WORKING PAPER 41

MATHEMATICA Policy Research

Working PAPER

BY DAVID BLAZAR (HARVARD GRADUATE SCHOOL OF EDUCATION AND 2015 MATHEMATICA SUMMER FELLOW) AND MATTHEW A. KRAFT (BROWN UNIVERSITY)

Teacher and Teaching Effects on Students' Academic Behaviors and Mindsets

December 2015

ABSTRACT

A growing body of evidence has identified a range of academic behaviors and mindsets other than test scores as important contributors to children's long-term success. We extend a complementary line of research focusing on the role that teachers play in developing these outcomes. We find that upper-elementary teachers have large effects on students' self-reported behavior in class, self-efficacy in math, and happiness in class that are similar in magnitude to effects on math test scores. However, teachers who are effective at improving these outcomes often are not the same as those who raise math test scores. We also find that these non-tested outcomes are predicted by teaching practices most proximal to these measures, including teachers' emotional support and classroom organization. Findings can inform policy around teacher development and evaluation.

Keywords: teacher effects, non-cognitive, behavior, self-efficacy, happiness, instruction

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. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Additional support came from the William T. Grant Foundation and Mathematica Policy Research's summer fellowship.

I. INTRODUCTION

A substantial body of evidence indicates that students' academic and lifelong success is a function of both their achievement on tests and a range of non-tested academic behaviors and mindsets (Borghans et al. 2008). 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 et al. 2012). Longitudinal studies document the strong predictive power of measures of motivation and childhood self-control on health and socioeconomic status in adulthood (Moffit et. al. 2011; Murayama et al. 2012). Economists and educators also have demonstrated the contribution of a range of non-tested outcomes, including disruptive class behaviors, emotional stability, and persistence, to educational achievement and labor market outcomes. In fact, these behaviors often are found to be stronger predictors of long-term outcomes than test scores (Chetty et al. 2011; Heckman and Rubinstein 2001; Lindqvist and Vestman 2011; Mueller and Plug 2006). Although these measures often are referred to as "non-cognitive" skills, our preference, like others (Duckworth and Yeager 2015; Farrington et al. 2012; Gehlbach 2015), is to refer to each competency by name. For brevity, we refer to them as "academic behaviors and mindsets" (Farrington et al. 2012) or "non-tested" outcomes to acknowledge the intellectual nature of many of these constructs.

The primary importance of academic behaviors and mindsets such as motivation, persistence, and self-control raises questions about how best to develop these outcomes in children. Given the substantial time that students spend in class, teachers may be in a particularly strong position to support their development in these areas. Indeed, several recent studies have found evidence that teachers can affect behaviors and mindsets beyond students' core academic knowledge and skills (Chetty et al. 2011; Jackson 2012; Jennings and DiPrete 2010; Koedel 2008; Ruzek et al. 2014). At the same time, important questions remain about the effect that teachers have on non-tested outcomes. To date, this line of research has focused on a relatively narrow set of measures, both within and across studies. Further, methodological challenges have made it difficult for prior research to separate teacher effects from broader class effects—that is, "the combined effect of teachers, peers, and any class-level shock" (Chetty et al. 2011, pp. 1596). Finally, very few studies have explored in a rigorous manner the specific characteristics of classrooms and teaching practices that produce these outcomes.

Better understanding the nature of how and when teachers affect students' academic behaviors and mindsets is critical for informing districts' and states' efforts to reform teacher development and evaluation systems. Currently, many education agencies reward teachers for improving test scores and executing instructional practices captured on observational instruments that have been shown to correlate with test score gains (Center on Great Teachers and Leaders 2013; Kane et al. 2013; Kane and Staiger 2012). However, it is not clear whether this approach also incentivizes teaching practices that support students' development in areas beyond their core academic skills. If teachers who are good at raising test scores do so without also affecting students' behavior, self-efficacy, or happiness, then current systems may seek to expand the measures they use to evaluate teachers. In particular, data linking specific teaching practices to non-tested outcomes would provide important validity evidence for observation instruments, many of which were designed to capture pedagogical practices thought to improve student behaviors and mindsets beyond test scores (Danielson 2011; Pianta and Hamre 2009).

In this study, we examine both teacher and teaching effects on a range of math test scores and non-tested outcomes. Throughout the paper, we refer to "teacher effects" to denote the unique effect of individual teachers on a range of student outcomes (for example, Chetty et al. 2011; Nye et al. 2004); similarly, we use the phrase "teaching effects" to refer to the relationship between specific classroom practices and these same outcomes. In the first part of our analyses, we estimate the effects of upper-elementary teachers on students' self-reported behavior in class, self-efficacy in math, and happiness in class, as well as absences from school. We then compare these effects with traditional teacher effects on both high- and low-stakes achievement tests in math. In the second part of our analyses, we examine whether certain dimensions of instruction help explain variation in these teacher effects by estimating the relationship between high quality teaching practices captured by two established observation instruments and our outcome measures.

We attempt to gain insight into the causal effect of both teacher and teaching effects on students' academic behaviors and mindsets by specifying education production function models that control for prior scores on our non-tested outcomes as well as students' prior achievement, and school fixed effects. Experimental and quasi-experimental studies suggest that this approach can minimize bias due to nonrandom sorting of students to teachers (Chetty et al. 2014; Kane et al. 2013; Kane et al. 2011). When estimating teaching effects, we specify models that also include out-of-year teaching quality scores and a range of other observable teacher characteristics. This method is advantageous because it eliminates bias that may be induced when students affect both current-year measures of teachers' practice and their own tested or non-tested outcomes. However, given the possibility of unmeasured teacher practices or characteristics biasing these models, we interpret our analyses of teaching effects conservatively as only suggestive evidence of any causal relationship.

Findings indicate that teachers have large impacts on students' non-tested outcomes. We estimate that the variation in teacher effects on students' behavior in class and self-efficacy in math is of similar magnitude to effects on math test scores. Teacher effects on students' happiness in class are even larger than those for test-based outcomes. However, although teachers impact both math test scores and non-tested outcomes, teacher effect estimates are only weakly correlated across outcome types (that is, effects on test scores versus non-tested outcomes). The largest of these unadjusted correlations is 0.19 between teacher effects on math test scores and effects on self-efficacy in math. This is notable, given the academic nature of both constructs and growing interest in measures of self-efficacy, persistence, and grit amongst educators and policymakers (Duckworth and Yeager 2015; Stecher and Hamilton 2014). Finally, a range of evidence suggests the importance of general elements of teachers' classroom instruction—namely, their interactions with and emotional support for students, as well as their classroom organization—in improving these outcomes. Errors in teachers' presentation of mathematical content also are negatively related to students' self-efficacy in math and happiness in class. Together, these findings point to specific teaching practices that may be a focus of development and evaluation efforts.

The remainder of this paper is organized as follows. In the second section, we discuss previous research on non-tested outcomes and the role of teachers in developing them. In the third section, we describe the data collected as part of this study. Fourth, we describe restrictions on our analytic sample. Fifth, we describe our analytic strategies for estimating teacher and

teaching effects on test scores and non-tested outcomes. Sixth, we provide our main results. We conclude by discussing the implications of our findings for ongoing research on teacher and teaching quality, and for policy around teacher development and evaluation.

II. BACKGROUND

A. The importance of non-tested outcomes

Recently, educators, researchers, and policymakers have called attention to students' non-tested outcomes to distinguish them from academic achievement on standardized tests. These outcomes encapsulate a range of factors, including belonging (Osterman 2000), conscientiousness (John and Srivastava 1999), grit and perseverance (Duckworth et al. 2007), growth mindset (Dweck 2006), happiness (Diener 2000), self-control (Tsukayama et al. 2013), self-esteem (Rosenberg 1989), and self-efficacy (Bandura 1977). Although each of these measures has been studied in depth on their own, many view these and other factors as interrelated. Together, they capture the complexity with which students develop academic knowledge and interact within their educational contexts (Farrington et al. 2012).

Researchers from a variety of disciplines have documented the centrality of many of these academic behaviors and mindsets to long-term life outcomes. For example, Heckman and Rubinstein (2001) identified the importance of non-tested outcomes to labor market success in an analysis of general equivalency diploma (GED) recipients versus high school dropouts. Drawing on interview data from the Swedish military, Lindqvist and Vestman (2011) also found that, conditional on cognitive ability, a composite measure of non-tested behaviors among 18 year-olds was positively related to wages and negatively related to unemployment. In a random sample of respondents in Wisconsin, Mueller and Plug (2006) found that each of the "Big Five" personality traits—openness, conscientiousness, extraversion, agreeableness, and neuroticism—predicted earnings, controlling for test scores. A review of research and meta-analysis by Borghans et al. (2008) came to similar conclusions. Academic behaviors and mindsets beyond test scores also hold long-term value outside of the labor market. Moffitt et al. (2011) followed a complete birth cohort in one city in New Zealand to age 32 and found that childhood measures of self-control predicted better physical health and lower substance dependence and criminal behaviors in adulthood. Results were similar when comparing outcomes within sibling pairs, suggesting that findings were not confounded with intelligence or social class.

B. The role of teachers in developing non-tested outcomes

Research highlighting the importance of a range of academic behaviors and mindsets has led to a discussion of how to develop them in children. In particular, this work has focused on the role of teachers. Psychologists Dweck et al. (2011) argued that "with greater awareness of non-cognitive factors, educators may be able to do relatively small things in classrooms that can make a big difference in their students' learning" (pp. 4–5). In their review of the literature, researchers at the Chicago Consortium on School Research concluded that "The essential question is not how to change students to improve their behavior but rather how to create [classroom] contexts that better support students in developing critical attitudes and learning strategies necessary for their academic success" (Farrington et al. 2012, p. 74).

Although teachers are obvious foci for developing these outcomes, only a handful of studies have examined these relationships empirically. Jennings and DiPrete (2010) used the Early Childhood Longitudinal Study—Kindergarten Cohort (ECLS-K) to estimate the role that teachers play in a composite measure of kindergarten and 1st-grade students' social and

behavioral outcomes. They found teacher effects on these outcomes that were even larger (0.35 standard deviations [sd]) than effects on academic achievement (0.20 sd and 0.30 sd for math and reading, respectively). In a study of 35 middle school math teachers, Ruzek et al. (2014) found small but meaningful teacher effects on motivation of between 0.03 sd and 0.08 sd among 7th graders. Additional studies have identified teacher effects on observed school behaviors, including absences, suspensions, grades, grade progression, and graduation (Gershenson forthcoming; Jackson 2012; Koedel 2008). With the exception of Gershenson (forthcoming) and Jackson (2012), however, these studies conflated teacher effects with class effects (that is, the combined effect of a teacher with the specific set of students in the classroom). Most examined only a narrow set of measures.

To date, evidence is mixed on the extent to which teachers who improve test scores also improve non-tested outcomes. Two of the studies described above found weak relationships between teacher effects on high-stakes standardized tests and other outcome measures. Compared to a correlation of 0.42 between teacher effects on math achievement versus teacher effects on reading achievement, Jennings and DiPrete (2010) found correlations of 0.15 between teacher effects on students' non-tested outcomes and teacher effects on either math or reading achievement. Jackson (2012) found that teacher effects in high school mathematics explained 5 percent or less of the variance in estimated teacher effects on suspensions, absences, and grade progression. Only 38 percent of teachers in the top quartile of teacher effects on test score outcomes were in the top quartile of teacher effects on these non-tested behaviors. Correlations from two other studies were larger. Ruzek et al. (2014) estimated a correlation of 0.50 between teacher effects on achievement versus effects on students' motivation in math class. Drawing on data from the Measures of Effective Teaching (MET) project, Mihaly et al. (2013) found a correlation of 0.57 between middle school teacher effects on students' self-reported effort versus effects on math test scores. At the same time, the correlation between middle school teacher effects on happiness in class and reading test scores was much lower, at 0.11.

Our analyses extended this body of research for additional non-tested outcomes captured by students in upper-elementary grades. We also were able to leverage data that offer the unique combination of a moderately sized sample of teachers and students with lagged measures of students' non-tested outcomes.

C. Teaching and non-tested outcomes

Although increased focus on the role of teachers in developing non-tested outcomes has been an important advance in the academic literature, to date, this line of research has yet to explore in a rigorous manner what teachers do in their classrooms to foster these behaviors in their students. Stemming from the "process-product" literature of the 1970s and 1980s, researchers have long hypothesized a relationship between teaching practice and student behaviors, focusing in particular on the extent to which students replicate behaviors first modeled by their teachers. For example, Willson (1973) found that students asked more complex questions when their teacher had been randomly assigned to a professional development program designed to develop more demanding questioning techniques. Despite the strong experimental design in this particular study, though, this literature provides little guidance, given the "unsystematic" and "methodologically unsophisticated" approaches of most other work (Centra and Potter 1980, p. 281).

More recent literature in this area suffers from similar limitations. For example, one meta-analysis described overall positive relationships between teachers' affective behavior and positive interactions with students, and students' observed engagement in class (Roorda et al. 2011). However, studies reviewed in this analysis generally were correlational in nature and thus did not take into account potential omitted variables that might have biased results. That is, teachers who have strong affective classroom practices might also engage in additional practices responsible for higher student outcomes.

Over the last several years, use of observation instruments has provided a unique opportunity to explore these relationships in more depth. In fact, many observation instruments focus on teaching behaviors identified because of theoretical links to social and emotional outcomes beyond test scores. For example, the Classroom Assessment Scoring System (CLASS) is organized around "meaningful patterns of [teacher] behavior (or behaviors) that are tied to underlying developmental processes [in students]" (Pianta and Hamre 2009, p. 112). With the goal of improving students' social and emotional functioning, one set of items captures the extent to which teachers create a positive environment conducive to learning. Additional items rate teachers on their classroom organization and are theoretically linked to students' own ability to self-regulate.

In this paper, we draw on data captured by two observation instruments to test these relationships. In particular, our work extends analyses conducted in the MET project, which identified moderate to large positive correlations (r = 0.33 to 0.57) between teacher scores on the Framework for Teaching (FFT)—an observation instrument also focused on teacher-student interactions—and their effectiveness in improving students' self-reported effort and happiness in class (Mihaly et al. 2013). Although these authors focused on a single composite score of teaching effectiveness from FFT, we extend our analyses to multiple subdomains thought to relate to each of our outcomes of interest. We also condition our estimates on a rich set of teacher covariates to minimize the threat due to omitted variables bias. We further compare these relationships to math-specific teaching effects on math test scores, which also have strong theoretical links (Hill et al. 2008; Lampert 2001; National Council of Teachers of Mathematics 1989, 1990, 2000).

III. DATA

Beginning in the 2010–2011 school year, the National Center for Teacher Effectiveness (NCTE) engaged in a three-year data collection process. Data came from participating 4th- and 5th-grade teachers (N = 310) in four anonymous medium or large districts that agreed to have their classes videotaped, complete a teacher questionnaire, and help collect a set of student outcomes. Although our study focuses on these teachers' math instruction, participants were generalists who taught all subject areas. Despite having a nonrandom sample, evidence from these same data indicates that teachers who participated in the study had similar value-added scores calculated from high-stakes standardized math tests as those who did not participate (Blazar 2015). We leverage this rich dataset to provide new insights into the relationship between teachers, teaching, and students' academic behaviors and mindsets. We describe each of these sources of data in turn, beginning with our outcomes of interest.

A. Non-tested outcomes

As part of the expansive data collection effort by NCTE, researchers collected a range of survey and administrate data that we use to construct non-tested outcomes. The measurement approach for these outcomes fell into two broad categories: (1) self-reported questionnaires and (2) observable behaviors, both of which have been shown to predict long-term outcomes (Diener 2000; Heckman and Rubinstein 2001; Jackson 2012; Mueller and Plug 2006).

We adapted self-report survey items (N = 18) from other large-scale surveys, including the TRIPOD survey project, the MET project, the National Assessment of Educational Progress (NAEP), and the Trends in International Mathematics and Science Study (TIMSS) (see Table A.1 for a full list of items). We rated all items on a five-point Likert scale where 1 = Totally Untrue and 5 = Totally True. We reverse coded items with negative valence to form composites with other items.

We utilized a combination of exploratory factor analyses and theory to construct a parsimonious set of measures from these 18 items. Exploratory factor analyses (see Table A.1) suggest that these items form two constructs.[1] The first consists of three items meant to hold together that we call *Behavior in Class* (internal consistency reliability is 0.74). Higher scores reflect better, less disruptive behavior. Teacher reports of students' classroom behavior have been found to relate to earnings, juvenile delinquency, and antisocial behaviors in adolescence (Chetty et al. 2011; Huesmann et al. 1984; Nagin and Tremblay 1999; Segal 2013; Tremblay et al. 1992). In turn, antisocial behavior predicts criminal behavior in adulthood (Loeber 1982; Moffitt 1993). Our analysis differs from these other studies in the self-reported nature of behavior outcomes, which is a limitation (Dunning et al. 2004). That said, the strong predictive

[1] We conducted factor analyses separately by year, given that there were fewer items in the first year. The NCTE project added additional items in subsequent years to help increase reliability. In the second and third years, each of the two factors 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 between 0.35 and 0.48—often taken as the minimum acceptable factor loading (Field 2013; Kline 1994) —this second factor explained roughly 20 percent more of the variation across teachers and thus had strong support for a substantively separate construct (Field 2013; Tabachnick and Fidell 2001). In the first year of the study, the eigenvalue on this second factor was less strong (0.78); the two items that load onto it also load onto the first factor.

power of in-school behavior cited above makes this an important construct to observe. Further, this measure is related to other outcomes in ways we would expect, with a positive and statistically significant correlation with test scores (r = 0.24 and 0.26 for high- and low-stakes tests, respectively) and a negative and statistically significant relationship to days absent (r = -0.13) (see Table III.1).

The second construct that emerged from the exploratory factor analyses consists of 15 total items whose face validity is less strong than the composite described above. For example, one item states, "Even when math is hard, I know I can learn it," and another states, "This math class is a happy place for me to be." Post hoc review of these items against the psychology literature from which they were derived suggests that they can be divided into two subdomains. *Self-Efficacy in Math* is a variation on well-known constructs related to students' effort, initiative, and perception that they can complete tasks (Bandura 1977; Duckworth et al. 2007; Schunk 1991). By asking students these items through the lens of mathematics, this construct is the most academic of our non-tested outcomes. However, moderate correlations to both high- and low-stakes math tests (r = 0.25 and 0.22, respectively) suggest that higher self-efficacy in math does not necessarily reflect mastery of the content (see Table III.1). The second sub-domain is *Happiness in Class*. As above, this measure is a school-specific version of well-known scales that capture affect and enjoyment (Diener 2000).[2] Both of these constructs have reasonably high internal consistency reliabilities (0.76 and 0.82, respectively), similar to those of non-tested behaviors explored in other studies (Duckworth et al. 2007; John and Srivastava 1999; Tsukayama et al. 2013). Further, self-reported measures of both of these broad constructs have been linked to long-term outcomes, including academic engagement, productivity, and earnings in adulthood, even conditioning on cognitive ability (King, McInerney, Ganotice, and Villarosa 2015; Lyubomirsky et al. 2005; Mueller and Plug 2006; Oswald, Proto, and Sgroi forthcoming).

For all non-tested outcomes, we created final scales by averaging raw student responses across all available items and standardizing measures to have a mean of zero and a standard deviation of one within each school year. We standardized within years, given that, for some measures, the set of survey items vary across years.

Studies have shown that some self-reported measures can be prone to reference bias and social desirability bias (Duckworth and Yeager 2015). For example, West and his colleagues (forthcoming) provide evidence that effective charter schools with explicit schoolwide norms around behavior and effort change the implicit standards of comparison that students use to judge their conscientious, grit, and self-control, but not their growth mindset. We attempted to minimize the potential threat posed by these possible biases by restricting our comparisons to teachers and students in the same school. This approach, which we describe in more detail below, helps to limit potential difference in reference groups and social norms across schools and districts that could confound our analyses.

[2] The *Happiness in Class* scale was not available for the first year of the study.

Table III.1. Descriptive statistics for test scores and non-tested outcomes

	Univariate statistics			Pairwise correlations					
	Mean	Standard deviation	Internal consistency reliability	High-stakes math test	Low-stakes math test	Behavior in class	Self-efficacy in math	Happiness in class	Days absent
High-stakes math test	0.10	0.91	0.90	1.00					
Low-stakes math test	0.61	1.10	0.82	0.70***	1.00				
Behavior in class	4.10	0.93	0.74	0.24***	0.26***	1.00			
Self-efficacy in math	4.17	0.58	0.76	0.25***	0.22***	0.35***	1.00		
Happiness in class	4.10	0.85	0.82	0.15***	0.10***	0.27***	0.62***	1.00	
Days absent	5.6	6.1	NA	-0.18***	-0.21***	-0.13***	-0.08***	-0.08***	1.00

Notes: For high-stakes math test, reliability varies by district. Behavior in class, self-efficacy in math, and happiness in class measured on Likert scale from 1 to 5. Statistics generated from all available data.

***p<0.001.

Finally, we constructed a measure of attendance, *Days Absent*, which is a count of total days that students were absent from school, collected from administrative records.[3] At the high school level, absences have been shown to predict school dropout, substance abuse, violence, and juvenile delinquency (Bell et al. 1994; Hawkins et al. 1998; Loeber and Farrington 2000; Robins and Ratcliff 1980). At the elementary level, as in our data, student absences may be a closer proxy for parental behaviors than student ones. Nevertheless, in Table III.1, we show that absences are significantly negatively correlated with all three survey constructs (r = -0.08 to -0.13), thus suggesting that absences can serve as an objective measure related to our self-reported measures of interest.

B. Mathematics lessons

As described by Blazar (2015), we captured teachers' mathematics lessons over a three-year period, with a maximum of three lessons per teacher per year. Capture occurred with a three-camera digital recording device and lasted between 45 and 60 minutes.[4] Trained raters scored

[3] Attendance data sometimes is disaggregated by excused versus unexcused absences to acknowledge differences in student behaviors for those who skip school versus those who are sick. However, the elementary students in our data were much less likely to skip school than middle or high school students. Further, districts take different approaches to reporting student absences. Thus, we focused on a single absence variable. A log transformation of these data—which has been used in similar work (Jackson 2012)—does not change trends in results. Attendance data were not available in District 2.

[4] As described by Blazar (2015), NCTE allowed teachers to choose the dates for capture in advance and directed them to select typical lessons and exclude days on which students were taking a test. Although it is possible that these lessons were unique from teachers' general instruction, teachers did not have any incentive to select lessons strategically, as no rewards or sanctions were involved with data collection. In addition, analyses from the MET

these lessons on two established observational instruments—the CLASS and the Mathematical Quality of Instruction (MQI). Analyses of these same data (Blazar et al. 2015) show that the items cluster into four main factors. The two dimensions from the CLASS instrument captured general teaching practices. *Emotional Support* focuses on teachers' interactions with students and the emotional environment in the classroom, and is thought to increase students' academic engagement and motivation, self-reliance, and ability to take risks. *Classroom Organization* focuses on behavior management and the productivity of the lesson, and is thought to improve students' self-regulatory behaviors (Pianta and Hamre 2009). The two dimensions from the MQI captured 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 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. Both dimensions are linked to teachers' mathematical knowledge for teaching and, in turn, to students' mathematics achievement (Hill et al. 2004, 2008).

We estimated reliability for these metrics by calculating the amount of variance in teacher scores attributable to the teacher—that is, the intraclass correlation [ICC]), adjusted for the modal number of lessons. These estimates were 0.53, 0.63, 0.74, and 0.56 for *Emotional Support, Classroom Organization, Ambitious Mathematics Instruction,* and *Mathematical Errors*, respectively (see Table III.2). Though some of these estimates were lower than conventionally acceptable levels (0.7), they were consistent with those generated from similar studies (Bell et al. 2012; Kane and Staiger 2012). Correlations between dimensions ranged from roughly 0 (between *Emotional Support* and *Mathematical Errors*) to 0.46 (between *Emotional Support* and *Classroom Organization*). (See Blazar et al. 2015 for further information on these instruments and dimensions.)

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. To avoid this source of bias, we utilized all lessons captured in years other than those in which student outcomes were measured. Then, to minimize noise in these observational measures, we utilized empirical Bayes estimation to shrink teacher-by-year scores back to the mean, based on their precision.[5] We

project indicate that teachers are ranked almost identically when they choose lessons themselves compared to when lessons are chosen for them (Ho and Kane 2013).

[5] Here, precision is a function of the number of lessons they provided to the study—between three and six, depending on the number of years each teacher participated in the project. The minimum number corresponds to recommendations by Hill et al. (2012) to achieve sufficiently high levels of predictive reliability. To estimate these scores, we specified the following hierarchical linear model separately for each school year:

$$OBSERVATION_{lj}^{-t} = \mu_j + \varepsilon_{ljt}$$

The outcome is the observation score for lesson *l* from teacher *j* in years other than *t*; μ_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 varied across time.

standardized final scores within the full sample of teachers to have a mean of zero and a standard deviation of one.

Table III.2. Descriptive statistics for CLASS and MQI dimensions

	Univariate statistics			Pairwise correlations			
	Mean	Standard deviation	Adjusted intraclass correlation	Emotional support	Classroom organization	Ambitious mathematics instruction	Mathematical errors
Emotional support	4.28	0.48	0.53	1.00			
Classroom organization	6.41	0.39	0.63	0.46***	1.00		
Ambitious mathematics instruction	1.27	0.11	0.74	0.22***	0.23***	1.00	
Mathematical errors	1.12	0.09	0.56	0.01	0.09	-0.27***	1.00

Notes: Intraclass correlations are adjusted for the model number of lessons. CLASS items (from emotional support and classroom organization) on a scale from 1 to 7. MQI items (from ambitious mathematics instruction and mathematical errors) on a scale from 1 to 3. Statistics generated from all available data.

***p<0.001.

C. Teacher characteristics

Information on teachers' background and knowledge were captured on a questionnaire administered by NCTE in the fall of each year. Survey items included years teaching math, which we collapsed to an indicator for whether teachers had three or fewer years of experience;[6] route to certification; and amount of undergraduate or graduate coursework in math and math courses for teaching (scored on a Likert scale from 1 to 4). For simplicity, we averaged these last two items to form one construct capturing teachers' mathematics coursework. Further, the survey included a test of teachers' mathematical content knowledge, with items from both the Mathematical Knowledge for Teaching assessment (Hill et al. 2004), which captures math-specific pedagogical knowledge, and the Massachusetts Test for Educator Licensure. Teacher scores were generated using IRTPro software and standardized them in these models, with a reliability of 0.92. (For more information about these constructs, see Hill et al. forthcoming.)

D. Additional student information

Student demographic and achievement data came from district administrative records. Demographic data included gender, race/ethnicity, free- or reduced-price lunch (FRPL) eligibility, limited English proficiency (LEP) status, and special education (SPED) status. These records also included current- and prior-year test scores in math and English Language Arts (ELA) on state assessments, which are standardized within a district by grade, subject, and year, using the entire sample of students in each district, grade, subject, and year.

Finally, the project administered a low-stakes mathematics assessment to all students in the study. We used this assessment in addition to high-stakes tests data, given evidence that teachers

[6] We bucketed teachers in this way, rather than using a more standard division between novices and veterans, as fewer than 15 teachers in our sample were in their first year in the classroom. A linear or quadratic specification of experience does not change results.

can be ranked differently depending on the specific test score measure utilized (Corcoran et al. 2012; Lockwood et al. 2007; Papay 2011). Validity evidence indicates an internal consistency reliability of 0.82 or higher for each form across grade levels and school years (Hickman et al. 2012). The low-stakes test is similar in content coverage to high-stakes math tests in all districts (Lynch et al. 2015). However, in addition to being low stakes, it differs from some of the high-stakes tests in its format (combining multiple-choice and short-response items) and cognitive demand (asking students to explore patterns rather than solving basic procedures). (See Lynch et al. 2015for an in-depth description of this low-stakes assessment and the high-stakes district tests.)

IV. SAMPLE RESTRICTIONS

In choosing our analysis sample, we faced a trade-off between precision and internal validity. Including all possible teachers would maximize the precision of our estimates. At the same time, we lacked data from some teachers' students—namely prior measures of our non-tested outcomes—that could have been used to guard against potential sources of bias. Thus, we chose to make two important restrictions to this original sample of teachers to strengthen the internal validity of our findings. First, for all analyses predicting non-tested outcomes, we only included 5th-grade teachers who happened to have students who also had been part of the project in the 4th grade, and therefore had prior-year scores for our non-tested outcomes. This group included between 51 and 111 teachers. For analyses predicting test score outcomes, we were able to maintain the full sample of 310 teachers, whose students all had test scores in the previous year. Second, in analyses relating domains of teaching practice to student outcomes, we further restricted our sample to teachers who themselves had been part of the study for more than one 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 when predicting non-tested outcomes, and 196 when predicting test scores.

In Table IV.1, we present descriptive statistics on teachers and their students in the full sample (column 1), as well as teachers and students who were ever in any of our non-tested outcomes samples (column 2). We find that teachers look roughly similar across these two analytic samples, with no statistically significant differences on any observable characteristics.[7] Sixteen percent of teachers were male, and 65 percent were white. Eight percent received their teaching certification through an alternative pathway. The average number of years of teaching experience was roughly 10. Value-added scores on the state math test were right around the mean for each district (0.01 sd).

[7] Descriptive statistics and formal comparisons of other samples show similar patterns and are available upon request.

Table IV.1. Participant demographics

	Full sample	Non-tested outcomes sample	*P*-value on difference
Teachers			
Male	0.16	0.16	0.949
African American	0.22	0.22	0.972
Asian	0.03	0.00	0.087
Hispanic	0.03	0.03	0.904
White	0.65	0.66	0.829
Mathematics coursework (1 to 4 Likert scale)	2.58	2.55	0.697
Mathematical content knowledge (standardized scale)	0.01	0.03	0.859
Alternative certification	0.08	0.08	0.884
Teaching experience (years)	10.29	10.61	0.677
Value added on high-stakes math test (standardized scale)	0.01	0.00	0.505
Observations	**310**	**111**	
Students			
Male	0.50	0.49	0.371
African American	0.40	0.40	0.421
Asian	0.08	0.07	0.640
Hispanic	0.23	0.20	0.003
White	0.24	0.28	<0.001
FRPL	0.64	0.59	0.000
SPED	0.11	0.09	0.008
LEP	0.20	0.14	<0.001
Prior score on high-stakes math test (standardized scale)	0.10	0.18	<0.001
Prior score on high-stakes ELA test (standardized scale)	0.09	0.20	<0.001
Observations	**10,575**	**1,529**	

We do observe some statistically significant differences between student characteristics in the full sample versus the subsample. For example, the percentage of students identified as limited English proficient was 20 percent in the full sample, compared to 14 percent in the sample of students who ever had been part of the analyses, drawing on our non-tested outcomes. Prior achievement scores were 0.10 sd and 0.09 sd in math and ELA in the full sample, respectively, compared to 0.18 sd and 0.20 sd in the subsample. 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.

V. EMPIRICAL STRATEGY

A. Estimating teacher effects

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

$$(1) \quad OUTCOME_{idsgjct} = \alpha f(A_{it-1}) + \pi X_{it} + \varphi \bar{X}_{it}^{c} + \tau_{dgt} + (\mu_j + \delta_{jc} + \varepsilon_{idsgjct})$$

$OUTCOME_{idsgjct}$ is used interchangeably for both math test scores and non-tested outcomes, which are 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 scores at this level. The error structure consists of both teacher- and class-level random effects, μ_j and δ_{jc}, respectively, and a student-specific error term, $\varepsilon_{idsgjct}$. Given our focus on elementary teachers, over 97 percent of teachers in our sample worked with just one set of students in a given year. Thus, class effects are estimated by observing teachers in multiple years and are analogous to teacher-by-year effects.[9]

The key identifying assumption of this model is that estimates are not biased by non-random sorting of students to teachers (Rothstein 2010). Recent experimental (Kane et al. 2013; Kane and Staiger 2008) 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 are sorted to teachers based on their in-class behavior in ways unrelated to their prior achievement. To address this possibility, we made two modifications to equation (2). First, we included school fixed effects, σ_s, to account for sorting of students and teachers across schools.[10] By restricting comparisons to those observed within schools, we also minimized the possibility of reference bias in our self-reported measures (Duckworth and Yeager 2015; West et al. forthcoming). Second, for models that predict each of our four non-tested outcomes, we included $OUTCOME_{it-1}$ on the right-hand side of the equation in addition to prior

[8] We controlled for prior-year scores only on the high-stakes assessment for two reasons. First, including previous low-stakes test scores would reduce our full sample by more than 2,200 students because the assessment was not given to students in District 4 in the first year of the study (N = 1,826 students). Further, an additional 413 students were missing fall test scores because they were not in class on the day it was administered. Second, prior-year scores on the high- and low-stakes test were correlated at 0.71, suggesting that including both would not help to explain substantively more variation in our outcomes.

[9] One exception was for *Happiness in Class*, where, in the smaller sample, each teacher only had one class; here, we exclude class random effects.

[10] A related concern may be that, within schools, teachers are sorted across grades due to a particularly strong or weak incoming class. This sorting would not be accounted for when using school fixed effects. Drawing on our full sample of teachers, though, we find that those who switched grades from one year to the next did not differ on their value-added in mathematics or instructional quality from those who did not switch (see Table A.2).

achievement—that is, when predicting students' *Behavior in Class*, we controlled for students' self-reported *Behavior in Class* in the prior year.[11]

Using equation (1), we conducted two sets of analyses. First, we estimated the variance of μ_j, which is the stable component of teacher effects, and reported the standard deviation of these estimates across outcomes. This parameter captures the magnitude of the variability of teacher effects. In our main analysis, we included but did not interpret δ_{jc} to separate out the time-varying portion of teacher effects, combined with peer effects and any other class-level shocks. In separate models, we excluded δ_{jc} to determine the extent to which teacher effects might be biased upward in analyses that conflate teacher and class effects. Because μ_j is measured imprecisely, given typical class sizes, unadjusted estimates would overstate the true variation in teacher effects. Thus, we utilized empirical Bayes estimation to shrink each score for teacher *j* back toward the mean based on its precision (Raudenbush and Bryk 2002), where precision is a function of the number of students attributed to each teacher or class. Like others interested in the variance of teacher effects (for example, Chetty et al. 2011), we specified this parameter as a random effect, which provides unbiased model-based estimates of the true population variance of teacher effects.[12] We also generated an estimate of the precision of each teacher effect variance estimate by calculating the signal-to-noise ratio.[13]

In our second analysis drawing on equation (2), we estimated μ_j for each outcome and teacher *j*, and then generated a correlation matrix of these teacher effects. For consistency, we continued to specify this parameter as a random effect rather than as fixed effects. 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). Recognizing this concern, we focused our analyses and discussion

[11] It is important to note that adding prior non-tested outcomes to the education production function is not entirely analogous to doing so with prior achievement scores. Whereas achievement outcomes have roughly the same reference group across administrations, the surveys do not because survey items often ask about students' experiences "in this class." All three *Behavior in Class* items and all five *Happiness in Class* items include this or similar language, as do five of the 10 items from *Self-Efficacy in Math*. That said, moderate year-to-year correlations of 0.53, 0.39, and 0.38 for *Behavior in Class*, *Self-Efficacy in Math*, and *Happiness in Class*, respectively, suggest that these items do serve as important controls. Comparatively, year-to-year correlations for the high-stakes test, the low-stakes test, and days absent are 0.75, 0.77, and 0.66, respectively.

[12] We estimated these variance components using restricted maximum likelihood estimation because full maximum likelihood estimates tend to be biased downward (Harville 1977; Raudenbush and Bryk 2002) and may be particularly problematic in our smaller subsample of students and teachers who had prior-year non-tested outcomes.

[13] We estimated the signal-to-noise ratio by calculating:

$$\frac{Var(\mu_j)}{Var(\mu_j) + \left(\frac{\sum_{j=1}^{n} se_j^{\,2}}{n}\right)}$$

The numerator is the observed variance in the teacher effect, or the squared value of the standard deviation of μ_j, which is our main parameter of interest. The denominator is an estimate of the true teacher-level variance, which we approximate as the sum of the observed variance in the teacher effect and the average squared error of the teacher effect. The number of teachers in the sample is detonated by *n*, and se_j is the standard error of the teacher effect for teacher *j*. See McCaffrey et al. (2009) for a similar approach.

on unadjusted correlations given that relative magnitude of the relationships between teacher effects across outcomes were largely unaffected by adjustments that account for attenuation.

B. Estimating teaching effects

Next, we examined whether certain types of instruction and teaching practices explain why some teachers have larger effects than others by estimating the relationship between high quality instruction and student outcomes, using equation (2):

$$(2) \quad OUTCOME_{idsgjct} = \beta \widehat{OBSERVATION}_j^{-t} + \theta T_j + \alpha f(A_{it-1}) + \pi X_{it} + \varphi \bar{X}_{it}^{c} + \sigma_s + \tau_{dgt} + (\mu_j + \delta_{jc} + \varepsilon_{idsgjct})$$

As above, we modeled each math test score or non-tested outcome for student *i* in district *d,* school *s,* and grade *g* with teacher *j* in class *c* at time *t* as a function of prior achievement, demographic characteristics, peer effects, school fixed effects, and district-by-grade-by-year fixed effects. In models that predict non-tested outcomes, we also controlled for prior scores, $OUTCOME_{it-1}$. We further included a vector of their teacher *j*'s observation scores, $\widehat{OBSERVATION}_j^{-t}$, in years other than *t*; these scores are predicted estimates. The coefficients on these variables are our main parameters of interest and can be interpreted as the standard deviation increase in each outcome associated with exposure to teaching practice one standard deviation above the mean.[14]

In this model, an additional concern for identification is the endogeneity of observed classroom quality. Randomly assigning teaching practice to teachers within their own classrooms raises a number of practical concerns; it was not possible in this study and likely would be challenging in any other study. Instead, our analytic approach attempted to account for potential sources of bias through T_j, a vector of observable teacher characteristics other than instructional quality. Including these characteristics helps isolate the relationship between teaching practice and student outcomes from others related both to the observation scores and student outcomes included in our data. Although we had access to an array of teacher characteristics, we focused on those that previous research suggests could induce bias in this type of analysis if omitted: mathematics coursework, mathematical content knowledge, alternative certification, and an indicator as to whether the teacher is a novice (that is, in his/her third year of teaching or fewer).[15] Given that we were not able to isolate teaching quality from all possible teacher

[14] Models were fit using full maximum likelihood, given our focus in this analysis on the fixed rather than the stochastic portion of the model; full maximum likelihood allowed use to compare estimates from the fixed portion of the equation between nested models (Harville 1977; Raudenbush and Bryk 2002).

[15] Review of previous research indicates that several observable characteristics are related both to student outcomes and instructional quality. Because current research has not yet explored the relationship between teacher characteristics and specific non-tested outcomes, we drew on evidence from studies that examine test score outcomes. As described by Blazar (2015), these studies indicated that students learn more mathematics from teachers with previous coursework in the content area (Wayne and Youngs 2003); stronger test scores (Metzler and Woessmann 2012); some forms of alternative certification, such as Teach for America (Clark et al. 2013; Decker et al. 2004); and more experience in the classroom (Chetty et al. 2011). These factors also appear to predict some dimensions of instruction in an analysis of the same data used in this study (Hill et al. forthcoming). Notably, a range of other characteristics sometimes thought to predict instructional quality, including specialized certifications, were not in fact related. Analyses of these same data (Blazar 2015; Hill et al. forthcoming) indicated that inclusion

characteristics, we consider this approach as providing suggestive rather than conclusive evidence on the underlying causal relationship between teaching practices and non-tested outcomes.

of these four characteristics limits some of the variation in instructional quality but these relationships are not overly strong, such that multicollinearity would be a concern.

VI. RESULTS

A. Teacher effects on test scores and non-tested outcomes

We begin by presenting results of the magnitude of teacher effects in Table VI.1. In Model 1, we estimate teacher effects separately from class effects, as specified in equation 2. Model 2 excludes class effects, providing an upward-bound estimate of teacher effects comparable to other studies that only observe one class per teacher. As expected, we find that teacher effects from Model 2, which exclude class effects, are between 13 to 36 percent larger in magnitude than effects from Model 1, which includes these class effects. This suggests that analyses that do not take into account classroom-level shocks likely produce upwardly biased estimates of stable teacher effects. For this reason, we focus our discussion on the latter rather than the former.

Table VI.1. Teacher effects on test scores and non-tested outcomes

	Observations		Model 1:Includes class effects	Model 2: Excludes class effects
	Teachers	Students		
High-stakes math test	310	10,575	0.18	0.21
Low-stakes math test	310	10,575	0.17	0.20
Behavior in class	111	1,529	0.15	0.17
Self-efficacy in math	108	1,433	0.14	0.19
Happiness in class	51	548	--	0.31
Days absent	86	1,076	0.00	0.02

Across both sets of models, we find comparable teacher effects on both test scores and non-tested outcomes. Consistent with a large body of literature (Hanushek and Rivkin 2010), in our preferred models that include both teacher and class effects, a one standard deviation difference in teacher effectiveness was equivalent to a 0.17 sd or 0.18 sd difference in students' academic achievement in math (high- and low-stakes math test, respectively). In other words, teachers at the 84th percentile of the distribution of teacher effectiveness moved the medium student up to roughly the 57th percentile of math achievement. Estimated teacher effects on students' self-reported *Behavior in Class* and *Self-Efficacy in Math* were 0.15 sd and 0.14 sd, respectively. The largest teacher effects we observe were on students' *Happiness in Class*, at 0.31 sd. This estimate comes from Model 2, given that we do not have multiple years of data to separate out class effects for this measure. Thus, we interpret this estimate as the upward bound of true teacher effects on *Happiness in Class.* Rescaling this by the ratio of teacher effects with and without class effects for *Self-Efficacy in Math* (0.14/0.19 = 0.74) produces an estimate of stable teacher effects on *Happiness in Class* of 0.23 sd, still larger than effects for other outcomes. We do not find evidence of any teacher effects on *Days Absent*. Though not reported in Table VI.1, we do observe a substantive and statistically significant class effect on this outcome of 0.06 standard deviations. However, as our data do not allow us to parse possible explanations for this effect—for example, teacher effects that vary over time, peer effects among classmates, or contagion effects from the flu—we do not interpret this estimate substantively. We excluded *Days Absent* from all subsequent analyses.

Next, we present estimates of signal-to-noise ratios to ensure that our variance estimates across outcomes are not driven by differences in reliability. We find that estimates of teachers' effectiveness are measured with broadly similar precision for non-tested outcomes as for test scores (see Table VI.2). To ensure that differences in reliability across outcome measures are not driven by sample sizes, we calculated these estimates of precision both in the original samples used to estimate effects shown in Table VI.1 (column 1) and then in a balanced sample of teachers and students who had complete data on all measures (column 2; N = 51 teachers and 548 students, the same sample for teacher effects on students' *Happiness in Class*). Focusing on this common sample, we find that precision ranged from 0.50 (for *Self-Efficacy in Math*) to 0.56 (for *Happiness in Class*). Estimates of precision for the high-stakes math test (0.54) and the low-stakes math test (0.50) fall within this narrow range. For outcomes in which we were able to include more teachers and students (all except *Happiness in Class*), the actual precision of our estimates presented in Table VI.1 is even greater. For example, for the high- and low-stakes math tests, which both draw on the full sample of 310 teachers, precision rises to 0.67 and 0.64, respectively.

Table VI.2. Signal-to-noise ratio of teacher effect estimates

	Original sample	Common sample
High-stakes math test	0.67	0.54
Low-stakes math test	0.64	0.50
Behavior in class	0.55	0.52
Self-efficacy	0.53	0.50
Happiness in class	0.56	0.56

Notes: See Table VI.1 for sample sizes across outcomes in the original samples. The common sample includes 51 teachers and 548 students.

Examining the correlations of teacher effect estimates reveals that individual teachers varied considerably in their ability to impact different students' outcomes (see Table VI.3). The fact that teacher effects are measured with error makes it difficult to interpret actual magnitudes of these relationships (Spearman, 1904). However, even a conservative adjustment that scales correlations by the inverse of the square root of the reliability leads to a similar overall pattern of results, particularly given our primary focus on comparing the relative magnitude of correlations across measures.[16] Unsurprisingly, we find the strongest correlations between teacher effects within outcome type. Specifically, we estimate a correlation of 0.64 between teacher effects on the two math achievement tests and a correlation of 0.49 between teacher effects on *Behavior in Class* and effects on *Self-Efficacy in Math*. Comparatively, 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 = 0.19). Importantly, the 95 percent confidence interval around the correlation between teacher effects on the two achievement measures [0.56, 0.72] does not

[16] We still observe much stronger relationships between teacher effects on the two math tests and between teacher effects on *Behavior in Class* and *Self-Efficacy in Math* than between other outcome measures. In some cases, these disattenuated correlations are close to 1, which we argue are unlikely to be the true relationships in the population. Overcorrections likely are driven by moderate reliabilities and moderate sample sizes (Zimmerman and Williams, 1997).

overlap with the 95 percent confidence interval of the correlation between teacher effects on the low-stakes math test and effects on *Self-Efficacy in Math* [-0.01, 0.39]. 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 non-tested outcomes. We caution against placing too much emphasis on the negative correlations between teacher effects on test scores and effects on *Happiness in Class* (r = -0.09 for the high-stakes test and -0.21 for the low-stakes test). Given limited precision of this relationship, we cannot rule out weak, positive or negative correlations among these measures. Interestingly, we also have some suggestive evidence that teachers may not be equally effective at raising different types of non-tested outcomes. We find relatively weak correlations between teacher effects on *Happiness in Class* with effects on *Behavior in Class* (0.21), as well as between this former construct and effects on *Self-Efficacy in Math* (0.26). While these correlations have large confidence intervals, we are able to distinguish them from the correlation between teacher effects on the two math tests.

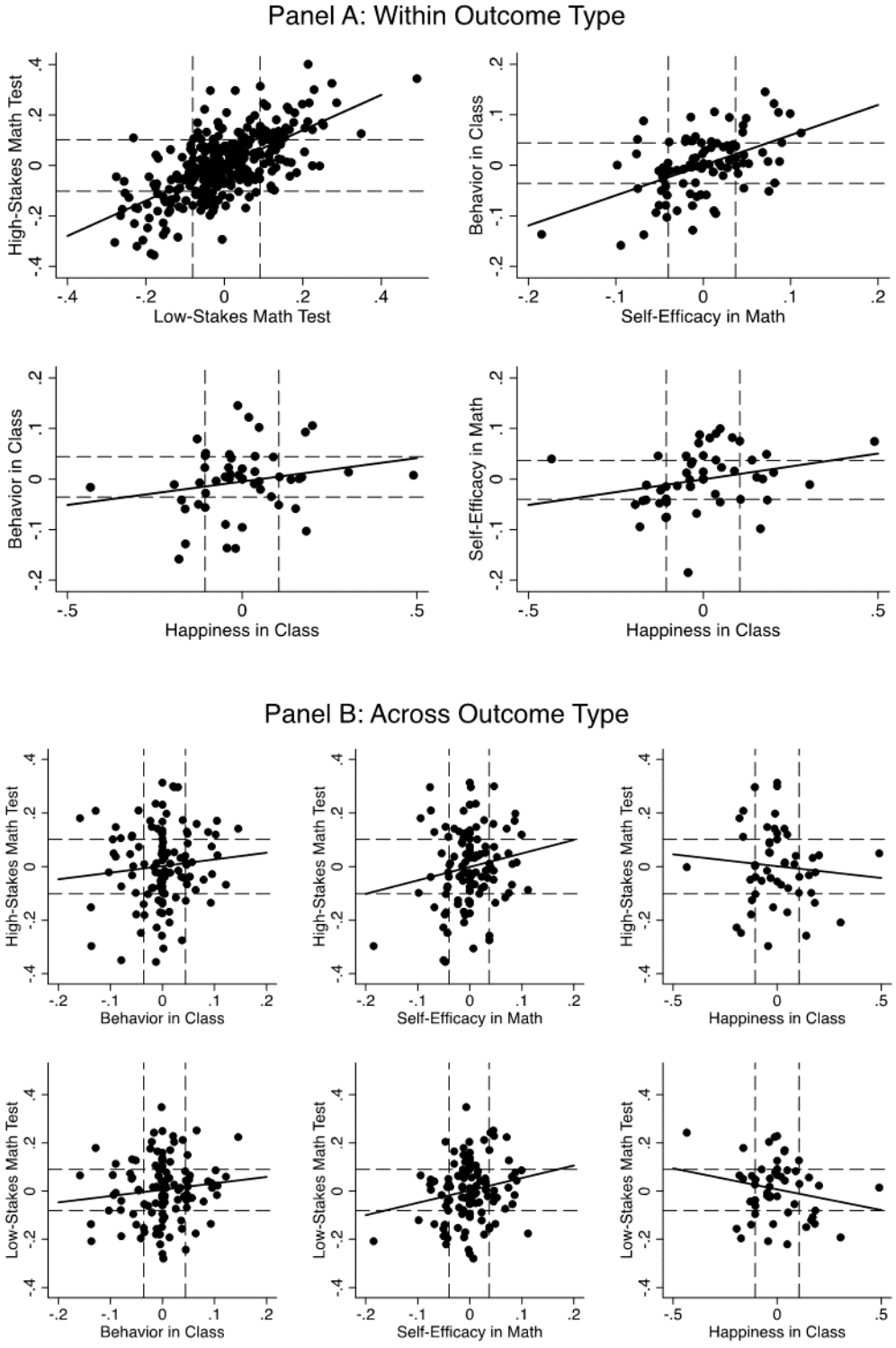
Table VI.3. Correlations between teacher effects on test scores and non-tested outcomes

	High-stakes math test	Low-stakes math test	Behavior in class	Self-efficacy in math	Happiness in class
High-stakes math test	1.00 --				
Low-stakes math test	0.64*** (0.04)	1.00 --			
Behavior in class	0.10 (0.10)	0.12 (0.10)	1.00 --		
Self-efficacy in math	0.16~ (0.10)	0.19* (0.10)	0.49*** (0.08)	1.00 --	
Happiness in class	-0.09 (0.14)	-0.21 (0.14)	0.21~ (0.14)	0.26~ (0.14)	1.00 --

Notes: ~ p < 0.10, * p < 0.05, ***p < 0.001. Standard errors in parentheses. See Table VI.1 for sample sizes used to calculate teacher effect estimates.

In Figure VI.1, we present scatter plots of these relationships to further illustrate the substantial variation in teacher effects across outcomes. In Panel A, we present relationships between teacher effects within outcome type (that is, effects on two different test scores or two types of non-tested outcomes). In Panel B, we depict relationships between teacher

Figure VI.1. Relationships between teacher effects



Notes: The figure shows scatter plots of teacher effects across outcomes. Solid lines represent the best-fit regression line. Vertical and horizontal dashed lines represent the 20th and 80th percentiles for each outcome.

effects across outcome type. In addition to showing best-fit lines, whose slope is analogous to the correlations presented above, we include dashed lines identifying the 20th and 80th percentile (that is, the divide between the first and second quintile and between the fourth and fifth) of each outcome. These thresholds align with how teacher effect estimates have been used in practice to make high-stakes decisions about tenure and compensation (Center on Great Teachers and Leaders 2013; Dee and Wyckoff 2015; Loeb et al. 2015), and illustrate how teachers move across rankings depending on the outcome measure used. In the bottom panel in particular, many teachers lie off the diagonal, indicating very weak relationships between teacher effects on test scores versus effects on non-tested outcomes. Even for the two outcome measures in this bottom panel in which we observe the strongest correlation (the low-stakes math test and *Self-Efficacy in Math*; $r = 0.19$), we still observe considerable movement across quintiles. Of the 22 teachers in the top quintile of effectiveness based on the low-stakes math test (of the 108 total teachers who have scores for both measures), only 9 (or 41 percent) remain in the top quintile of effectiveness based on *Self-Efficacy in Math*; 32 percent are in the lowest two quintiles (see Table VI.4). Together, these results indicate that teachers have meaningful impacts on both test scores and non-tested outcomes, but that individual teachers often are not equally effective at improving all measures.

Table VI.4. Quintiles of effectiveness for low-stakes math test versus self-efficacy in math

Low-stakes math test quintile	Self-efficacy in math quintile				
	Top quintile	Second	Third	Fourth	Bottom quintile
Top quintile	9	3	3	4	3
Second quintile	3	6	5	4	4
Third quintile	5	3	2	7	4
Fourth quintile	4	5	3	6	4
Bottom quintile	1	5	8	1	6

Notes: Sample includes 108 teachers.

B. Teaching effects on test scores and non-tested outcomes

Next, we explore whether certain characteristics of teachers' instructional practice help explain the sizable teacher effects described above. In Table VI.5, we present estimates from equation (2) relating each of our four dimensions of instructional practice to students' math test scores and non-tested outcomes. Estimates are conditioned on the other three teaching practices, a select set of teacher characteristics, and student and class characteristics, and are presented as standardized effect sizes.

We find statistically significant relationships between both general and content-specific teaching practices and a range of non-tested outcomes. For general teaching practices, our findings support hypotheses from existing literature, with *Emotional Support* positively associated with the two closely related student constructs of *Self-Efficacy in Math* (0.13 sd) and *Happiness in Class* (0.37 sd). This finding makes 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. Our estimate for *Happiness in Class* is quite large in magnitude but is imprecisely estimated, given the smaller sample size available for this outcome. *Classroom Organization*, which captures teachers' behavior management skills and productivity, also is positively related to students' reports of their own *Behavior in Class* (0.08 sd).

Table VI.5. Teaching effects on test scores and non-tested outcomes

	High-stakes math test	Low-stakes math test	Behavior in class	Self-efficacy in math	Happiness in class
Teaching characteristics					
Emotional support	0.015 (0.014)	0.020 (0.015)	0.034 (0.032)	0.133*** (0.036)	0.367*** (0.104)
Classroom organization	-0.023 (0.014)	-0.019 (0.015)	0.079* (0.037)	-0.009 (0.044)	-0.282* (0.119)
Ambitious mathematics instruction	0.014 (0.015)	0.015 (0.016)	-0.027 (0.040)	-0.026 (0.045)	0.091 (0.079)
Mathematical errors	-0.021 (0.014)	0.001 (0.015)	-0.034 (0.036)	-0.094* (0.041)	-0.245~ (0.128)
Teacher characteristics					
Mathematics coursework	0.023 (0.024)	0.013 (0.025)	-0.079 (0.070)	-0.017 (0.080)	-0.025 (0.168)
Mathematical content knowledge	0.007 (0.017)	0.023 (0.018)	-0.024 (0.045)	-0.004 (0.052)	0.032 (0.095)
Alternative certification	-0.005 (0.078)	0.019 (0.079)	0.025 (0.134)	0.046 (0.152)	-0.145 (0.269)
Experience (three years or fewer)	0.056 (0.059)	0.067 (0.062)	0.184 (0.262)	-0.304 (0.293)	0.613 (0.559)
Teacher observations	**196**	**196**	**93**	**90**	**47**
Student observations	**8,660**	**8,660**	**1,362**	**1,275**	**517**

Notes: ~ p < 0.10, *p < 0.05, ***p < 0.001. Columns contain estimates from separate regressions. All models control for student and class characteristics, and include school fixed effects and teacher random effects. Models predicting all outcomes except for Happiness in Class also include class random effects.

Further, we find relationships between domains of teaching practice and non-tested outcomes beyond those theorized in previous research. The degree to which teachers commit mathematical errors was negatively related to *Self-Efficacy in Math* (-0.09 sd) and *Happiness in Class* (-0.25 sd). These findings illuminate how a teacher's ability to present mathematics with clarity and without serious mistakes is related to their students' self-confidence in math and enjoyment in class. We also find that *Classroom Organization* was negatively associated with *Happiness in Class* (-0.28 sd), which suggests that classrooms overly focused on routines and

management are negatively related to students' social and emotional experiences in class.[17] This relationship stands in contrast to the positive relationship between *Classroom Organization* and *Behavior in Class*, which we discuss below in our conclusion.

Comparatively, when predicting scores on both math tests, we find no statistically significant relationships for any domain of teaching practice. For *Emotional Support, Ambitious Mathematics Instruction*, and *Mathematical Errors*, estimates generally were signed in the way we would expect—positive for the first two teaching constructs and negative for the latter, with higher scores indicating worse instruction. However, magnitudes were no larger than 0.02 sd. For *Classroom Organization*, estimates were signed opposite from expectation, although similarly small in magnitude. Given the consistency of estimates across the two math tests and our restricted sample size, it is possible that non-significant results were due to limited statistical power. This may be particularly true for *Ambitious Mathematics Instruction*; in this case, point estimates were smaller than those found by Blazar (2015), who conducted similar analyses in a subset of the NCTE data.[18] However, even if true relationships exist between math-specific (or other) teaching practices and students' math test scores, these relationships likely are smaller than those between general teaching practices and non-tested outcomes. We find that the 95 percent confidence intervals relating *Classroom Emotional Support* to both *Self-Efficacy in Math* [0.062, 0.204] and *Happiness in Class* [0.163, 0.571] did not overlap with the 95 percent confidence intervals for any of the point estimates predicting math test scores.

Finally, we find that none of the four teacher characteristics—mathematics coursework, mathematical content knowledge, alternative certification, and an indicator for whether a teacher has three or fewer years of experience—was related to any outcome, conditional on measures of teaching practices. This finding is consistent with a large body of literature in which very few characteristics about teachers' backgrounds predict test scores (for example, Boyd et al. 2009; Harris and Sass 2011; Wayne and Youngs 2003).

[17] When we adjusted *p*-values for estimates presented in Table VI.4 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* remained statistically significant.

[18] In that analysis, *Ambitious Mathematics Instruction* was a statistically significant predictor of scores on the low-stakes math test—the only outcome measure used—at 0.11 sd. Significantly, the 95 percent confidence interval around this point estimate overlapped with the 95 percent confidence interval 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 (-0.04 sd for *Emotional Support*, -0.002 sd for *Classroom Organization*, and -0.03 sd for *Mathematical Errors*). Differences between the two studies likely emerge from the fact that we drew on a larger sample with an additional year of data as well as slight modifications to our identification strategy. See Blazar (2015) for more details.

VII. DISCUSSION AND CONCLUSION

A. Relationship between our findings and previous research

The teacher effectiveness literature has shaped education policy profoundly over the last decade and has served as the catalyst for sweeping reforms around teacher recruitment, evaluation, development, and retention. We extend this literature here by estimating teacher and teaching effects on both math test scores and a range of academic behaviors and mindsets. To our knowledge, this study is the first to identify teacher effects on self-efficacy in math and happiness in class, as well as on a self-reported measure of student behavior. Further, this is the only study to estimate teaching effects on non-tested outcomes using high quality measures of teaching practice and models that limit the degree to which estimates might be driven by omitted variables bias.

In many ways, our findings align with conclusions drawn from previous research (Jackson 2012; Jennings and DiPrete 2010; Koedel 2008; Ruzek et al. 2014). Consistent with these studies, we find strong evidence of teacher effects on a range of outcomes beyond test scores. As expected, our estimates are slightly smaller than those calculated by Jennings and DiPrete (2010), whose estimates conflated teacher and class effects. Further, like these authors and Jackson (2012), we find weak correlations between teachers' effectiveness ratings across outcome types. For example, more than 25 percent of the teachers ranked in the top quintile of effectiveness when evaluated using the low-stakes math test fell in the bottom two quintiles of effectiveness when evaluated using students' self-efficacy in math. Interestingly, we also find relatively weak correlations between teacher effects on students' self-reported self-efficacy in math and their happiness in class, indicating that attention to just one type of non-tested outcome may be insufficient to describe teachers' skill in the classroom. Finally, our estimates of teacher effects across high- and low-stakes tests are similar to Corcoran et al. (2012) but substantively larger than those found by Lockwood et al. (2007) and Papay (2011).

An additional contribution of this work is our focus on teaching and its relation to students' academic behaviors and mindsets. Findings linking general instructional pedagogy to closely related student outcomes suggest that students likely are influenced by the behavior first modeled by their teacher. In particular, we find that teachers' social and emotional support for students is related quite strongly to a range of non-tested outcomes, including their self-efficacy in math and happiness in class. However, there seems to be a tradeoff for some teaching practices. High quality instruction around classroom organization is positively related to students' self-reported behavior in class but negatively related to their happiness in class. Given that both of these outcomes predict labor market success (Chetty et al. 2011; Lyubomirsky et al. 2005), further research will be critical to gain a better understanding of how teachers can develop classroom environments that engender both constructive classroom behavior and students' happiness in class. Finally, our results also suggest that teachers who make mathematical errors not only fail to deliver accurate academic content but also likely lower students' self-efficacy in math and happiness in class. These relationships make sense and support the importance of efforts to increase teachers' content knowledge. Relationships for other teaching practices are more challenging to interpret. Together, these findings provide important validity evidence for observation instruments, which are now widely used in teacher development programs and evaluation systems, and point to specific teaching practices that may be a focus of these efforts.

B. Implications for policy

Together, these findings lend empirical support to decades of research on the multidimensional and complex nature of teaching and learning (Cohen 2011; Lampert 2001; Leinhardt 1993), and thus the need for policymakers and administrators to recognize and account for this complexity. This need presents a challenging trade-off between the need to expand the ways in which teacher effectiveness is conceptualized and measured while recognizing the current limitations of existing measures of non-tested student outcomes.

Our evidence may generate interest among some policymakers to incorporate teacher effect estimates on students' academic behavior and mindsets into high-stakes personnel decisions. Although we find that teacher effects on non-tested outcomes can be measured with reasonably similar precision to effects on test scores, evidence suggests that current measures derived from student self-reported questionnaires may be prone to reference bias and social desirability bias (Duckworth and Yeager 2015; West et al. forthcoming). We attempted to minimize these potential biases in our analyses by restricting our comparisons to teachers within the same schools. However, this approach to calculating teacher effects may not be realistic in policy contexts where teachers often are compared across rather than within schools. Further, little is known about how these outcomes might function under high-stakes conditions. Student responses on self-reported questionnaires could easily be coached. Such incentives would likely also render teacher assessments of their students' behavior inappropriate. Thus, there is a clear need for additional research on the reliability and validity of non-tested measures, as well as the development of objective performance measures that can capture these outcomes.

Given these challenges, another approach to evaluation efforts may be to place more weight on teaching practices that are predictive of non-tested outcomes. For example, these systems may place greater emphasis on affective and organizational behaviors as well as precision of math content, which we find to be related to a range of non-tested outcomes. 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. Further, increased emphasis on these teaching practices likely will have stronger face validity for teachers than test-based metrics of effectiveness because of the intuitive relationship between these teaching behaviors and an array of student outcomes.

Results linking teaching behaviors to students' non-tested outcomes may also be useful in teacher development efforts, which many argue should be a primary focus of evaluation systems (Darling-Hammond 2013; Hill and Grossman 2013; Odden 2004; Papay 2012). One possibility would be to pair lower-skilled teachers with programs designed specifically to strengthen their interpersonal relationships with students and their classroom organization. For example, coaching programs such as My Teaching Partner have been shown to improve teachers' relationships with students in experimental trials (Allen et al. 2011; Gregory et al. 2013). In some settings, MATCH Teacher Coaching can improve classroom management skills (Blazar and Kraft forthcoming). In turn, these programs may also impact students' behavior, self-efficacy, and happiness.

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 academic behaviors and mindsets that are equally important for students' long-term success.

REFERENCES

Allen, J.P., R.C. Pianta, A. Gregory, A.Y. Mikami, and J. Lun. "An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement." *Science*, vol. 333, no. 6045, 2011, pp. 1034–1037.

Bandura, A. "Self-Efficacy: Toward a Unifying Theory of Behavioral Change." *Psychological Bulletin*, vol. 84, no. 2, 1977, pp. 191–215.

Baron, J. "Personality and Intelligence." In *Handbook of Human Intelligence* (pp. 308–351), edited by R.J. Sternberg. New York: Cambridge University Press, 1982.

Bell, C.A., D.H. Gitomer, D.F. McCaffrey, B.K. Hamre, R.C. Pianta, and Y. Qi. "An Argument Approach to Observation Protocol Validity." *Educational Assessment*, vol. 17, nos. 2–3, 2012, pp. 62–87.

Bell, A.J., L.A. Rosen, and D. Dynlacht. "Truancy Intervention." *Journal of Research and Development in Education*, vol. 27, 1994, pp. 203–211.

Blazar, D. "Effective Teaching in Elementary Mathematics: Identifying Classroom Practices that Support Student Achievement." *Economics of Education Review*, vol. 48, 2015, pp. 16–29.

Blazar, D., D. Braslow, C.Y. Charalambous, and H.C. Hill. "Attending to General and Content-Specific Dimensions of Teaching: Exploring Factors Across Two Observation Instruments." Working Paper. Cambridge, MA: National Center for Teacher Effectiveness, 2015.

Blazar, D., and M.A. Kraft. "Exploring Mechanisms of Effective Teacher Coaching: A Tale of Two Cohorts from a Randomized Experiment." *Educational Evaluation and Policy Analysis*, forthcoming.

Borghans, L., A.L. Duckworth, J.J. Heckman, and B. Ter Weel. "The Economics and Psychology of Personality Traits." *Journal of Human Resources*, vol. 43, no. 4, 2008, pp. 972–1059.

Boyd, D.J., P.L. Grossman, H. Lankford, S. Loeb, and J. Wyckoff. "Teacher Preparation and Student Achievement." *Educational Evaluation and Policy Analysis*, vol. 31, no. 4, 2009, pp. 416–440.

Center on Great Teachers and Leaders. "Databases on State Teacher and Principal Policies." 2013. Available at http:// http://resource.tqsource.org/stateevaldb.

Centra, J.A., and D.A. Potter. "School and Teacher Effects: An Interrelational Model." *Review of Educational Research*, vol. 5, no. 2, 1980, pp. 273–291.

Chetty, R., J.N. Friedman, N. Hilger, E. Saez, D. Schanzenbach, and D. Yagan. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics*, vol. 126, no. 4, 2011, pp. 1593–1660.

Chetty, R., J.N. Friedman, and J.E. Rockoff. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review,* vol. 104, no. 9, 2014, pp. 2593–2632.

Clark, M.A., H.S. Chiang, T. Silva, S. McConnell, K. Sonnenfeld, A. Erbe, and M. Puma. "The Effectiveness of Secondary Math Teachers from Teach for America and the Teaching Fellows Programs." Washington, DC: U.S. Department of Education, 2013.

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

Corcoran, S.P., J.L. Jennings, and A.A. Beveridge. "Teacher Effectiveness on High- and Low-Stakes Tests." 2012. Unpublished manuscript. Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.269.5537&rep=rep1&type=pdf.

Danielson, C. *Enhancing Professional Practice: A Framework for Teaching*. Alexandria, VA: Association for Supervision and Curriculum Development, 2011.

Darling-Hammond, L. *Getting Teacher Evaluation Right: What Really Matters for Effectiveness and Improvement.* New York: Teachers College Press, 2013.

Decker, P.T., D.P. Mayer, and S. Glazerman. "The Effects of Teach for America on Students: Findings from a National Evaluation." Princeton, NJ: Mathematica Policy Research, 2004.

Dee, T.S., and J. Wyckoff. "Incentives, Selection, and Teacher Performance: Evidence from IMPACT." *Journal of Policy Analysis and Management*, vol. 34, no. 2, 2015, pp. 267–297.

Diener, E. "Subjective Well-Being: The Science of Happiness and a Proposal for a National Index." *American Psychologist,* vol. 55, no. 1, 2000, pp. 34–43.

Duckworth, A.L., C. Peterson, M.D. Matthews, and D.R. Kelly. "Grit: Perseverance and Passion for Long-Term Goals." *Journal of Personality and Social Psychology*, vol. 92, no. 6, 2007, pp. 1087–1101.

Duckworth, A.L., P.D. Quinn, and E. Tsukayama. "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,* vol. 104, no. 2, 2012, pp. 439–451.

Duckworth, A.L., and D.S. Yeager. "Measurement Matters: Assessing Personal Qualities Other than Cognitive Ability for Educational Purposes." *Educational Researcher*, vol. 44, no. 4, 2015, pp. 237–251.

Dunn, O.J. "Multiple Comparisons Among Means." *Journal of the American Statistical Association*, vol. 56, no. 293, 1961, pp. 52–64.

Dunning, D., C. Heath, and J.M. Suls. "Flawed Self-Assessment Implications for Health, Education, and the Workplace." *Psychological Science in the Public Interest*, vol. 5, no. 3, 2004, pp. 69–106.

Dweck, C.S. *Mindset: The New Psychology of Success*. New York: Random House, 2006.

Dweck, C.S., G.M. Walton, and G.L. Cohen. "Academic Tenacity: Mindsets and Skills that Promote Long-Term Learning." White paper prepared for the Gates Foundation. Seattle, WA, 2011.

Farrington, C.A., M. Roderick, E. Allensworth, J. Nagaoka, T.S. Keyes, D.W. Johnson, and N.O. Beechum. "Teaching Adolescents to Become Learners: The Role of Non-Cognitive Factors in Shaping School Performance, a Critical Literature Review." Chicago: University of Chicago Consortium on Chicago School Reform, 2012.

Field, A. *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). London: SAGE publications, 2013.

Gehlbach, H. "Name that Baby: Why 'Non-Cognitive' Factors Need a New Name." *Education Week.* April 15, 2015. Available at http://blogs.edweek.org/edweek/rick_hess_straight_up/2015/04/non-cognitive_factors_need_new_name.html.

Gershenson, S. "Linking Teacher Quality, Student Attendance, and Student Achievement." *Education Finance and Policy*, forthcoming.

Gregory, A., J.P. Allen, A.Y. Mikami, C.A. Hafen, and R.C. Pianta. "Effects of a Professional Development Program on Behavioral Engagement of Students in Middle and High School." *Psychology in the Schools,* 40(3), 2013, pp. 1–22.

Hanushek, E.A., and S.G. Rivkin. "Generalizations About Using Value-Added Measures of Teacher Quality." *American Economic Review,* vol. 100, no. 2, 2010, pp. 267–271.

Harris, D., and T. Sass. "Teacher Training, Teacher Quality, and Student Achievement." *Journal of Public Economics*, vol. 95, 2011, pp. 798–812.

Harville, D.A. "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems." *Journal of the American Statistical Association,* vol. 72, no. 358, 1977, pp. 320–338.

Hawkins, J.D., T. Herrenkohl, D.P. Farrington, D. Brewer, R.F. Catalano, and T.W. Harachi. "A Review of Predictors of Youth Violence." In *Serious and Violent Juvenile Offenders* (pp. 30–46), edited by R. Loeber and D.P. Farrington. Thousand Oaks, CA: Sage, 1998.

Heckman, J.J., and Y. Rubinstein. "The Importance of Non-Cognitive Traits: Lessons from the GED Testing Program." *American Economic Review,* vol. 91, no. 2, 2001, pp. 145–149.

Hickman, J.J., J. Fu, and H. C. Hill. "Technical Report: Creation and Dissemination of Upper-Elementary Mathematics Assessment Modules." Princeton, NJ: Educational Testing Service, 2012.

Hill, H.C., D. Blazar, and K. Lynch. "Resources for Teaching: Examining Personal and Institutional Predictors of High-Quality Instruction." *AERA Open*, forthcoming.

Hill, H.C., M.L. Blunk, C.Y. Charalambous, J.M. Lewis, G.C. Phelps, L. Sleep, and D.L. Ball. "Mathematical Knowledge for Teaching and the Mathematical Quality of Instruction: An Exploratory Study." *Cognition and Instruction*, vol. 26, no. 4, 2008, pp. 430–511.

Hill, H.C., C.Y. Charalambous, and M.A. Kraft. "When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study." *Educational Researcher*, vol. 41, no. 2, 2012, pp. 56–64.

Hill, H.C., and P. Grossman. "Learning from Teacher Observations: Challenges and Opportunities Posed by New Teacher Evaluation Systems." *Harvard Educational Review*, vol. 83, no. 2, 2013, pp. 371–384.

Hill, H.C., S.G. Schilling, and D.L. Ball. "Developing Measures of Teachers' Mathematics Knowledge for Teaching." *Elementary School Journal*, vol. 105, 2004, pp. 11–30.

Ho, A.D., and T.J. Kane. "The Reliability of Classroom Observations by School Personnel." Seattle, WA: Measures of Effective Teaching Project, Bill and Melinda Gates Foundation, 2013.

Huesmann, L.R., L.D. Eron, M.M. Lefkowitz, and L.O. Walder. "Stability of Aggression over Time and Generations." *Developmental Psychology*, vol. 20, no. 6, 1984, pp.1120–1134

Jackson, C.K. "Non-Cognitive Ability, Test Scores, and Teacher Quality: Evidence from Ninth Grade Teachers in North Carolina." NBER Working Paper No. 18624. Cambridge, MA: National Bureau for Economic Research, 2012.

Jennings, J.L., and T.A. DiPrete. "Teacher Effects on Social and Behavioral Skills in Early Elementary School." *Sociology of Education*, vol. 83, no. 2, 2010, pp.135–159.

John, O.P., and S. Srivastava. "The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives." *Handbook of Personality: Theory and Research*, vol. 2, 1999, pp. 102–138.

Kane, T.J., D.F. McCaffrey, T. Miller, and D.O. Staiger. "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, 2013.

Kane, T.J., and D.O. Staiger. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." NBER Working Paper No. 14607. Cambridge, MA: National Bureau for Economic Research, 2008.

Kane, T.J., and D.O. Staiger. "Gathering Feedback for Teaching." Seattle, WA: Measures of Effective Teaching Project, Bill and Melinda Gates Foundation, 2012.

Kane, T.J., E.S. Taylor, J.H. Tyler, and A.L. Wooten. "Identifying Effective Classroom Practices Using Student Achievement Data." *Journal of Human Resources,* vol. 46, no. 3, 2011, pp. 587–613.

King, R.B., D.M., McInerney, F.A. Ganotice, and J.B. Villarosa. "Positive Affect Catalyzes Academic Engagement: Cross-Sectional, Longitudinal, and Experimental Evidence. *Learning and Individual Differences*, vol. 39, 2015, pp. 64-72.

Kline, P. *An Easy Guide to Factor Analysis*. London: Routledge, 1994.

Koedel, C. "Teacher Quality and Dropout Outcomes in a Large, Urban School District." *Journal of Urban Economics*, vol. 64, no. 3, 2008, pp. 560–572.

Lampert, M. *Teaching Problems and the Problems of Teaching*. New Haven, CT: Yale University Press, 2011.

Leinhardt, G. "On Teaching." In *Advances in Instructional Psychology* (vol. 4, pp. 1–54), edited by R. Glaser. Hillsdale, NJ: Lawrence Erlbaum Associates, 1993.

Lindqvist, E., and R. Vestman. "The Labor Market Returns to Cognitive and Non-Cognitive Ability: Evidence from the Swedish Enlistment." *American Economic Journal: Applied Economics*, vol. 3, no. 1, 2011, pp. 101–128.

Lockwood, J.R., D.F. McCaffrey, L.S. Hamilton, B. Stecher, V. Le, and J.F. Martinez. "The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures." *Journal of Educational Measurement*, vol. 44, no. 1, 2007, pp. 47–67.

Loeb, S., L.C. Miller, and J. Wyckoff. "Performance Screens for School Improvement: The Case of Teacher Tenure Reform in New York City." *Educational Researcher*, vol. 44, no. 4, 2015, pp. 199–212.

Loeber, R. "The Stability of Antisocial and Delinquent Child Behavior: A Review." *Child Development*, vol. 53, no. 6, 1982, pp. 1431–1446.

Loeber, R., and D.P. Farrington. "Young Children Who Commit Crime: Epidemiology, Developmental Origins, Risk Factors, Early Interventions, and Policy Implications." *Development and Psychopathology*, vol. 12, 2000, pp. 737–762.

Lynch, K., M. Chin, and D. Blazar. "Relationship Between Observations of Elementary Teacher Mathematics Instruction and Student Achievement: Exploring Variability Across Districts." Working Paper. Cambridge, MA: National Center for Teacher Effectiveness, 2015.

Lyubomirsky, S., L. King, and E. Diener. "The Benefits of Frequent Positive Affect: Does Happiness Lead to Success?" *Psychological Bulletin*, vol. 131, no. 6, 2005, pp. 803–855.

McCaffrey, D.F., T.R. Sass, J.R. Lockwood, and K. Mihaly. "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy*, vol. 4, no. 4, 2009, pp. 572–606.

Metzler, J., and L. Woessmann. "The Impact of Teacher Subject Knowledge on Student Achievement: Evidence from Within-Teacher Within-Student Variation." *Journal of Development Economics,* vol. 99, no. 2, 2012, pp. 486–496.

Mihaly, K., D.F. McCaffrey, D.O. Staiger, and J.R. Lockwood. "A Composite Estimator of Effective Teaching." Seattle, WA: Measures of Effective Teaching Project, Bill and Melinda Gates Foundation, 2013.

Moffitt, T.E. "Adolescence-Limited and Life-Course-Persistent Antisocial Behavior: A Developmental Taxonomy." *Psychological Review,* vol. 100, no. 4, 1993, pp. 674–701.

Moffitt, T.E., L. Arseneault, D. Belsky, N. Dickson, R.J. Hancox, H. Harrington, R. Houts, R. Poulton, B.W. Roberts, and S. Ross. "A Gradient of Childhood Self-Control Predicts Health, Wealth, and Public Safety." *Proceedings of the National Academy of Sciences,* vol. 108, no. 7, 2011, pp. 2693–2698.

Mueller, G., and E. Plug. "Estimating the Effect of Personality on Male and Female Earnings." *Industrial & Labor Relations Review,* vol. 60, no. 1, 2006, pp. 3–22.

Murayama, K., R. Pekrun, S. Lichtenfeld, and R. vom Hofe. "Predicting Long-Term Growth in Students' Mathematics Achievement: The Unique Contributions of Motivations and Cognitive Strategies." *Child Development,* vol. 84, no. 4, 2012, pp. 1475–1490.

Nagin, D., and R.E. Tremblay. "Trajectories of Boys' Physical Aggression, Opposition, and Hyperactivity on the Path to Physically Violent and Nonviolent Juvenile Delinquency." *Child Development,* vol. 70, no. 5, 1999, pp. 1181–1196.

National Council of Teachers of Mathematics. *Curriculum and Evaluation Standards for School Mathematics.* Reston, VA: 1989.

National Council of Teachers of Mathematics. *Professional Standards for Teaching Mathematics.* Reston, VA: 1991.

National Council of Teachers of Mathematics. *Principles and Standards for School Mathematics*. Reston, VA: 2000.

National Governors Association Center for Best Practices. *Common Core State Standards for Mathematics*. Washington, DC: 2010.

Nye, Konstantopoulos, and Hedges. "How Large are Teacher Effects?" *Educational Evaluation and Policy Analysis,* vol. 26, no. 3, 2004, pp. 237–257.

Odden, A. "Lessons Learned About Standards-Based Teacher Evaluation Systems." *Peabody Journal of Education,* vol. 79, no. 4, 2004, pp. 126–137.

Osterman, K.F. "Students' Need for Belonging in the School Community." *Review of Educational Research,* vol. 70, no. 3, 2000, pp. 323–367.

Oswald, A.J., E. Proto, and D. Sgroi. "Happiness and Productivity." *Journal of Labor Economics,* forthcoming.

Papay, J.P. "Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures." *American Educational Research Journal,* vol. 48, no. 1, 2011, pp. 163–193.

Papay, J.P. "Refocusing the Debate: Assessing the Purposes and Tools of Teacher Evaluation." *Harvard Educational Review,* vol. 82, no. 1, 2012, pp. 123–141.

Pianta, R.C., and B.K. Hamre. "Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity." *Educational Researcher,* vol. 38, no. 2, 2009, pp. 109–119.

Raudenbush, S.W., and A.S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods. Second Edition*. Thousand Oaks, CA: Sage Publications, 2002.

Robins, L.N., and K.S. Ratcliff. "The Long-Term Outcome of Truancy." In *Out of School: Modern Perspectives in Truancy and School Refusal* (pp. 65–83), edited by L. Hersov and I. Berg. New York: John Wiley, 1980.

Roorda, D.L., H.M. Koomen, J.L. Spilt, and F.J. Oort. "The Influence of Affective Teacher-Student Relationships on Students' School Engagement and Achievement: A Meta-Analytic Approach." *Review of Educational Research,* vol. 81, no. 4, 2011, pp. 493–529.

Rosenberg, M. *Society and the Adolescent Self-Image*. Middletown, CT: Wesleyan University Press, 1989.

Rothstein, J. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *The Quarterly Journal of Economics,* vol. 125, no. 1, 2010, pp. 175–214.

Ruzek, E.A., T. Domina, A.M. Conley, G.J. Duncan, and S.A. Karabenick. "Using Value-Added Models to Measure Teacher Effects on Students' Motivation and Achievement." *The Journal of Early Adolescence,* vol. 35, nos. 5–6, 2014, pp. 852–882.

Segal, C. "Misbehavior, Education, and Labor Market Outcomes." *Journal of the European Economic Association,* vol. 11, no. 4, 2013, pp. 743–779.

Šidák, Z. "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions." *Journal of the American Statistical Association,* vol. 62, no. 318, 1967, pp. 626–633.

Spearman, C. "'General Intelligence,' Objectively Determined and Measured." *The American Journal of Psychology,* vol. 15, no. 2, 1904, pp. 201–292.

Stecher, B.M., and L.S. Hamilton. *Measuring Hard-to-Measure Student Competencies: A Research and Development Plan.* Santa Monica, CA: RAND Corporation, 2014.

Tabachnick, B.G., and L.S. Fidell. *Using Multivariate Statistics* (4th ed.). New York: Harper Collins, 2001.

Tremblay, R.E., B. Masse, D. Perron, M. LeBlanc, A.E. Schwartzman, and J.E. Ledingham. "Early Disruptive Behavior, Poor School Achievement, Delinquent Behavior, and Delinquent Personality: Longitudinal Analyses." *Journal of Consulting and Clinical Psychology*, vol. 60, no. 1, 1992, p. 64–72.

Tsukayama, E., A.L. Duckworth, and B. Kim. "Domain-Specific Impulsivity in School-Age Children." *Developmental Science,* vol. 16, no. 6, 2013, pp. 879–893.

Wayne, A.J., and P. Youngs. "Teacher Characteristics and Student Achievement Gains: A Review." *Review of Educational Research,* vol. 73, no. 1, 2003, pp. 89–122.

West, M.R., M.A. Kraft, A.S. Finn, R. Martin, A.L. Duckworth, C.F. Gabrieli, and J.D. Gabrieli. "Promise and Paradox: Measuring Students' Non-Cognitive Traits and the Impact of Schooling." *Educational Evaluation and Policy Analysis*, forthcoming.

Willson, I.A. "Changes in Mean Levels of Thinking in Grades 1–8 Through Use of an Interaction Analysis System Based on Bloom's Taxonomy." *Journal of Educational Research,* vol. 66, no. 9, 1973, pp. 423–429.

Zimmerman, D.W., and R.H. Williams. "Properties of the Spearman Correction for Attenuation for Normal and Realistic Non-Normal Distributions." *Applied Psychological Measurement,* vol. 21, no. 3, 1997, pp. 253–270.

APPENDIX

Table A.1. 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
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
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
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, I focus on learning, not the time the 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

Notes: Estimates drawn from all available data. Loadings of roughly 0.4 or higher are highlighted to identify patterns.

Table A.2. Differences between teachers who switch grade assignments and those who do not

	Never switch	Switch	*P*-value on difference
Value added on high-stakes math test	0.01	0.01	0.871
Emotional support	-0.02	0.10	0.491
Classroom organization	0.06	0.08	0.869
Ambitious mathematics instruction	0.00	-0.03	0.777
Mathematical errors	-0.07	0.03	0.464
Join test		*F*-statistic	0.26
		p-value	0.937
Teacher-year observations	**506**	**66**	

Notes: Means and *p*-values calculated from regression framework that controls for school-by-year blocks.

About the series

Policymakers and researchers require timely, accurate, evidence-based research as soon as it's available. Further, statistical agencies need information about statistical techniques and survey practices that yield valid and reliable data. To meet these needs, Mathematica's working paper series offers access to our most current work.

For more information about this paper, contact info@mathematica-mpr.com.

Suggested citation: Blazar, David, and Kraft, Matthew A. "Teacher and Teaching Effects on Students' Academic Behaviors and Mindsets," Working Paper 41. Cambridge, MA: Mathematica Policy Research, December 2015.

Note: David Blazar was a summer fellow at Mathematica in 2015.

www.mathematica-mpr.com

Improving public well-being by conducting high-quality, objective research and surveys

PRINCETON, NJ - ANN ARBOR, MI - CAMBRIDGE, MA - CHICAGO, IL - OAKLAND, CA - WASHINGTON, DC

MATHEMATICA
Policy Research