

VALIDATING TEACHER EFFECTS ON STUDENTS' ATTITUDES AND BEHAVIORS: EVIDENCE FROM RANDOM ASSIGNMENT OF TEACHERS TO STUDENTS

David Blazar
University of Maryland

Abstract

There is growing interest among researchers, policymakers, and practitioners in identifying teachers who are skilled at improving student outcomes beyond test scores. However, important questions remain about the validity of these teacher effect estimates. Leveraging the random assignment of teachers to classes, I find that teachers have causal effects on their 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. Weak correlations between teacher effects on different student outcomes indicate that these measures capture unique skills that teachers bring to the classroom. Teacher effects calculated in non-experimental data are related to these same outcomes following random assignment, revealing that they contain important information content on teachers. However, for some non-experimental teacher effect estimates, large and potentially important degrees of bias remain. These results suggest that researchers and policymakers should proceed with caution when using these measures. They likely are more appropriate for low-stakes decisions, such as matching teachers to professional development, than for high-stakes personnel decisions and accountability.

Suggested Citation:

Blazar, D. (Forthcoming). Validating teacher effects on students' attitudes and behaviors: Evidence from random assignment of teachers to students. *Education Finance and Policy*.

Correspondence regarding the paper can be sent to David Blazar at dblazar@umd.edu; Department of Teaching and Learning, Policy and Leadership, 2311 Benjamin Building, University of Maryland College of Education, College Park, MD, 20742. I would like to thank Martin West, Thomas Kane, Heather Hill, and anonymous reviewers for their feedback on earlier drafts. The data collection effort described here was supported 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 author and do not represent views of the Institute or the U.S. Department of Education. Research support came from the Mathematica Policy Research summer fellowship and the Smith Richardson Foundation.

1. Introduction

Decades worth of research on education production have narrowed in on the importance of teachers to student outcomes (Murnane and Phillips 1981; Todd and Wolpin 2003). Over the last several years, these studies have coalesced around two key findings. First, teachers vary considerably in their ability to improve students' academic performance (Hanushek and Rivkin 2010; Nye, Konstantopoulos, and Hedges 2004), which in turn influences a variety of long-term outcomes including teenage pregnancy rates, college attendance, and earnings in adulthood (Chetty, Friedman, and Rockoff 2014b). Second, experimental and quasi-experimental studies indicate that "value-added" approaches to estimating teachers' contribution to student test scores are valid ways to identify effective teachers (Bacher-Hicks et al. 2017; Chetty, Friedman, and Rockoff 2014a; Glazerman and Protik 2015; Kane et al. 2013; Kane and Staiger 2008). In other words, on average, these teacher effect estimates are not confounded with the non-random sorting of teachers to students, the specific set of students in the classroom, or factors beyond teachers' control. Policymakers have taken notice of these findings, leading to widespread changes in teacher evaluation, compensation, and promotion.

While the studies described above have focused predominantly on teachers' impact on students' academic performance, the research community is starting to have evidence that teachers also vary in their contributions to a variety of other student outcomes in ways that are only weakly related to their effects on test scores (Gershenson 2016; Jackson 2012; Jennings and DiPrete 2010; Kraft forthcoming). For example, in work drawing on the study that generated data used in this paper, Blazar and Kraft (2017) found that teachers identified as 1 standard deviation (SD) above the mean in the distribution of effectiveness improved students' self-reported behavior in class, self-efficacy in math, and happiness in class by between 0.15 SD to

0.30 SD; these effects are similar to or larger than teacher effects on students' test scores (Hanushek and Rivkin 2010). However, teachers who were effective at improving these outcomes often were not equally effective at improving students' math test scores, with correlations between teacher effect estimates no higher than 0.19. Jackson (2012) came to similar conclusions using additional student outcomes, and also found that teacher effects on non-test score outcomes captured in ninth grade predicted longer-run outcomes including high-school completion above and beyond teachers' effects on test scores (Jackson, 2016). Together, these findings lend empirical evidence to the multidimensional nature of teaching and, thus, the need for policymakers to account for this sort of complexity.

Given that the research base examining teachers' contributions to student outcomes beyond test scores is relatively new, important questions remain about the validity of these measures. In the value-added literature more broadly, researchers have asked about the sensitivity of teacher effects to different model specifications and the specific set of covariates included in the model (Goldhaber and Theobald 2012), as well as the most appropriate ways to calculate these scores in light of measurement error (Guarino et al. 2015). Further, it is not clear whether the key identifying assumption underlying the estimation of teacher effects – that estimates are not biased by non-random sorting of students to teachers (Chetty, Friedman, and Rockoff 2014a; Kane et al. 2013) – holds when test scores are replaced with other student outcomes. Researchers who estimate value-added to students' test scores typically control for prior achievement because it captures many of the pre-determined factors that also affect current achievement, including the schools students attend, the neighborhoods they live in, and the family members with whom they interact. However, it is possible that there are additional factors

not captured by prior test scores or by prior measures of the outcome variable that lead to bias in teacher effects on these outcomes.

I examine these issues by drawing on a dataset in which participating students completed a survey that asked about a range of attitudes and behaviors in class. In the third year of the study, a subset of participating teachers ($N = 41$) was randomly assigned to class rosters within schools. Together, these data allow me to examine the extent to which teachers vary in their contribution to students' attitudes and behaviors, even after random assignment; the sensitivity of teacher effects on students' attitudes and behaviors to different model specifications, including those that control for students' prior academic performance versus prior attitudes and behaviors; and, ultimately, whether non-experimental estimates of teacher effects on these attitudes and behaviors predict these same outcomes following random assignment, which produces a measure of forecast bias.

Findings indicate that teachers have causal effects on students' self-reported behavior in class, self-efficacy in math, and happiness in class. The magnitude of the teacher-level variation on these outcomes is similar to or larger than effects on math test scores (e.g., Hanushek and Rivkin 2010), with estimates as large as 0.35 SD. Weak correlations between teacher effects on different student outcomes indicate that these measures capture unique skills that teachers bring to the classroom. However, value-added approaches to estimating these teacher effects appear to be insufficient to account for all sources of bias in some cases. One exception is teacher effects on students' behavior in class, where predicted differences come close to actual differences following random assignment. In the observational portion of this study, teacher effects are not particularly sensitive to models that control for students' prior achievement, student demographic characteristics, or prior survey responses. Given that these are the tools and data typically

available to the econometrician, it will be important for researchers and policymakers to use these estimates of teacher effectiveness with caution. In the conclusion, I describe some potential uses of these measures, focusing on low-stakes decision making, such as matching teachers to professional development, rather than high-stakes decisions, such as teacher evaluation and promotion.

2. Validating Methods for Estimating Teacher Effects on Student Outcomes

Over the last decade, several experimental and quasi-experimental studies have tested the validity of non-experimental methods for estimating teacher effects on student achievement. In the first of these, Kane and Staiger (2008) described the rationale and set up for such a study: “Non-experimental estimates of teacher effects attempt to answer a very specific question: If a given classroom of students were to have teacher A rather than teacher B, how much different would their average test scores be at the end of the year?” (p. 1). However, as these sorts of teacher effects estimates are derived from conditions where non-random sorting is the norm (Clotfelter, Ladd, and Vigdor 2006; Rothstein 2010), these models assume that statistical controls (e.g., students’ prior achievement, demographic characteristics) are sufficient to isolate the talents and skills of individual teachers rather than “principals’ preferential treatment of their favorite colleagues, ability-tracking based on information not captured by prior test scores, or the advocacy of engaged parents for specific teachers” (Kane and Staiger 2008, p. 1).¹

Random assignment of teachers to classes offers a way to test this assumption. If non-experimental teacher effects are causal estimates that capture true differences in quality between teachers, then non-experimental or predicted differences should be equal, on average, to actual differences following the random assignment of teachers to classes. In other words, a 1 SD increase in predicted differences in achievement across classrooms should result in a 1 SD

¹ See Bacher-Hicks et al. (2017) for an analysis of persistent sorting in the classroom data used in this study.

increase in observed differences, on average. Estimates that are statistically significantly greater than 0 SD indicate that non-experimental teacher effects contain some information content about teachers' underlying talents and skills. However, deviations from the 1:1 relationship would signal that these scores also are influenced by factors beyond teachers' control, including students' background and skill, the composition of students in the classroom, or strategic assignment policies. These deviations often are referred to as "forecast bias."

Results from Kane and Staiger (2008) and other experimental studies (Bacher-Hicks et al. 2017; Glazerman and Protik 2015; Kane et al. 2013) have accumulated to provide strong evidence against bias in teacher effects on students' test scores. Pooling results from three experimental studies with the same research design (i.e., teachers randomly assigned to class rosters within schools²), Bacher-Hicks et al. (2017) found an estimate of 0.96 SD relating predicted, non-experimental teacher effects on students' math achievement to actual differences in this same outcome. Predicted teacher effects were calculated from models that controlled for students' prior achievement. Given the nature of their meta-analytic approach, the standard error around this estimate (0.099) was much smaller than in each individual study, and the corresponding 95% confidence interval included 1 SD, indicating little bias. This result was quite similar to findings from quasi-experimental studies in much larger administrative datasets, which leveraged plausibly exogenous variation in teacher assignments due to staffing changes at the school-grade level (Bacher-Hicks, Kane, and Staiger 2014; Chetty, Friedman, and Rockoff 2014a).

² In a fourth experimental study, Glazerman and Protik (2015) exploited random assignment of teachers across schools as part of a merit pay program. Here, findings were more mixed. In the elementary sample, the authors estimated a standardized effect size relating non-experimental value-added scores (stacking across math and reading) to student test scores following random assignment of roughly 1 SD. However, in their smaller sample, the standard error was large (0.34), meaning that they could not rule out potentially large degrees of bias. Further, in the middle school sample, they found no statistically significant relationship between non-experimental and experimental teacher effect estimates.

Following a long line of inquiry around the sensitivity of value-added estimates to different model specifications and which may be most appropriate for policy (Aaronson, Barrow, and Sander 2007; Blazar, Litke, and Barmore 2016; Goldhaber and Theobald 2012; Newton et al. 2010), many of these studies also examined the predictive validity of alternative methods for estimating teacher effects. For example, some have advocated for controlling for the composition of students in the classroom, which is thought to influence test scores beyond teachers themselves (Hanushek et al. 2003; Kupermintz 2003). Others have specified models that only compare teachers within schools in order to limit bias due to sorting of teachers and students across schools (Rivkin, Hanushek, and Kain 2005); however, this approach can lead to large differences in teacher rankings relative to models that compare teachers across schools (Goldhaber and Theobald 2012). The general conclusion across validation studies is that controlling for students' prior achievement is sufficient to account for the vast majority of bias in teacher effect estimates on achievement (Chetty, Friedman, and Rockoff 2014a; Kane et al. 2013; Kane and Staiger 2008).

To my knowledge, only one study has examined the predictive validity of teacher effects on student outcomes beyond test scores. Drawing on the quasi-experimental design described by Chetty, Friedman, and Rockoff (2014a), Backes and Hansen (2015) examined the validity of teacher effects on a range of observed school behaviors captured in administrative records.³ They found that teacher effects on students' suspensions and percent of classes failed did not contain bias when pooling across all grade levels. However, teacher effects on unexcused absences,

³ Two additional studies have examined teacher effects on students' attitudes and behaviors using the random assignment portion of the Measures of Effective Teaching (MET) project. Kraft (forthcoming) found sizeable teacher effects on students' grit, growth mindset, and effort in class (0.10 to 0.17 SD). Correlations between teacher effects on students' academic performance versus effects on other outcomes were no higher than 0.22. Kane et al. (2013) found that a composite measure of teacher effectiveness based on observational data predicted student effort following random assignment. However, measures of students' attitudes and behaviors were collected in only one year. Therefore, it was not possible to relate teacher effects calculated under non-experimental conditions to teacher effects on this same outcome calculated under experimental ones.

grade point average, and on-time grade progression did contain moderate to large degrees of bias, as least in some grade levels. For both unexcused absences and on-time grade progression, predicted differences in student outcomes at the elementary level overstated actual differences (i.e., coefficient less than 1 SD), likely due to sorting of higher-performing students to higher-performing teachers in a way that could not be controlled for in the model. The opposite was true at the high school level, where predicted differences understated actual differences (i.e., coefficient greater than 1 SD). This suggests that bias in teacher effects on outcomes beyond test scores may not be easily quantified or classified across contexts.

3. Data and Sample

As in Blazar and Kraft (2017) and Bacher-Hicks et al. (2017), this paper draws on data from the National Center for Teacher Effectiveness (NCTE), whose goal was to develop valid measures of effective teaching in upper-elementary mathematics. Over the course of three school years (2010-11 through 2012-13), the project collected data from participating fourth- and fifth-grade teachers ($N = 310$) in four anonymous districts from three states on the East coast of the United States. Participants were generalists who taught all subject areas. This is important, as it provided an opportunity to estimate the contribution of individual teachers to students' attitudes and behaviors that was not confounded with the effect of another teacher with whom a student engaged with in the same year. Teacher-student links were verified for all study participants based on class rosters provided by teachers.

Measures of students' attitudes and behaviors came from a survey administered in the spring of each school year (see Appendix Table 1 for items and descriptive statistics). Based on theory and exploratory factor analyses (see Blazar and Kraft 2017), I divided items into three constructs: *Behavior in Class* (internal consistency reliability [α] is 0.74), *Self-Efficacy in Math*

($\alpha = 0.76$), and *Happiness in Class* ($\alpha = 0.82$). Teacher reports of student behavior and self-reports of versions of the latter two constructs have been linked to labor market outcomes even controlling for cognitive ability (Chetty et al. 2011; Lyubomirsky, King, and Diener 2005; Mueller and Plug 2006), lending strong consequential validity to these metrics. Blazar and Kraft (2017) describe additional validity evidence, including convergent validity, for these constructs. For each of these outcomes, I created final scales by reverse coding items with negative valence, averaging student responses across all available items, and then standardizing to mean of 0 and SD of 1.⁴ Standardization occurred within school year but across grades.

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 reading on state assessments, which were standardized within district by grade, subject, and year using the entire population of students in each district, grade, subject, and year.

I focus on two subsamples from the larger group of 310 teachers. The primary analytic sample includes the subset of 41 teachers who were part of the random assignment portion of the NCTE study in the third year of data collection. I describe this sample and the experimental design in detail below. The second sample includes the set of students (and their teachers) who took the project-administered survey in both the current and prior years. This allowed me to test the sensitivity of teacher effect estimates to different model specifications, including those that controlled for students' prior survey responses, from a balanced sample of teachers and students. As noted above, the student survey only was administered in the spring of each year; therefore,

⁴ For all three outcomes, composite scores that average across raw responses are correlated at 0.99 and above with scales that incorporate weights from the factor analysis.

this sample consisted of the group of fifth-grade teachers who happened to have students who also were part of the NCTE study in the fourth grade (N = 51 teachers; N = 548 students).⁵

Generally, I found that average teacher characteristics, including their gender, race, math course taking, math knowledge, route to certification, years of teaching experience, and value-added scores calculated from state math tests were similar across samples (see Table 1).⁶ Given that teachers self-selected into the NCTE study, I also tested whether these samples differed from the full population of fourth- and fifth-grade teachers in each district with regard to value-added scores on the state math test (see Equation (1) for more details on these value-added predictions). Although I found a marginally significant difference between the full NCTE sample and the district populations ($p = .065$), I found no difference between the district populations and either the experimental or non-experimental subsamples used in this analysis ($p = .890$ and $.652$, respectively; not shown in Table 1). These similarities lend external validity to findings presented below.

4. Experimental Design

In the spring of 2012, the NCTE project team worked with staff at participating schools to randomly assign sets of teachers to class rosters of the same grade level (i.e., fourth- or fifth-

⁵ This sample size is driven by teachers whose students had current- and prior-year survey responses for *Happiness in Class*, which was only available in two of the three years of the study. Additional teachers and students had current- and prior-year data for *Behavior in Class* (N = 111) and *Self-Efficacy in Math* (N = 108), both of which were available in all three years of the study. For consistency, I limit this sample to teachers and students who had current- and prior-year scores for all three survey measures. I did not place any restriction on the number of students each teacher needed to have in order to be included in this sample. Although others advocate excluding teachers with fewer than five students per class (Kane and Staiger 2008), for example, this would have further reduced my sample. Instead, I rely on the fact that shrinkage estimators shrink estimates more for teachers with few students than for teachers with larger classes.

⁶ Background information on teachers was captured on a questionnaire administered in the fall of each year. Survey items included years teaching math, route to certification, amount of undergraduate or graduate coursework in math and math courses for teaching (1 = No classes, 2 = One or two classes, 3 = Three to five Classes, 4 = Six or more classes). For simplicity, I 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 and the Massachusetts Test for Educator Licensure. Teacher scores were generated by IRTPro software and standardized in these models, with a reliability of 0.92. For more information about these constructs, see Hill, Blazar, and Lynch 2015.

grade) that were constructed by principals or school leaders. To be eligible for randomization, teachers had to work in schools and grades in which there was at least one other participating teacher. In addition, their principal had to consider these teachers as capable of teaching any of the rosters of students designated for the group of teachers.

In order to fully leverage this experimental design, it was important to limit the most pertinent threats to internal validity: attrition and non-compliance amongst participating teachers and students (Murnane and Willett 2011). My general approach was to focus on randomization blocks in which attrition and non-compliance were not a concern. As these blocks are analogous to individual experiments, dropping individual ones should not threaten the internal validity of results. First, I restricted the sample to blocks where teachers and their randomization block partner(s) had both current-year student outcomes and prior-year, non-experimental teacher effect estimates. Of the original 79 teachers who agreed to participate and were randomly assigned to class rosters within schools⁷, I dropped seven teachers who left the study before the beginning of the 2012-13 school year for reasons unrelated to the experiment (i.e., leaving the district or teaching, maternity leave, change in teaching assignment); 11 teachers who only were part of the study in the third year and, therefore, did not have the necessary data from prior years to calculate non-experimental teacher effects on students' attitudes and behaviors; and seven teachers whose random assignment partner left the study for either of the two reasons above.⁸

⁷ Two other teachers from the same randomization block also agreed to participate. However, the principal decided that it was not possible to randomly assign rosters to these teachers. Thus, I exclude them from all analyses.

⁸ One concern with dropping teachers in this way is that they may differ from other teachers on post-randomization outcomes, which could bias results. Comparing attriters for whom I had post-randomization data ($N = 21$, which excludes the four teachers who either left teaching, left the district, moved to third grade and therefore out of my dataset, or were on maternity leave) to the remaining teachers ($N = 54$) on their observed effectiveness at raising students' math achievement in the 2012-13 school year, I found no difference ($p = .899$). Further, to ensure strong external validity, I compared attriters to the experimental sample on each of the teacher characteristics listed in Table 1 and found no difference on any.

Next, I restricted the remaining sample to randomization blocks with low levels of non-compliance amongst participating students. Here, non-compliance refers to the fact that some students switched out of their randomly assigned teacher's classroom. Other studies that exploit random assignment between teachers and students have accounted for this form of non-compliance through instrumental variables estimation and calculation of treatment on the treated (Bacher-Hicks et al. 2017; Glazerman and Protik 2015; Kane et al. 2013). However, this approach was not possible in this study, given that students who transferred out of an NCTE teacher's classroom no longer had survey data to calculate teacher effects on these outcomes. Further, I would have needed to have prior student survey responses for these students' actual teachers, which I did not. In total, 28% of students moved out of their randomly assigned teachers' classroom (see Appendix Table 2 for information on reasons for and patterns of non-compliance). At the same time, non-compliance was nested within a small subset of six randomization blocks. In these blocks, rates of non-compliance ranged from 40% to 82% due primarily to principals and school leaders who made changes to the originally constructed class rosters. By eliminating these blocks, I am able to focus on a sample with a much lower rate of non-compliance (11%) and where patterns of non-compliance are much more typical. The remaining 18 blocks had a total of 67 non-compliers and an average rate of non-compliance of 9% per block; three randomization blocks had full compliance.

In Table 2, I confirm the success of the randomization process among the teachers in my final analytic sample ($N = 41$) and the students on their randomly assigned rosters ($N = 598$).⁹ In a traditional experiment, one can examine balance at baseline by calculating differences in average student characteristics between the treatment and control groups. In this context, though,

⁹ Thirty-eight students were hand placed in these teachers' classrooms after the random assignment process. As these students were not part of the experiment, they were excluded from all analyses.

treatment consisted of multiple possible teachers within a given randomization block. Thus, to examine balance, I examined the relationship between the assigned teacher's predicted effectiveness at improving students' state math test scores in years prior to the experiment and baseline student characteristics. Specifically, I regressed these teacher effect estimates on a vector of observable student characteristics and fixed effects for randomization block. As expected, observable student characteristics were not related to teacher effects on state math tests, either tested individually or as a group ($p = .808$), supporting the fidelity of the randomization process. Even though this sample includes some non-compliers, these students looked similar to compliers on observable baseline characteristics, as well as the observed effectiveness of their randomly assigned teacher at improving state math test scores in years prior to random assignment (see Appendix Table 3).¹⁰ As such, I am less concerned about having to drop the few non-compliers left in my sample from all subsequent analyses.

5. Empirical Strategy

For all analyses, I began with the following model of student production:

$$OUTCOME_{idsajt} = \alpha f(A_{it-1}) + \zeta OUTCOME_{it-1} + \pi X_{it} + \varphi \bar{X}_{it}^c + \varphi \bar{X}_{it}^s + \varepsilon_{idsajt} \quad (1)$$

$OUTCOME_{idsajt}$ was used interchangeably for each survey construct – i.e., *Behavior in Class*, *Self-Efficacy in Math*, and *Happiness in Class* – for student i in district d , school s , grade g taught by teacher j in year t . I also specified models that use students' math achievement as an outcome as a point of comparison. Throughout the paper, I test a variety of alternative models that include different combinations of control variables. The full set of controls includes a cubic

¹⁰ Twenty-six students were missing baseline data on at least one characteristic. In order to retain all students, I imputed missing data to the mean of the students' randomization block. I take the same approach to missing data in all subsequent analyses. This includes the 19 students who were part of my main analytic sample but happened to be absent on the day that project managers administered the student survey and, thus, were missing outcome data. This approach to imputation seems reasonable given that there was no reason to believe that students were absent on purpose to avoid taking the survey.

function of students' prior academic achievement, A_{it-1} , in both math and reading; a prior measure of the outcome variable, $OUTCOME_{it-1}$; student demographic characteristics, X_{it} , including gender, race, free or reduced-price lunch eligibility, special education status, and limited English proficiency; these same test-score variables and demographic characteristics averaged to the class level, \bar{X}_{it}^c , and to the school level, \bar{X}_{it}^s ; and school fixed effects, σ_s , which replace school characteristics in some models.

To generate teacher effect estimates, $\hat{\tau}_{jt}^S$, I took two approaches, each with strengths and limitations. First, I calculated teacher effects by fitting Equation (1) using ordinary least squares (OLS) regression and then averaging student-level residuals to the teacher level. I did so separately for each outcome measure, as well as with several different model specifications denoted by the superscript, S . This approach is intuitive, as it creates an estimate of the contribution of teachers to student outcomes above and beyond factors already controlled for in the model. It also is computationally simple.¹¹ At the same time, measurement error in these estimates due to small class sizes, sampling idiosyncrasies, measurement error in students' survey responses, etc. may lead me to overstate the variance of true teacher effects; it also could attenuate the relationship between different measures of teacher effectiveness (e.g., measures at two points in time), even if they capture the same underlying construct.

Therefore, I also calculated a form of empirical Bayes estimates that take into account measurement error and shrink teacher effects back toward the mean based on their precision. To do so, I included a teacher-level random effect in the model, which I fit using restricted maximum likelihood. This approach is similar to a two-step approach described by others (e.g.,

¹¹ An alternative fixed-effects specification is preferred by some because it does not assume that teacher assignment is uncorrelated with factors that predict student outcomes (Guarino et al. 2015). However, in these data, this approach returned similar estimates in models where it was feasible to include teacher fixed effects in addition to the other set of control variables, with correlations of 0.99 or above (see Blazar and Kraft 2017 for more details).

Chetty, Friedman, and Rockoff 2014a; Guarino et al. 2015; Kane et al. 2013) that first calculates the unshrunk teacher effects using OLS and then multiplies these by a shrinkage factor. The shrinkage factor generally is calculated by looking at variation in teacher effects within teachers and across classrooms. It was not possible to use this two-step approach here given that data from multiple classrooms from the same teacher was not available in the experimental portion of the study; elementary teachers in the sample all worked with just one class in a given year. Instead, the one-step random effects approach I use shrinks estimates back toward the mean based on the variance of the observed data (Raudenbush and Bryk 2002). While shrinking teacher effects is commonplace in both research and policy (Koedel, Mihaly, and Rockoff 2015), theory and simulated analyses show that shrunken estimates are biased downward relative to the size of the measurement error (Jacob and Lefgren 2005). I refer to these two sets of estimates as “unshrunk” and “shrunken” teacher effects.

I utilized these teacher effect estimates for three subsequent analyses. First, I estimated the variance of $\hat{\tau}_{jt}^S$ in order to examine whether teachers vary in their contributions to students’ attitudes and behaviors. I focused on the experimental sample in order to be assured that estimates were not biased by non-random sorting. Given that the variance of true teacher effects is bounded between the unshrunk and shrunken estimates (Raudenbush and Bryk 2002), I present both. The latter are model-based estimates reported directly from the random effects model. I also examined whether teachers who improve one student outcome were equally effective at improving others by calculating pairwise correlations between these teacher effect estimates.

Second, I examined the sensitivity of $\hat{\tau}_{jt}^S$ to different model specifications. I began with a baseline model that calculated teacher effects controlling only for students’ prior academic

achievement, as this is the measure typically used to account for non-random sorting when test scores are the outcome of interest (Chetty, Friedman, and Rockoff 2014a; Kane et al. 2013; Kane and Staiger 2008). I also considered a model that conditioned estimates on a lagged measure of students' survey response, which is a more direct analog of the value-added approach by looking at gains in student outcomes. Additional variations of these models include ones that control for student, class, or school characteristics. Finally, in order to address concerns about "reference bias" in self-reported measures (Duckworth and Yeager 2015; West et al. 2016), I replaced school characteristics with school fixed effects. By making within-school comparisons, I am able to difference out school-level factors including norms around behavior or engagement that can create an implicit standard of comparison that students use when they judge their own behavior or engagement. It was not possible to run these analyses in the experimental sample, given that only a small subset of students and teachers in that sample had lagged survey measures. There also was no guarantee that a teacher who had students with prior survey measures had a randomization block partner whose students had these measures. Instead, I focused on the balanced sample of teachers and students with all possible control variables from the larger observational dataset.

In my third and final set of analyses, I examined whether non-experimental teacher effect estimates calculated in years prior to 2012-13 predicted student outcomes following random assignment. The randomized design allowed for a straightforward analytic model:

$$OUTCOME_{ijsg2012-13} = \delta \hat{\tau}_{jt < 2012-13}^S + v_{sg} + \epsilon_{ijsgt} \quad (2)$$

$OUTCOME_{ijsg2012-13}$ was used interchangeably for each outcome measure for student i in teacher j 's classroom in the 2012-13 school year. I predicted these measures in the random assignment year with predicted, non-experimental teacher effect estimates, $\hat{\tau}_{jt < 2012-13}^S$. That is,

when *Behavior in Class* is the outcome of interest, $\hat{\tau}_{jt < 2012-13}^S$, represents a non-experimental estimate of teachers' effectiveness at improving students' *Behavior in Class* in prior years; when *Self-Efficacy in Math* is the outcome of interest, $\hat{\tau}_{jt < 2012-13}^S$, represents a non-experimental estimate of teachers' effectiveness at improving *Self-Efficacy in Math* in prior years. Following the research design, I included fixed effects for each randomization block, ν_{sg} . In order to increase the precision of my estimates, I calculated non-experimental teacher effects using all available teacher-years prior to the experiment. For the same reason, in Equation (2) I also controlled for students' prior achievement, demographic characteristics, and class characteristics captured from the randomly assigned rosters. I clustered standard errors at the class level to account for the nested structure of the data.

My parameter of interest is δ , which describes the relationship between non-experimental teacher effect estimates and current student outcomes. As in Kane and Staiger (2008), I examined whether these estimates had any predictive validity (i.e., whether they were statistically significantly different from 0 SD) and whether they contained some degree of bias (i.e., whether they were statistically significantly different from 1 SD).

6. Results

6.1. *Experimental Teacher Effects on Students' Attitudes and Behaviors*

In Table 3a, I present results describing the extent to which teachers vary in their contribution to students' attitudes and behaviors, as well as their math achievement. Estimates represent the standard deviation of the teacher-level variance, with Panel A and Panel B presenting unshrunk and shrunken estimates, respectively. In Table 3b, I present correlations between corresponding unshrunk and shrunken estimates. All models focus on the experimental sample in which teachers were randomly assigned to class rosters within schools,

and therefore include randomization block (i.e., school-by-grade) fixed effects to match this design. Model 1 includes no additional controls, while Model 2 adds students' prior achievement in math and reading, which is standard practice when estimating teacher effects. In Model 3, I add additional student characteristics, as well as class characteristics that aim to remove the contribution of peer effects from the teacher effect estimates. It was not possible to model classroom-level shocks directly, as random assignment data were not available over multiple classes or school years. Class characteristics describe the set of students included on the randomly assigned rosters rather than the students who ultimately stayed in that classroom.

I begin by describing the magnitude of teacher effects on students' math performance on state tests, which have been well documented in the academic literature (for a review, see Hanushek and Rivkin 2010) and thus provide a point of comparison for teacher effects on students' attitudes and behaviors. I find that a 1 SD increase in teacher effectiveness is equivalent to between a 0.13 SD and 0.28 SD increase in students' math achievement. Results are fairly similar between the corresponding unshrunk and shrunken estimates, particularly when controlling for students' prior achievement. In Models 2 and 3, the magnitude of the teacher-level variation for these unshrunk and shrunken estimates are almost identical to two decimal places, and correlations between them range from 0.84 to 0.95. These results are quite similar to those found by Guarino et al. (2015), who argued that the "effect of shrinkage itself does not appear to be practically important for properly ranking teachers or to ameliorate the performance of the [unshrunk] estimator" (p. 212). While I do not find large differences between shrunken and unshrunk estimates, I do observe that the variance of teacher effects is substantively larger in Model 1 (0.28 and 0.22 SD for unshrunk and shrunken estimates, respectively), which only controls for randomization block fixed effects, than in Model 3 (0.13 SD for both unshrunk

and shrunken estimates), which also controls for student and class characteristics. This is consistent with other literature suggesting that controlling for observable class or peer characteristics produces a conservative estimate of the magnitude of teacher effects on student test scores (Kane et al. 2013; Thompson, Guarino, and Wooldridge 2015). In Model 3, both shrunken and unshrunken teacher effect estimates of 0.13 SD indicate that, relative to an average teacher, teachers at the 84th percentile of the distribution of effectiveness move the medium student up to roughly the 55th percentile of math achievement.

I also find that teachers have substantive impacts on self-reported measures of students' attitudes and behaviors. The largest of these teacher effects is on students' *Happiness in Class*, where a 1 SD increase in teacher effectiveness leads to a roughly 0.30 SD increase in this outcome. Similar to teacher effects on students' math performance, results for teacher effects on students' *Happiness in Class* are fairly consistent between Panel A and Panel B, indicating that shrinkage does not necessarily boost performance. Correlations between unshrunken and shrunken estimates range from 0.87 to 0.95. For the unshrunken teacher effects, estimates are smaller when controlling for student and class characteristics in Models 3 (0.26 SD) compared to estimates from the other two models (0.35 SD). This is not the case for the shrunken teacher effects, where estimates across all three models are roughly 0.33 SD. In Model 3, the variance of the unshrunken teacher effects on students' *Happiness in Class* is slightly larger than the variance of the analogous shrunken estimates. This is possible given that, as described above, the shrunken estimates are not derived by directly shrinking the unshrunken estimates. Rather, these estimates come from a separate model that includes a teacher-level random effect and generates model-based estimates of the variance component.

The evidence also points to sizeable teacher effects on students' *Behavior in Class* and *Self-Efficacy in Math*, though results are less consistent between unshrunk and shrunk estimators. Without shrinkage, the magnitude of teacher effects on these two outcomes generally is larger than teacher effects on students' math performance but smaller than teacher effects on students' *Happiness in Class*: between 0.14 SD and 0.28 SD for teacher effects on students' *Behavior in Class*, and between 0.19 SD and 0.29 SD for teacher effects on students' *Self-Efficacy in Math*. Shrunk estimates are considerably smaller. For example, in Model 3, these shrunk estimates are 0.05 SD for teacher effects on *Behavior in Class* and 0.08 SD for teacher effects on *Self-Efficacy in Math*. Correlations between the unshrunk and shrunk estimates still are strong, but never above 0.87. Models that exclude class characteristics and use shrinkage to calculate teacher effects on students' *Self-Efficacy in Math* produce estimates close to 0 SD. For this reason, I exclude from Table 3b correlations between unshrunk and shrunk estimates for this outcome and these models.

It is counterintuitive that models that include class characteristics produce estimates that are larger than those that exclude these control variables. It is possible that the error structure for students' self-reported *Self-Efficacy in Math* is quite different from the error structure for other measures, which in turn leads to challenges when implementing shrinkage through a random effects model fit using restricted maximum likelihood estimation. Although restricted maximum likelihood aims to address concerns that full maximum likelihood tends to produce variance estimates that are biased downward, this may also be a concern in the relatively small sample of teachers and students (Harville 1977; Raudenbush and Bryk 2002). Mixed models can result in singular fits (i.e., variance-covariance components that are exactly zero) in several instances,

including small number of random effects and complex random effects models (Gelman, 2006). This topic is beyond the scope of this paper but is an important one for future research.

In Table 3c, I present a correlation matrix of teacher effects on different student outcomes. Here, teacher effects come from Model 3 (see Tables 3a and 3b), which controls for prior achievement, student characteristics, and class characteristics. I focus on this model given that shrunken and unshrunken teacher effects are greater than 0 SD for all outcomes. One concern when estimating relationships between different measures of teacher quality is that individual teacher effect estimates are measured with error, which will attenuate these correlations (Spearman 1904). Indeed, correlations between the unshrunken teacher effect estimates in Panel A generally are larger than correlations between the shrunken ones in Panel B; by design, shrinkage attempts to minimize measurement error in individual teacher effect estimates. At the same time, differences in correlations between these two panels are not large, suggesting that additional approaches to address attenuation due to measurement error are unlikely to change overall patterns of results.

The largest of these correlations is between teacher effects on different measures of students' attitudes and behaviors. For example, teacher effects on students' *Self-Efficacy in Math* and teacher effects on the other two non-tested outcomes fall between 0.44 and 0.65. However, teachers do not appear to be equally effective at improving all three attitudes and behaviors. The correlation between teacher effects on students' *Happiness in Class* and students' *Behavior in Class* is weak and non-significant. Correlations between teacher effects on students' attitudes and behaviors versus effects on students' math achievement are similarly weak; most are not statistically significant. One exception is the relationship between teacher effects on students' math performance and teacher effects on students' *Happiness in Class*, which is negative and

statistically significantly correlated when using the shrunken estimates ($r = -0.38$). This suggests that teachers who are skilled at improving students' math achievement may do so in ways that make students less happy or less engaged in class.

Overall, these findings provide strong evidence that teachers impact several student attitudes and behaviors in addition to their academic performance. Weak, non-significant correlations between many of these teacher effect estimates indicate that these measures identify unique skills that teachers bring to and engage in in the classroom.

6.2. *Sensitivity of Teacher Effects Across Model Specifications*

In Tables 4a and 4b, I present results describing the relationship between teacher effects on students' attitudes and behaviors across model specifications. Panel A shows correlations for unshrunken estimates, while Panel B shows correlations for shrunken estimates. Because patterns of results are quite similar for the unshrunken and shrunken estimates, I focus my discussion on the latter for simplicity. This analysis includes the balanced sample of teachers and students with all possible control variables from the larger observational dataset. In Appendix Table 4, I present the magnitude of the teacher-level variation on all student outcomes using this sample, and find that results are similar to those presented in Table 3a using the experimental sample.

In the first of these tables (Table 4a), I examine the correlations between teacher effects on students' attitudes and behaviors that control for prior achievement (Model 1), a prior measure of the survey outcome (Model 2), or both (Model 3). Because this analysis examines the sensitivity of teacher effects to inclusion or exclusion of lagged measures of the outcome variable, I focus only on teacher effects on the three measures of students' attitudes and behaviors; for teacher effects on students' math performance used elsewhere in the paper, all

models control for lagged achievement. Here, I find correlations of teacher effects across model specifications above 0.86. As expected, the smallest of these correlations describe the relationship between teacher effects that control either for prior achievement (Model 1) or for students' prior survey responses (Model 2): 0.90 for teacher effects on *Behavior in Class*, 0.86 for *Self-Efficacy in Math*, and 0.96 for *Happiness in Class*. However, these correlations still are quite strong. Correlations between teacher effects from models that have overlapping sets of controls (i.e., between Models 1 and 3 or between Models 2 and 3) are stronger, between 0.90 and 0.99. This suggests that teacher effects on these attitudes and behaviors are not particularly sensitive to inclusion of prior achievement or prior survey responses. In light of these findings, I exclude prior measures of students' attitudes and behaviors from most subsequent analyses, allowing me to retain the largest possible sample of teachers and students.

Next, I examine the sensitivity of teacher effects from this baseline model (Model 1) to models that control for additional student, class, or school characteristics (see Table 4b). In the table, empty cells indicate instances where the teacher-level variation is close to 0 SD (i.e., for shrunken teacher effects on students' *Self-Efficacy in Math* generated from Models 5 through 7; see Appendix Table 4). I find that teacher effects on students' math performance and on the three measures of students' attitudes and behaviors are not sensitive to student demographic characteristics but are sensitive to additional control variables. Correlations between teacher effect estimates from Model 1 (which controls for prior test scores) and from Model 4 (which builds on Model 1 by adding student demographic characteristics) all are greater than or equal to 0.95. For teacher effects on students' math performance, *Behavior in Class*, and *Self-Efficacy in Math*, correlations between estimates from Model 1 and from Model 5 (which builds on previous models by adding classroom characteristics) are substantively smaller, at 0.76, 0.69, and 0.82,

respectively. For teacher effects on students' *Happiness in Class*, the correlation stays above 0.90.

Adding school characteristics to teacher effect specifications appears to have the largest impact on teacher rankings. Correlations between estimates from Model 1 and from Model 6 (which builds on previous models by adding observable school characteristics) range from 0.54 (for teacher effects on students' math performance) to 0.71 (for teacher effects on students' *Happiness in Class*). Correlations between estimates from Model 1 and from Model 7 (which replaces observable school characteristics with school fixed effects) range from 0.41 (for teacher effects on students' *Behavior in Class*) to 0.66 (for teacher effects on students' *Happiness in Class*). Correlations between estimates from Models 6 and 7 (not shown in Table 4b) are 0.94, 0.70, 0.71, and 0.92 for teacher effects on students' math performance, *Behavior in Class*, *Self-Efficacy in Math*, and *Happiness in Class*, respectively. Reference bias is one possible explanation for lower correlations in models that do and do not control for school fixed effects. At the same time, these estimates are well within the range reported in studies looking at the sensitivity of teacher effects on test scores across models that control for school characteristics or school fixed effects, between roughly 0.5 and 0.9 (Aaronson, Barrow, and Sander 2007; Goldhaber and Theobald 2012; Hill, Kapitula, and Umland 2011).

6.3. Predictive Validity of Non-Experimental Teacher Effects

In Table 5, I report estimates describing the relationship between non-experimental teacher effects on student outcomes and these same measures following random assignment. Cells contain estimates from separate regression models where the dependent variable is the student attitude or behavior listed in each column. The independent variable of interest is the non-experimental teacher effect on this same outcome estimated in years prior to random

assignment. All models include fixed effects for randomization block to match the experimental design. In order to increase the precision of my estimates, models also control for students' prior achievement in math and reading, student demographic characteristics, and classroom characteristics from randomly assigned rosters. In Panel A, non-experimental teacher effects are unshrunk estimates, while in Panel B these non-experimental teacher effects are shrunk. Non-experimental teacher effects are modeled from five separate equations discussed above, each with different sets of covariates. I exclude teacher effects calculated from Models 2 and 3 (described in Table 4a), both of which controlled for prior measures of students' attitudes and behaviors that were not available for many teachers' students in the experimental portion of the study. However, below I describe results from additional analyses that estimate the relationship between non-experimental teacher effects that do control for imputed lagged survey measures; results are consistent with main results. Stars indicate whether point estimates are statistically significantly different from 0 SD, while p -values testing the null hypothesis that effect sizes are equal to 1 SD are presented next to each estimate. The sample sizes for *Happiness in Class* is reduced by one teacher who did not have non-experimental teacher effects on this outcome.

Validity evidence for teacher effects on students' math performance are consistent with other experimental studies (Kane et al. 2013; Kane and Staiger 2008), where predicted differences in teacher effectiveness in observational data come close to actual differences following random assignment of teachers to classes. The non-experimental teacher effect estimate that comes closest to a 1:1 relationship is the shrunk estimate that controls for students' prior achievement and other demographic characteristics (0.995 SD). Despite a relatively small sample of teachers, the standard error for this estimate (0.084) is substantively smaller than those in other studies – including the meta-analysis conducted by Bacher-Hicks et

al. (2017) – and allows me to rule out relatively large degrees of bias in teacher effects calculated from this model. A likely explanation for greater precision in this study relative to others is the fact that other studies generate estimates through instrumental variables estimation to calculate treatment on the treated. Instead, I use OLS regression and account for non-compliance by narrowing in on randomization blocks in which very few, in any, students moved out of their randomly assigned teachers' classroom. Non-experimental teacher effects calculated without shrinkage are related less strongly to current student outcomes, though differences in estimates and associated standard errors between Panel A and Panel B are not large. All corresponding estimates (e.g., Model 1 from Panel A versus Panel B) have overlapping 95% confidence intervals.

Results examining forecast bias in teacher effects on students' *Behavior in Class* are not substantively different from what I would expect based on the math test-score outcome. I find that teacher effects with the best predictive validity are the shrunken estimates from Model 1, which calculates non-experimental teacher effects only controlling for students' prior achievement.¹² Here, I find an estimate of 1.00 SD that matches the hypothesis described by Kane and Staiger (2008) where predicted differences across classroom should equal observed differences. However, the standard error around this estimate is substantively larger (0.25) than the standard error for test-score estimates.

¹² Results in Panel B suggest that adding class- and school-level controls to the model calculating non-experimental teacher effects on students' *Behavior in Class* may in fact add bias. Point estimates describing the relationship between current student outcomes and non-experimental teacher effects calculated from Models 3 through 5 all are greater than 1.2 SD. These patterns are somewhat consistent with findings from the MET project, in which researchers suggested that some models "over control," resulting in removal of peer effects that actually predict important differences in teacher performance (Kane et al. 2013). This explanation makes sense here as well, where teachers' ability to improve individual students' behavior likely is closely related to the control they have over peer-to-peer relationships. At the same time, I do not want to place too much emphasis on these differences across models, given that standard errors are large and, thus, point estimates have overlapping 95% confidence intervals.

Comparison of estimates between Panel A and Panel B provide some insight here and, in particular, the tradeoff between accuracy and precision. In Panel B, estimates relating non-experimental, shrunken teacher effect estimates on students' *Behavior in Class* to current student outcomes are notably larger than estimates in Panel A relating unshrunk estimates to current outcomes. This makes sense, as shrunken estimates are adjusted for the amount of measurement error the unshrunk estimates contain. Measurement error will attenuate the relationship between two teacher effect estimates, even if the true relationship is equal to 1 SD. Indeed, earlier in the paper, I showed that teacher effects on students' *Behavior in Class* generally underwent more shrinkage than teacher effects on students' math test scores (see Table 3a). At the same time, relationships between shrunken teacher effect estimates and current student outcomes in Panel B are measured with considerably less precision than relationships drawing on unshrunk teacher effect estimates in Panel A. Standard errors in Panel B are roughly two to three times as large as those in Panel A. This also makes sense, as shrunken estimates provide a lower bound on the variation of true teacher effects, particularly when measurement error is large (Jacob and Lefgren 2005); decreased variation in the independent variable decreases statistical power. Considering results from Panel A and Panel B jointly provides evidence that non-experimental methods for estimating teacher effects on students' *Behavior in Class* account for a large degree of bias due to non-random sorting and factors beyond teachers' control.

For both *Self-Efficacy in Math* and *Happiness in Class*, non-experimental teacher effect estimates have moderate predictive validity. Generally, I can distinguish estimates from 0 SD, indicating that they contain some information content on teachers. The exception is shrunken estimates for *Self-Efficacy in Math*. Although estimates are similar in magnitude to the unshrunk estimates in Panel A, between 0.42 SD and 0.58 SD, standard errors are large and

95% confidence intervals cross 0 SD. I also can distinguish many estimates from 1 SD. This indicates that non-experimental teacher effects on students' *Self-Efficacy in Math* and *Happiness in Class* contain potentially large and important degrees of bias. For both measures of teacher effectiveness, point estimates around 0.5 SD suggest that they contain roughly 50% bias.

One concern with these results is that non-experimental teacher effects do not control for prior survey responses, and such a model might reduce bias. Earlier in the paper, I show that in the observational portion of the study teacher effects that control for some combination of prior achievement and/or prior survey responses do not return markedly different teacher rankings. However, in some instances correlations are lower than 0.90 (see Table 4a), leaving open the possibility that controlling for lagged survey responses may boost performance. To address this concern in the experimental data, I conduct a robustness check that calculates non-experimental teacher effects after imputing students' lagged survey responses. To impute, I fit a regression model that predicts students' lagged survey responses with all other available data (i.e., prior test scores in math and reading, and the student demographic characteristics listed in Table 2) for the sample of students with these data used elsewhere in this paper.¹³ Then, I use this model to infer predicted values for all students without lagged survey responses. Finally, I calculate non-experimental teacher effects on students' attitudes and behaviors controlling for these lagged measures, an indicator for whether or not the lagged survey measure was imputed, and students' prior test scores in some instances. As in Table 4b, Model 2 calculates teacher effects controlling for students' prior survey response, while Model 3 calculates teacher effects controlling for prior

¹³ Students' prior test scores are statistically significant predictors of all three measures of students' attitudes and behavior, as are several demographic characteristics. Together, prior achievement and demographic characteristics explain a sizeable amount of the variation in prior survey responses: 22% for *Behavior in Class*, 10% for *Self-Efficacy in Math*, and 7% for *Happiness in Class*. Further, imputed measures of students' prior survey responses are strongly related to current survey responses, with standardized regression coefficients between 0.43 SD (*Happiness in Class*) and 0.63 SD (*Behavior in Class*).

achievement and prior survey response. I present results that relate these non-experimental teacher effects to student outcomes following random assignment in Appendix Table 5. Patterns of results are very similar to those presented in Table 5, suggesting that controlling for lagged measures of the outcome variable when calculating non-experimental teacher effects on students' attitudes and behaviors does not appear to change inferences regarding bias in these measures. These results are similar to those from the quasi-experimental validation study by Backes and Hansen (2016), where the authors controlled for prior measures of their non-tested outcomes in all models and still found large degrees of bias in some instances.

7. Discussion and Conclusion

Where does this leave policy, practice, and research? Should these measures be used in policy settings, despite concerns about bias? This is not an easy question to answer. For some, the relationships presented in this paper could point to considerable policy usefulness for the non-experimental estimates. If one were to rank teachers using experimental and non-experimental estimates, results would be similar. Thus, ignoring reference bias problems and possible gaming, a teacher de-selection policy using biased measures would still improve outcomes on average.

Another possible reason to incorporate measures of students' attitudes and behaviors and teachers' ability to improve them into selection and accountability policy, in spite of bias, would be to create clear incentives for improving these skills in school. Many, including myself, see students' social and emotional development as a central goal of teachers' and schools' work (e.g., Durlak et al. 2011; Farrington et al. 2012; Pianta and Hamre 2009). Yet, accountability systems that focus predominantly or exclusively on student achievement send a message that the skills captured on these tests are the ones that policymakers want students to have when they

leave school. Broadening what it means to be a successful student and “making the development of the whole child central to the mission of education” (Garcia 2014, p. 4) clearly is good policy.

At the same time, lessons learned from new teacher evaluation systems that incorporate teacher effects on students’ test scores highlight several reasons why making high-stakes policy decisions based on teacher effects on students’ attitudes and behaviors may not be appropriate or advantageous. Despite convincing evidence against bias in teacher effects on students’ academic performance, teachers still are skeptical about their use and the fairness of these measures (Jiang, Sporte, and Luppescu 2015). One reason for this skepticism discussed in the academic literature is that, even if teacher effects are unbiased, they often are quite noisy measures of teachers’ effectiveness (Ballou and Springer 2015). Large confidence intervals around individual teachers’ scores – due to the number of students attached to that teacher and error in the student-level assessment itself – means that a teacher’s underlying ability often is statistically indistinguishable from others’. This likely would be an even greater issue for teacher effects on students’ attitudes and behaviors given well-documented concerns about error in student- and even teacher- or parent-reports of these measures (Duckworth and Yeager 2015). Even if systems were to incorporate classical measurement error into teachers’ effectiveness ratings, which most do not, there still would be lingering concerns about other sources of error. In particular, there are bound to be concerns about cheating (Campbell 1977; Koretz 2008). In this study, student surveys were administered under low-stakes conditions where student responses were not visible to the teacher or other students in the classroom. It is possible that estimates of bias might differ – likely to increase – under high-stakes settings where survey responses could be coached or influenced by other pressures.

Despite concern about using teacher effects on students' attitudes and behaviors in high-stakes policy settings, I believe that there are other uses of these measures that fall within and would enhance existing school practices. In particular, measures of teachers' effectiveness at improving students' attitudes and behaviors could be used to identify areas for professional growth and connect teachers with targeted professional development. Bringing costly but effective development programs such as teacher coaching (Kraft, Blazar, and Hogan forthcoming) to scale require at least two key pieces of information that measures such as those used in this study could provide. First, it would be useful to know which teachers require immediate support in order to allocate professional development dollars to these teachers, as opposed to investing in lower-cost but less-effective programs that reach all teachers. Second, the individualized nature of coaching and related development programs requires that school leaders know teachers' individual strengths and weaknesses in order to facilitate appropriate teacher-coach or teacher-team matches where members have complementary skill sets (Papay et al. 2016). Observation rubrics provide one source of data for this purpose, yet have logistical constraints including needing school leaders who have the time and knowledge to assess multiple teachers on multiple teaching skills (Hill and Grossman 2013). In light of moderate to strong relationships between teachers' observed classroom behaviors captured on established observation rubrics and teacher effects on several student attitudes and behaviors (Blazar and Kraft 2017), it is possible that the latter could be used as a lower-cost proxy for the former. In these instances, biased measures are less likely to be a concern than in settings where teachers' jobs are on the line.

Finally, supporting teachers in the work of developing students' attitudes and behaviors will require investments in research in addition to changes in policy and practice. Based on

experimental studies of teacher effects on student achievement, newer research is starting to examine related questions (e.g., how instructional supports impact students' achievement; Kane et al. 2016) using observational and value-added approaches that generally are less expensive and considerably more tractable than randomized control trials. Making this important decision in the context of research on students' attitudes and behaviors will require close consideration of the tradeoffs between two key issues: bias and availability of data. The analyses presented here suggest that value-added approaches likely will reduce some but not all of the sorting bias that could influence estimates of the impact of different inputs on measures of students' behavior, self-efficacy, and happiness. Even if some degree of bias remains, this approach likely would improve upon much of the existing body of research to date that lacks convincing evidence about what works in education (Kane 2015; Murnane and Willett 2011). At the same time, such studies would not have the benefit of easily accessible administrative data that has made this type of work possible when examining gains in student achievement outcomes. Building up administrative datasets that include rich measures of students' attitudes and behaviors in addition to their academic performance is, in my opinion, a worthy goal (see West 2016 for how this is starting to happen in some education agencies including the CORE districts in California). Until that happens, though, researchers will need to continue to collect these measures themselves. In turn, we likely will want to conduct more and learn as much as possible from random assignment studies.

References

- Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics* 25(1): 95-135.
- Bacher-Hicks, Andrew, Mark Chin, Thomas J. Kane, and Douglas O. Staiger. 2017. An evaluation of bias in three measures of teacher quality: Value-added, classroom observations, and student surveys. NBER Working Paper 23478.
- Bacher-Hicks, Andrew, Thomas J. Kane, and Douglas O. Staiger. 2014. Validating teacher effect estimates using changes in teacher assignments in Los Angeles. NBER Working Paper 20657.
- Backes, Ben, and Michael Hansen. 2015. Teach for America impact estimates on nontested student outcomes. National Center for Analysis of Longitudinal in Education Research Working Paper 146.
- Ballou, Dale, and Matthew G. Springer. 2015. Using student test scores to measure teacher performance some problems in the design and implementation of evaluation systems. *Educational Researcher* 44(2): 77-86.
- Blazar, David, and Matthew A. Kraft. 2017. Teacher and teaching effects on students' academic behaviors and mindsets. *Educational Evaluation and Policy Analysis* 29(1): 146-170.
- Blazar, David, Erica Litke, and Johanna Barmore. 2016. What does it mean to be ranked a "high" or "low" value-added teacher? Observing differences in instructional quality across districts. *American Educational Research Journal* 53(2): 324-359.
- Campbell, Donald T. 1979. Assessing the impact of planned social change. *Evaluation and Program Planning* 2(1): 67-90.

- Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics* 126(4): 1593-1660.
- Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014a. Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review* 104(9): 2593-2632.
- . 2014b. Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review* 104(9): 2633-2679.
- Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources* 41(4):778-820.
- Duckworth, Angela L., and David Scott Yeager. 2015. Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher* 44(4): 237-251.
- Durlak, Joseph A., Roger P. Weissberg, Allison B. Dymnicki, Rebecca D. Taylor, and Kriston B. Schellinger. 2011. The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development* 82(1): 405-432.
- Farrington, Camille A., Melissa Roderick, Elaine Allensworth, Jenny Nagaoka, Tasha Seneca Keyes, David W. Johnson, and Nicole O. Beechum. 2012. *Teaching adolescents to become learners: The role of noncognitive factors in shaping school performance – A critical literature review*. Consortium on Chicago School Research.

- Garcia, Emma. 2014. The need to address noncognitive skills in the education policy agenda. Briefing Paper # 386. Economic Policy Institute.
- Gelman, Andrew. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1(3): 515-534.
- Gershenson, Seth. 2016. Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy* 11(2): 125-149.
- Glazerman, Steven, and Ali Protik. 2015. Validating value-added measures of teacher performance. Working Paper.
- Goldhaber, Dan, and Roddy Theobald. 2012. Do different value-added models tell us the same things? Carnegie Knowledge Network.
- Guarino, Cassandra M., Michelle Maxfield, Mark D. Reckase, Paul N. Thompson, and Jeffrey M. Wooldridge. 2015. An evaluation of Empirical Bayes' estimation of value-added teacher performance measures. *Journal of Educational and Behavioral Statistics* 40(2): 190-222.
- Hanushek, E. A., John F. Kain, Jacob M. Markman, and Steven G. Rivkin. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics*, 18(5), 527–544.
- Hanushek, Eric A., and Steven G. Rivkin. 2010. Generalizations about using value-added measures of teacher quality. *The American Economic Review*, 100(2): 267-271.
- Harville, David A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* 72(358): 320-338.
- Hill, Heather C., David Blazar, and Kathleen Lynch. 2015. Resources for teaching: Examining personal and institutional predictors of high-quality instruction. *AERA Open* 1(4): 1-23.

- Hill, Heather, and Pam Grossman. 2013. Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review* 83(2): 371-384.
- Hill, Heather C., Laura Kapitula, and Kristin Umland. 2011. A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal* 48(3): 794-831.
- Jackson, C. Kirabo. 2012. Non-cognitive ability, test scores, and teacher quality: Evidence from ninth grade teachers in North Carolina. NBER Working Paper 18624.
- Jackson, C. Kirabo. 2016. What do test scores miss? The importance of teacher effects on non-test score outcomes. NBER Working Paper 22226.
- Jacob, Brian, and Lars Lefgren. 2005. Principals as agents: Subjective performance assessment in education. NBER Working Paper 11463.
- Jennings, Jennifer L., and Thomas A. DiPrete. 2010. Teacher effects on social and behavioral skills in early elementary school. *Sociology of Education* 83(2): 135-159.
- Jiang, Jennie Y., Susan E. Spote, and Stuart Luppescu. 2015. Teacher perspectives on evaluation reform Chicago's REACH students. *Educational Researcher* 44(2): 105-116.
- Kane, Thomas J. 2015. Frustrated with the pace of progress in education? Invest in better evidence. The Brookings Institution.
- Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. Have we identified effective teachers? Validating measures of effective teaching using random assignment. Bill and Melinda Gates Foundation.

- Kane, Thomas J., Antoniya M. Owens, William H. Marinell, Daniel R. C. Thal, and Douglas O. Staiger. 2016. Teaching higher: Educators' perspectives on Common Core implementation. Cambridge, MA: Center for Education Policy Research.
- Kane, Thomas J., and Douglas O. Staiger. 2008. Estimating teacher impacts on student achievement: An experimental evaluation. NBER Working Paper 14607.
- Koedel, Corey. 2008. Teacher quality and dropout outcomes in a large, urban school district. *Journal of Urban Economics* 64(3): 560-572.
- Koedel, Corey, Kata Mihaly, and Jonah E. Rockoff. 2015. Value-added modeling: A review. *Economics of Education Review* 47: 180-195.
- Koretz, Daniel M. 2008. *Measuring up*. Cambridge, MA: Harvard University Press.
- Kraft, Matthew A., David Blazar, and Dylan Hogan. Forthcoming. The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*.
- Kraft, Matthew A. Forthcoming. Teaching effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*.
- Kupermintz, Haggai. 2003. Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis* 25(3): 287-298.
- Lyubomirsky, Sonja, Laura King, and Ed Diener. 2005. The benefits of frequent positive affect: Does happiness lead to success? *Psychological Bulletin* 131(6): 803-855.
- Mueller, Gerrit, and Erik Plug. 2006. Estimating the effect of personality on male and female earnings. *Industrial & Labor Relations Review* 60(1): 3-22.
- Murnane, Richard J., and Barbara R. Phillips. 1981. What do effective teachers of inner-city

- children have in common? *Social Science Research* 10(1): 83-100.
- Murnane, Richard J., and John B. Willett. 2010. *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.
- Newton, Xiaoxia A., Linda Darling-Hammond, Edward Haertel, and Ewart Thomas. 2010. Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives* 18(23): 1–27.
- Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges. 2004. How large are teacher effects? *Educational Evaluation and Policy Analysis* 26(3): 237-257.
- Papay, John P., Eric S. Taylor, John H. Tyler, and Mary Laski. 2016. Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. NBER Working Paper 21986. National Bureau of Economic Research.
- Pianta, Robert C., and Bridget K. Hamre. 2009. Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher* 38(2): 109-119.
- Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical linear models: Applications and data analysis methods. Second edition*. Thousand Oaks, CA: Sage Publications.
- Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. Teachers, schools, and academic achievement. *Econometrica* 73(2): 417–458.
- Rothstein, Jesse. 2010. Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics* 125(1): 175-214.
- Spearman, Charles. 1904. “General Intelligence,” objectively determined and measured. *The American Journal of Psychology* 15(2): 201-292.
- Thompson, Paul N., Cassandra M. Guarino, and Jeffrey M. Wooldridge. 2015. An evaluation of

teacher value-added models with peer effects. Unpublished Working Paper.

Todd, Petra E., and Kenneth I. Wolpin. 2003. On the specification and estimation of the production function for cognitive achievement. *The Economic Journal* 113(485): F3-F33.

West, M. R. 2016. Should non-cognitive skills be included in school accountability systems? Preliminary evidence from California's CORE districts. The Brookings Institution.

West, Martin R., Matthew A. Kraft, Amy S. Finn, Rebecca E. Martin, Angela L. Duckworth, Christopher F. Gabrieli, and John D. Gabrieli. 2016. Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis* 38(1): 148-170.

Tables

Table 1
Demographic Characteristics of Participating Teachers

	Full NCTE Sample	Experimental Sample		Non-Experimental Sample		District Populations	
		Mean	<i>P</i> -Value on Difference	Mean	<i>P</i> -Value on Difference	Mean	<i>P</i> -Value on Difference
Male	0.16	0.15	0.95	0.19	0.604	--	--
African-American	0.22	0.18	0.529	0.24	0.790	--	--
Asian	0.03	0.05	0.408	0.00	0.241	--	--
Hispanic	0.03	0.03	0.866	0.02	0.686	--	--
White	0.65	0.70	0.525	0.67	0.807	--	--
Mathematics Coursework	2.58	2.62	0.697	2.54	0.735	--	--
Mathematical Content Knowledge	0.01	0.05	0.816	0.07	0.671	--	--
Alternative Certification	0.08	0.08	0.923	0.12	0.362	--	--
Teaching Experience	11.04	14.35	0.005	11.44	0.704	--	--
Value Added on State Math Test	0.02	0.00	0.646	0.01	0.810	0.00	0.065
<i>P</i> -value on Joint Test			0.533		0.958		NA
Teachers	310	41		51		3,454	

Note: *P*-value refers to difference from the full NCTE sample.

Table 2
Balance Between Randomly Assigned Teacher Effectiveness in Math and Student Characteristics

	Teacher Effects on State Math Scores from Randomly Assigned Teacher
Male	-0.005 (0.009)
African American	0.028 (0.027)
Asian	0.030 (0.029)
Hispanic	0.043 (0.028)
White	0.010 (0.028)
FRPL	0.002 (0.011)
SPED	-0.023 (0.021)
LEP	0.004 (0.014)
Prior Achievement on State Math Test	0.009 (0.007)
Prior Achievement on State Reading Test	-0.001 (0.007)
<i>P</i> -Value on Joint Test	0.316
Teachers	41
Students	598

Notes: The regression model includes fixed effects for randomization block. Robust standard errors in parentheses.

Table 3a
Standard Deviation of Teacher-Level Variance

	(1)	(2)	(3)
<u>Panel A: Unshrunk Estimates</u>			
State Math Test	0.28	0.19	0.13
Behavior in Class	0.28	0.24	0.14
Self-Efficacy in Math	0.29	0.27	0.19
Happiness in Class	0.35	0.35	0.26
<u>Panel B: Shrunk Estimates</u>			
State Math Test	0.22	0.18	0.13
Behavior in Class	0.13	0.09	0.05
Self-Efficacy in Math	0.00	0.00	0.08
Happiness in Class	0.33	0.33	0.34
Prior Achievement		X	X
Student Characteristics			X
Class Characteristics			X
School-by-Grade Fixed Effects	X	X	X
Teachers	41	41	41
Students	531	531	531

Table 3b
Correlations Between Unshrunk and Shrunk Teacher Effect Estimates

	(1)	(2)	(3)
Teacher Effects on State Math Test	0.93	0.95	0.84
Teacher Effects on Behavior in Class	0.87	0.84	0.87
Teacher Effects on Self-Efficacy in Math	--	--	0.83
Teacher Effects on Happiness in Class	0.95	0.95	0.87
Prior Achievement		X	X
Student Characteristics			X
Class Characteristics			X
School-by-Grade Fixed Effects	X	X	X
Teachers	41	41	41

Note: Empty cells indicate that variation in unshrunk or shrunk teacher effects is close to 0 SD (see Table 3a).

Table 3c
Correlations Between Teacher Effects on Different Student Outcomes

	Teacher Effects on State Math Test	Teacher Effects on Behavior in Class	Teacher Effects on Self-Efficacy in Math	Teacher Effects on Happiness in Class
<u>Panel A: Unshrunk Estimates</u>				
Teacher Effects on State Math Test	1.0			
Teacher Effects on Behavior in Class	0.16	1.0		
Teacher Effects on Self-Efficacy in Math	0.17	0.48**	1.0	
Teacher Effects on Happiness in Class	-0.22	0.17	0.44**	1.0
<u>Panel B: Shrunk Estimates</u>				
Teacher Effects on State Math Test	1.0			
Teacher Effects on Behavior in Class	0.17	1.0		
Teacher Effects on Self-Efficacy in Math	-0.03	0.65***	1.0	
Teacher Effects on Happiness in Class	-0.38*	0.17	0.59***	1.0

Notes: * p<.05, ** p<.01, ***p<.001. Teacher effects are calculated from Model 3 from Tables 3a and 3b, which controls for prior achievement, student characteristics, class characteristics, and randomization block fixed effects. Samples include 41 teachers.

Table 4a
Pairwise Correlations Between Teacher Effects Across Model Specifications

	$\rho_{Model\ 1,Model\ 2}$	$\rho_{Model\ 1,Model\ 3}$	$\rho_{Model\ 2,Model\ 3}$
<u>Panel A: Unshrunk Estimates</u>			
Teacher Effects on Behavior in Class	0.89***	0.89***	1.00***
Teacher Effects on Self-Efficacy in Math	0.88***	0.91***	0.98***
Teacher Effects on Happiness in Class	0.96***	0.97***	0.99***
<u>Panel B: Shrunk Estimates</u>			
Teacher Effects on Behavior in Class	0.90***	0.91***	1.00***
Teacher Effects on Self-Efficacy in Math	0.86***	0.90***	0.97***
Teacher Effects on Happiness in Class	0.96***	0.96***	0.99***

Notes: ***p<.001. Model 1 calculates teacher effectiveness ratings that control for students' prior achievement in math and reading. Model 2 controls for a prior measure of students' attitude or behavior. Model 3 controls for prior scores on both prior achievement and prior attitude or behavior. Samples include 51 teachers.

Table 4b
Pairwise Correlations Between Unshrunk Teacher Effects from Model 1 and Other Model Specifications

	$\rho_{Model\ 1,Model\ 4}$	$\rho_{Model\ 1,Model\ 5}$	$\rho_{Model\ 1,Model\ 6}$	$\rho_{Model\ 1,Model\ 7}$
<u>Panel A: Unshrunk Estimates</u>				
Teacher Effects on State Math Test	0.98***	0.72***	0.64***	0.49***
Teacher Effects on Behavior in Class	0.95***	0.74***	0.64***	0.38***
Teacher Effects on Self-Efficacy in Math	0.99***	0.84***	0.78***	0.46***
Teacher Effects on Happiness in Class	0.97***	0.85***	0.52***	0.49***
<u>Panel B: Shrunk EB Estimates</u>				
Teacher Effects on State Math Test	0.99***	0.76***	0.54***	0.53***
Teacher Effects on Behavior in Class	0.98***	0.69***	0.63***	0.41***
Teacher Effects on Self-Efficacy in Math	0.99***	--	--	--
Teacher Effects on Happiness in Class	0.99***	0.90***	0.71***	0.66***

Notes: ***p<.001. Baseline model to which others are compared (Model 1) calculates teacher effectiveness ratings that only control for students' prior achievement in math and reading. Model 4 adds student demographic characteristics, including gender, race, free or reduced-price lunch eligibility, special education status, and limited English proficiency status; Model 5 adds classroom characteristics; Model 6 adds school characteristics; Model 7 replaces school characteristics with school fixed effects. Empty cells indicate that variation in teacher effects from one of the models is close to 0 SD (see Appendix Table 4). Samples include 51 teachers.

Table 5
Relationship Between Current Student Outcomes and Prior, Non-Experimental Teacher Effect Estimates

	State Math Test		Behavior in Class		Self-Efficacy in Math		Happiness in Class	
	Estimate/SE	P-value on Difference from 1 SD	Estimate/SE	P-value on Difference from 1 SD	Estimate/SE	P-value on Difference from 1 SD	Estimate/SE	P-value on Difference from 1 SD
<u>Panel A: Unshrunk Estimates</u>								
Teacher Effects Calculated from Model 1	0.846*** (0.075)	0.046	0.685*** (0.165)	0.064	0.418~ (0.216)	0.010	0.350* (0.142)	0.000
Teacher Effects Calculated from Model 4	0.869*** (0.081)	0.115	0.706*** (0.148)	0.054	0.414~ (0.219)	0.011	0.361* (0.140)	0.000
Teacher Effects Calculated from Model 5	0.914*** (0.094)	0.366	0.719*** (0.144)	0.058	0.447~ (0.244)	0.029	0.349* (0.133)	0.000
Teacher Effects Calculated from Model 6	0.915*** (0.092)	0.361	0.742*** (0.142)	0.077	0.461~ (0.247)	0.035	0.399** (0.136)	0.000
Teacher Effects Calculated from Model 7	0.903*** (0.095)	0.314	0.768*** (0.145)	0.118	0.448~ (0.234)	0.023	0.399** (0.134)	0.000
<u>Panel B: Shrunk Estimates</u>								
Teacher Effects Calculated from Model 1	0.960*** (0.078)	0.609	1.003*** (0.266)	0.992	0.514 (0.369)	0.195	0.427* (0.177)	0.003
Teacher Effects Calculated from Model 4	0.995*** (0.084)	0.953	1.090*** (0.268)	0.738	0.507 (0.372)	0.192	0.438* (0.175)	0.003
Teacher Effects Calculated from Model 5	1.055*** (0.100)	0.585	1.240*** (0.305)	0.436	--	--	0.416* (0.167)	0.001
Teacher Effects Calculated from Model 6	1.079*** (0.101)	0.435	1.472*** (0.368)	0.207	--	--	0.487** (0.174)	0.005
Teacher Effects Calculated from Model 7	1.084*** (0.102)	0.419	1.789*** (0.458)	0.092	--	--	0.522** (0.172)	0.008
Teachers	41		41		41		40	
Students	531		531		531		509	

Notes: ~ p< .10, * p<.05, ** p<.01, ***p<.001. Cells include estimates from separate regression models that control for students' prior achievement in math and reading, student demographic characteristics, classroom characteristics from randomly assigned rosters, and fixed effects for randomization block. Robust standard errors clustered at the class level in parentheses. Model 1 calculates teacher effectiveness ratings that only control for students' prior achievement in math and reading; Model 4 adds student demographic characteristics; Model 5 adds classroom characteristics; Model 6 adds school characteristics; Model 7 replaces school characteristics with school fixed effects. Empty cells indicate that variation in non-experimental teacher effects is close to 0 SD (see Appendix Table 4).

Appendix

Table A.1
Univariate and Bivariate Descriptive Statistics for Student Survey

	Univariate Statistics			Pairwise Correlations		
	Mean	SD	Cronbach's Alpha	Behavior in Class	Self-Efficacy in Math	Happiness in Class
<u>Behavior in Class</u>	4.10	0.93	0.74	1.00		
My behavior in this class is good.	4.23	0.89				
My behavior in this class sometimes annoys the teacher.	3.80	1.35				
My behavior is a problem for the teacher in this class.	4.27	1.13				
<u>Self-Efficacy in Math</u>	4.17	0.58	0.76	0.35***	1.00	
I have pushed myself hard to completely understand math in this class.	4.23	0.97				
If I need help with math, I make sure that someone gives me the help I need.	4.12	0.97				
If a math problem is hard to solve, I often give up before I solve it.	4.26	1.15				
Doing homework problems helps me get better at doing math.	3.86	1.17				
In this class, math is too hard.	4.05	1.10				
Even when math is hard, I know I can learn it.	4.49	0.85				
I can do almost all the math in this class if I don't give up.	4.35	0.95				
I'm certain I can master the math skills taught in this class.	4.24	0.90				
When doing work for this math class, focus on learning not time work takes.	4.11	0.99				
I have been able to figure out the most difficult work in this math class.	3.95	1.09				
<u>Happiness in Class</u>	4.10	0.85	0.82	0.27***	0.62***	1.00
This math class is a happy place for me to be.	3.98	1.13				
Being in this math class makes me feel sad or angry.	4.38	1.11				
The things we have done in math this year are interesting.	4.04	0.99				
Because of this teacher, I am learning to love math.	4.02	1.19				
I enjoy math class this year.	4.12	1.13				

Notes: ***p<.001. Statistics are generated from all available data. All survey items are on a scale from 1 to 5. Statistics drawn from all available data.

Table A.2
Summary of Random Assignment Student Compliance

	Number of Students	Percent of Total
Remained with randomly assigned teacher	677	0.72
Switched teacher within school	168	0.18
Left school	40	0.04
Left district	49	0.05
Not sure	9	0.01
Total	943	1.00

Table A.3
Comparison of Student Compliers and Non-Compliers in Randomization Blocks with Low Levels of Non-Compliance

	Non-Compliers	Compliers	<i>P</i> -Value on Difference
<u>Student Characteristics</u>			
Male	0.38	0.49	0.044
African American	0.38	0.33	0.374
Asian	0.12	0.15	0.435
Hispanic	0.15	0.21	0.128
White	0.31	0.27	0.403
FRPL	0.64	0.66	0.572
SPED	0.06	0.05	0.875
LEP	0.11	0.21	0.016
Prior Achievement on State Math Test	0.30	0.26	0.689
Prior Achievement on State Reading Test	0.28	0.30	0.782
<i>P</i> -Value on Joint Test			0.146
<u>Teacher Characteristics</u>			
Prior Teacher Effects on State Math Scores	-0.01	-0.01	0.828
Students	67	531	

Note: Means and *p*-values are calculated from regression framework that controls for randomization block.

Table A.4
Standard Deviation of Teacher-Level Variance in Non-Experimental Sample

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
<u>Panel A: Unshrunk Estimates</u>							
State Math Test	0.26	NA	NA	0.25	0.19	0.15	0.13
Behavior in Class	0.48	0.38	0.38	0.41	0.31	0.29	0.16
Self-Efficacy in Math	0.35	0.30	0.30	0.33	0.37	0.30	0.14
Happiness in Class	0.48	0.45	0.44	0.43	0.43	0.34	0.27
<u>Panel B: Shrunk Estimates</u>							
State Math Test	0.17	NA	NA	0.17	0.13	0.11	0.14
Behavior in Class	0.33	0.22	0.22	0.26	0.19	0.21	0.10
Self-Efficacy in Math	0.15	0.14	0.14	0.13	0.00	0.00	0.00
Happiness in Class	0.34	0.31	0.31	0.34	0.34	0.30	0.33
Prior Achievement	X		X	X	X	X	X
Prior Survey Response		X	X				
Student Characteristics				X	X	X	X
Class Characteristics					X	X	X
School Characteristics						X	
School Fixed Effects							X
Teachers	51	51	51	51	51	51	51
Students	548	548	548	548	548	548	548

Table A.5
 Relationship Between Student Outcomes Following Random Assignment and Prior, Non-Experimental Teacher Effect Estimates that Control for an Imputed Measure of Students' Prior Survey Response

	Behavior in Class		Self-Efficacy in Math		Happiness in Class	
	Estimate/SE	<i>P</i> -value on Difference from 1 SD	Estimate/SE	<i>P</i> -value on Difference from 1 SD	Estimate/SE	<i>P</i> -value on Difference from 1 SD
<u>Panel A: Unshrunk Estimates</u>						
Teacher Effects Calculated from Model 2	0.692*** (0.159)	0.059	0.452* (0.197)	0.008	0.346* (0.141)	0.000
Teacher Effects Calculated from Model 3	0.702*** (0.165)	0.078	0.447* (0.213)	0.013	0.347* (0.141)	0.000
<u>Panel B: Shrunk Estimates</u>						
Teacher Effects Calculated from Model 2	1.067*** (0.269)	0.805	0.604 (0.360)	0.279	0.423* (0.175)	0.002
Teacher Effects Calculated from Model 3	1.073*** (0.286)	0.799	0.563 (0.367)	0.241	0.421* (0.175)	0.002
Teachers		41		41		40
Students		531		531		509

Notes: * $p < .05$, *** $p < .001$. Cells include estimates from separate regression models that control for students' prior achievement in math and reading, student demographic characteristics, classroom characteristics from randomly assigned rosters, and fixed effects for randomization block. Robust standard errors clustered at the class level in parentheses. Model 2 calculates non-experimental teacher effects controlling for a prior measure of students' attitude or behavior. Model 3 controls for prior scores on both prior achievement and prior attitude or behavior. Both models include an indicator for whether or not students' lagged survey response was imputed.