacher performance* (Working paper). Cambridge, MA: National Center for Teacher Effectiveness. Retrieved from http://cepr.harvard.edu/files/cepr/files/sree2015_simulation_working_paper.pdf?m=1436541369

Cohen, D. K. (2011). *Teaching and its predicaments*. Cambridge, MA: Harvard University Press.

Corcoran, S. P., & Jennings, J. L. (2012). *Teacher effectiveness on high- and low-stakes tests*. Unpublished manuscript. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.269.5537&rep=rep1&type=pdf

Council of the Great City Schools. (2013). *Beating the odds: Analysis of student performance on state assessments, results from the 2012–2013 school year*. Washington, DC: Author.

Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. New York, NY: Teachers College Press.

Diener, E. (2000). Subjective well-being: The science of happiness and a proposal for a national index. *American Psychologist*, *55*, 34–43.

Downer, J. T., Rimm-Kaufman, S., & Pianta, R. C. (2007). How do classroom conditions and children's risk for school problems contribute to children's behavioral engagement in learning? *School Psychology Review*, *36*, 413–432.

Duckworth, A. (2016, March 26). Don't grade schools on grit. *The New York Times*. Retrieved from http://www.nytimes.com/2016/03/27/opinion/sunday/dont-grade-schools-on-grit.html

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*, 1087–1101.

Duckworth, A. L., Quinn, P. D., & Tsukayama, E. (2012). What no child left behind leaves behind: The roles of IQ and self-control in predicting standardized achievement test scores and report card grades. *Journal of Educational Psychology*, *104*, 439–451.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*, 237–251.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*, 52–64.

Epstein, D., & Miller, R. T. (2011). *Slow off the mark: Elementary school teachers and the crisis in science, technology, engineering, and math education*. Washington, DC: Center for American Progress.

The Every Student Succeeds Act, Public Law 114-95, 114th Cong., 1st sess. (2015, December 10). Retrieved from https://www.congress.gov/bill/114th-congress/senate-bill/1177/text

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching adolescents to become learners. The role of non-cognitive factors in shaping school performance: A critical literature review*. Chicago, IL: University of Chicago Consortium on Chicago School Reform.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). London, England: SAGE.

Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy*, *11*, 125–149.

Goodman, R. (2001). Psychometric properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*, 1337–1345.

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

Guarino, C. M., Maxfield, M., Reckase, M. D., Thompson, P. N., & Wooldridge, J. M. (2015). An evaluation of Empirical Bayes' estimation of value-added teacher performance measures. *Journal of Educational and Behavioral Statistics*, *40*, 190–222.

Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., & Pianta, R. C. (2015). Teaching through interactions in secondary school classrooms: Revisiting the factor structure and practical application of the Classroom Assessment Scoring System–Secondary. *The Journal of Early Adolescence*, *35*, 651–680.

Hamre, B., 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. (2001). Early teacher–child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development*, *72*, 625–638.

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

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*, 267–271.

Hickman, J. J., Fu, J., & Hill, H. C. (2012). *Technical report: Creation and dissemination of upper-elementary mathematics assessment modules*. Princeton, NJ: Educational Testing Service.

Hill, H. C., Blazar, D., & Lynch, K. (2015). Resources for teaching: Examining personal and institutional predictors of high-quality instruction. *AERA Open*, *1*(4), 1–23.

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

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

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

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal*, *105*, 11–30.

Hitt, C., Trivitt, J., & Cheng, A. (2016). When you say nothing at all: The predictive power of student effort on surveys. *Economics of Education Review*, *52*, 105–119.

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

Jackson, C. K. (2012, December). *Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina* (NBER Working Paper No. 18624). Cambridge, MA: National Bureau for Economic Research.

Jacob, B. A., & Lefgren, L. (2005). *Principals as agents: Subjective performance assessment in education* (NBER Working Paper No. 11463). Cambridge, MA: National Bureau for Economic Research.

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

Jennings, J. L., & DiPrete, T. A. (2010). Teacher effects on social and behavioral skills in early elementary school. *Sociology of Education*, *83*, 135–159.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). New York, NY: Guilford Press.

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: Measures of Effective Teaching Project, Bill and Melinda Gates Foundation.

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

King, R. B., McInerney, D. M., Ganotice, F. A., & Villarosa, J. B. (2015). Positive affect catalyzes academic engagement: Cross-sectional, longitudinal,

and experimental evidence. *Learning and Individual Differences*, *39*, 64–72.

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

Koedel, C. (2008). Teacher quality and dropout outcomes in a large, urban school district. *Journal of Urban Economics*, *64*, 560–572.

Kraft, M. A., & Grace, S. (2016). *Teaching for tomorrow's economy? Teacher effects on complex cognitive skills and social-emotional competencies* (Working paper). Providence, RI: Brown University. Retrieved from http://scholar.harvard.edu/files/mkraft/files/teaching_for_tomorrows_economy_-_final_public.pdf

Ladd, H. F., & Sorensen, L. C. (2015, December). *Returns to teacher experience: Student achievement and motivation in middle school* (Working Paper No. 112). Washington, DC: National Center for Analysis of Longitudinal in Education Research. Retrieved from http://www.caldercenter.org/sites/default/files/WP%20112%20Update_0.pdf

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

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*, 47–67.

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

Lynch, K., Chin, M., & Blazar, D. (2015). *Relationship between observations of elementary teacher mathematics instruction and student achievement: Exploring variability across districts* (Working paper). Cambridge, MA: National Center for Teacher Effectiveness.

Lyubomirsky, S., King, L., & Diener, E. (2005). The benefits of frequent positive affect: Does happiness lead to success? *Psychological Bulletin*, *131*, 803–855.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., . . . Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development*, *79*, 732–749.

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

Miles, S. B., & Stipek, D. (2006). Contemporaneous and longitudinal associations between social behavior and literacy achievement in a sample of low-income elementary school children. *Child Development*, *77*, 103–117.

Miller, C. C. (2015, October 16). Why what you learned in preschool is crucial at work. *The New York Times*. Retrieved from http://www.nytimes.com/2015/10/18/upshot/how-the-modern-workplace-has-become-more-like-preschool.html?_r=0

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., . . . Ross, S. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences*, *108*, 2693–2698.

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

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

National Governors Association Center for Best Practices. (2010). *Common core state standards for mathematics*. Washington, DC: Author.

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, *48*, 163–193.

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

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

Pianta, R. C., La Paro, K., Payne, C., Cox, M., & Bradley, R. (2002). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *Elementary School Journal*, *102*, 225–238.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: SAGE.

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

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

Ruzek, E. A., Domina, T., Conley, A. M., Duncan, G. J., & Karabenick, S. A. (2015). Using value-added models to measure teacher effects on students' motivation and achievement. *The Journal of Early Adolescence*, *35*, 852–882.

Segal, C. (2013). Misbehavior, education, and labor market outcomes. *Journal of the European Economic Association*, *11*, 743–779.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62*, 626–633.

Spearman, C. (1904). "General Intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*, 201–292.

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*, *38*, 293–317.

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

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*, F3–F33.

Tremblay, R. E., Masse, B., Perron, D., LeBlanc, M., Schwartzman, A. E., & Ledingham, J. E. (1992). Early disruptive behavior, poor school achievement, delinquent behavior, and delinquent personality: Longitudinal analyses. *Journal of Consulting and Clinical Psychology*, *60*, 64–72.

Tsukayama, E., Duckworth, A. L., & Kim, B. (2013). Domain-specific impulsivity in school-age children. *Developmental Science*, *16*, 879–893.

Usher, E. L., & Pajares, F. (2008). Sources of self-efficacy in school: Critical review of the literature and future directions. *Review of Educational Research*, *78*, 751–796.

West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F., & Gabrieli, J. D. (2016). Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis*, *38*, 148–170.

Whitehurst, G. J. (2016). *Hard thinking on soft skills*. Washington, DC: Brookings Institute. Retrieved from http://www.brookings.edu/research/reports/2016/03/24-hard-thinking-soft-skills-whitehurst

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations: Lessons learned in four districts*. Washington, DC: Brown Center on Education Policy at the Brookings Institute. Retrieved from http://www.brookings.edu/~/media/research/files/reports/2014/05/13-teacher-evaluation/evaluating-teachers-with-classroom-observations.pdf

Wigfield, A., & Meece, J. L. (1988). Math anxiety in elementary and secondary school students. *Journal of Educational Psychology*, *80*, 210–216.

Zernike, K. (2016, February 29). Testing for joy and grit? Schools nationwide push to measure students' emotional skills. *The New York Times*. Retrieved from http://www.nytimes.com/2016/03/01/us/testing-for-joy-and-grit-schools-nationwide-push-to-measure-students-emotional-skills.html?_r=0

Authors

DAVID BLAZAR is a postdoctoral research fellow at the Center for Education Policy Research at Harvard University. His research focuses on teacher and teaching quality, and the effect of policies aimed at improving both.

MATTHEW A. KRAFT is an assistant professor of education and economics at Brown University. He studies teacher effectiveness and organizational contexts in K–12 urban public schools. His recent work explores how teachers and schools develop students' social and emotional competencies.

Manuscript received November 18, 2015
First revision received March 28, 2016
Second revision received June 2, 2016
Third revision received August 18, 2016
Accepted August 23, 2016