Explaining Teacher Effects on Achievement

Using Commonly Found Teacher-Level Predictors

Andrew Bacher-Hicks, Mark J. Chin, and Heather C. Hill

Harvard University

Douglas O. Staiger

Dartmouth College

Author Note

Andrew Bacher-Hicks, John F. Kennedy School of Government, Harvard University; Mark Chin, Harvard Graduate School of Education, Harvard University; Heather C. Hill, Harvard Graduate School of Education, Harvard University; Douglas O. Staiger, Department of Economics, Dartmouth College.

Correspondence concerning this article should be addressed to Andrew Bacher-Hicks, John F. Kennedy School of Government, 79 John F. Kennedy Street, Cambridge, MA 02138. E-mail: abacherhicks@g.harvard.edu.

Acknowledgments

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C090023 to the President and Fellows of Harvard College to support the National Center for Teacher Effectiveness. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

Teachers have large effects on students' academic outcomes, but there is little understanding of which features of teachers and teaching best explain these effects. In this study, we collected 22 measures of teacher and teaching quality from three broad categories: classroom instruction (e.g., quality, content), teacher personal characteristics (e.g., knowledge, self-efficacy), and teacher background (e.g., education, experience). All three categories explained some unique variation in teacher effects, though measures of instruction explained the least, even after correcting for measurement error. Thus, although a diverse set of measures best distinguishes teachers' effectiveness, a cost-effective starting point might be to focus on the less expensive but relatively information-rich measures of teacher personal characteristics and background.

Keywords: Education production; teacher effectiveness; value-added models

1. Introduction

Though teachers have large effects on their students' academic and long-term outcomes (e.g., Chetty et al. 2014a, 2014b; Kane et al. 2013), we lack a comprehensive understanding of what explains this variation in teacher effects. Hundreds of studies have examined this topic, with some using easily quantifiable measures—such as teaching experience and credentials—to explain student outcomes; however, these measures largely fail to account for differences among teachers (e.g., Hanushek 1986; Wayne and Youngs 2003). Other studies develop, collect, and test more job-embedded measures (e.g., pedagogical content knowledge, instructional quality), but similarly explain little variation in teacher effects (e.g., Boonen et al. 2014; Hill et al. 2005; Kunter et al. 2013; Palardy and Rumberger 2008).

There are at least three reasons for the lack of explanatory power. First, most prior studies examine only a few predictors at a time; we thus do not know whether a more comprehensive set of measures explains substantial variance in teacher effects. For example, if each individual measure accounts for a small but unique portion of the variation in teacher effects, a diverse set of measures may collectively account for substantial variation across teachers. Relatedly, without directly comparing different types of measures (e.g., instructional quality versus teachers' knowledge) in the same study, it is difficult to evaluate their relative explanatory power and the extent of overlap between them. Second, some of these commonly-used measures—such as those derived from classroom observation—contain substantial measurement error, yet little is known about the extent to which measurement error attenuates the explained variance in teacher effects. Finally, we do not know whether these issues hold across student outcomes on high- and low-stakes assessments. Because high-stakes assessments may lead to distorted teacher effect estimates (e.g., due to test-focused preparation, Koretz 2008), and because measures of teacher

and teaching quality may not align with high-stakes assessments, analyzing results from low-stakes assessments may confirm the value of the predictors of teacher quality assembled by the field.

We address these issues in this paper by collecting administrative records, teacher surveys, and videos of classroom observation from 283 teachers and 7,843 students in fourth- and fifth-grade mathematics classrooms across four large East Coast public school districts. Using a value-added approach, we estimated teacher effects on two types of student assessments: state standardized assessments and a predictor-aligned low-stakes assessment. We then generated 22 commonly used measures of teacher and teaching quality from three broad categories: instruction, teacher personal characteristics, and teacher background. Using these data, we ask the following three research questions:

1. How much variance in teacher effects does a diverse set of predictors collectively explain, and how well do specific types of predictors explain such variance? In addition, how much overlap exists between measures?
2. To what extent does measurement error in the predictor variables attenuate the variance explained?
3. Do predictors account for similar amounts of variance across the state tests and the predictor-aligned low-stakes test?

Our results yield three key findings. First, we show that all three categories of measures explained some unique variation in teacher effects, yet no single measure substantially differentiated teachers. This suggests that, when possible, researchers and policymakers should examine a diverse set of measures to best distinguish between teachers. However, among the measures studied here, observation-based measures of instruction explained the least amount of

variation in teacher effects. Thus, given the high cost of collecting observation-based measures, a cost-effective approach to distinguishing teachers may be to rely on the more predictive and less expensive survey-based measures (e.g., teacher knowledge, background). Second, we conclude that the attenuation in explained variance due to measurement error is minor. Thus, developing new measures is likely to be more fruitful in explaining additional variation than improving the reliability of existing ones. Finally, we explained substantially more variance on a low-stakes predictor-aligned test than on high-stakes state tests. This finding highlights the importance of alignment between predictors and outcome variables, and suggests that theoretically important measures may be not be as well suited in explaining teacher quality based on high-stakes state tests.

2. Background and Motivation

Scholars have invested considerable effort in identifying and measuring features of teachers and teaching that relate to student performance on standardized tests. Such features fall into three general categories: instruction (e.g., quality, content), teacher personal characteristics (e.g., knowledge, self-efficacy) and teacher background (e.g., coursework, experience). Although measures from multiple categories are not typically examined in the same study, there is ample evidence that measures from all three categories individually predict student learning.

2.1 Instruction

Over the past four decades, researchers have examined the relationship between the activities that occur inside classrooms and student learning, with the latter usually measured by performance on standardized tests. One line of research has focused on classroom organization

and general pedagogical methods (e.g., classroom climate, managing classroom behavior, assessing student learning), which consistently relate to student outcomes (e.g., Pianta et al. 2008; Reyes et al. 2012; Stronge et al. 2011). Other scholars, by contrast, have focused on more nuanced instructional strategies that respond to the nature of the discipline and students' learning of that discipline. These activities are often less strongly related to student outcomes, yet still significantly predict student outcomes in most cases (Bell et al. 2012; Garet et al. 2016; Good and Grouws 1977; Grossman et al. 2013; Tyler et al. 2010). Finally, a third set of research focuses on the alignment between lesson content and tested topics. In some classrooms, for example, teachers routinely expose students to released test items, creating opportunities to learn that closely mimic test content (Hamilton 2003; Koretz 2008). Several studies indeed show that students in classrooms that devote more time to tested topics have higher test scores (Cooley and Leinhardt 1980; Gamoran et al. 1997).

2.2 Personal Characteristics

Because measures of instructional practice are typically costly to devise and implement, other research has focused on identifying and measuring personal characteristics of teachers theorized to more indirectly affect student learning. Perhaps the most widely studied of these characteristics is teachers' knowledge. Theory suggests that disciplinarily rich, error-free instruction requires that teachers possess both pure content and teaching-specific content knowledge (Ball et al. 2008; Shulman 1986), and empirical investigations have generally found a positive relationship between such knowledge and student outcomes (Carlisle et al. 2011; Hill et al. 2005; Rockoff et al. 2011). A second widely studied type of knowledge focuses on teachers' knowledge of their students (e.g., students' common misunderstandings), which can also

facilitate effective instruction. Teachers' knowledge of their students similarly positively relates to student outcomes (Carpenter et al. 1988; Hoge and Coladarci 1989; Sadler et al. 2013).

Other research on teachers' personal characteristics besides knowledge has focused on teachers' effort and self-efficacy. Though effort is rarely measured directly, self-efficacy (i.e., the extent to which teachers believe they have the capacity to affect student performance) theoretically relates to effort (Bandura, 1977), and research generally finds a positive relationship between teachers' self-efficacy and student outcomes (Armor et al. 1976; Kunter et al. 2013; Ross 1992). Other studies have focused on teacher performance pay, which provides incentives for increased effort. Although the effects of performance pay have been mixed, studies that find a positive impact show that gains in student achievement appear to be partly due to increased teacher effort (Lavy 2009; Muralidharan and Sundararaman 2011).

2.3 Teacher Background

Rather than developing measures of instructional quality or assessing teacher personal characteristics, another line of research has relied on pre-existing teacher background variables from administrative datasets (e.g., teacher pre-service coursework, degree, certification status, and experience). In theory, these variables should predict student achievement because they proxy for skills and knowledge that teachers need to be effective in the classroom. The empirical results, however, have been largely inconsistent; only teacher experience consistently relates to student outcomes, with the most pronounced gains occurring in the early years of one's teaching career (Boonen et al. 2014; Clotfelter et al. 2007; Harris and Sass 2011; Kane et al. 2008; Papay and Kraft 2015). The results for other measures are mixed, including for educational attainment (e.g., Aaronson et al. 2007; Goldhaber and Brewer 2000; Harris and Sass 2011; Rowan et al.

1997, 2002), certification (for a review, see Cochran-Smith et al. 2012) and post-secondary mathematics coursework (e.g., Harris and Sass 2011; Hill et al. 2005; Monk 1994; Wayne and Youngs 2003).

2.4 Explaining Variation in Teacher Effects

A considerable body of work has explored relationships between individual measures and teacher effects, yet few studies use multiple measures from each category to explain teacher-level variation in student outcomes. Further, these studies tend to explain little variation in teacher effects. Studies using only easily quantifiable background variables (e.g., teaching experience, educational attainment) explained less than 5% of the variation across teachers (Goldhaber et al. 1999; Nye et al. 2004). Other studies using surveys to collect more detailed sets of predictors (e.g., teachers' content knowledge, self-efficacy) explain about one fifth of the variance in teacher effects at best (Carlisle et al. 2009; Hill et al. 2005; Kunter et al. 2013; Palardy and Rumberger 2008). However, even these studies have included only a small handful of predictors from a single data source, and none have included observations of teachers' instructional practice. These gaps in the literature motivate our study.

3. Data and Methods

3.1 Design and Data Sources

This study included 283 teachers and 7,843 students from two cohorts of fourth- and fifth-grade mathematics classrooms across four large East Coast public school districts in three states. Over the course of two academic years, 2010–11 and 2011–12, we observed classrooms using up to three videos of instruction per year from each teacher. Teachers selected the dates for

recording their classes under the restriction that they choose lessons typical of their teaching, that they choose lessons longer than 20 minutes, and that they not choose lessons during which student testing would occur.[1] Recorded lessons lasted approximately one hour and typically consisted of the presentation of new tasks and material as opposed to computational practice or lengthy test preparation. In addition to collecting videos of classroom instruction, we administered surveys to all participating teachers in the fall and spring of both years to collect measures of teacher characteristics and teacher background. Finally, we conducted a follow-up survey in 2012–13 with additional questions on teacher knowledge.

We collected student-level data from two sources: de-identified administrative data provided by the participating school districts and a project-administered low-stakes mathematics assessment. We received the following administrative data for all fourth- and fifth-grade students in participating districts from the school years 2010–11 and 2011–12: (a) teacher of record, (b) demographic information, and (c) performance on state standardized mathematics and reading exams. In addition to these administrative data, we administered additional fourth- and fifth-grade assessments to students in participating classrooms in 2010–11 and 2011–12. The project-administered test was jointly developed by [omitted for blind review] and focused on three mathematical domains—number and operations, algebra, and geometry and measurement—in order to align with the fourth- and fifth-grade Common Core mathematics standards. The state tests used in our analysis have a range of reliability estimates from 0.90 to 0.93 and the reliability estimates for the project-administered assessment range from 0.82 to 0.89.

3.2 Student-Level Measures

In this study, student-level data were used as control and outcome variables. We used the

following variables from state administrative datasets as controls: students' prior achievement on state standardized mathematics and reading tests, as well as indicators of race, gender, subsidized-price lunch eligibility, English language learner status, and special education status. For our two outcome variables, we used students' current-year performance on state standardized mathematics tests and on the project-administered mathematics test. Because the state tests varied across the three states in the study, we standardized students' scores on these exams within district, grade, and academic year using van der Waerden rank-based standardization methods to generate *z*-scores (Conover 1999). For interpretability, we also transformed students' scores from project-administered assessment into *z*-scores.

3.3 Teacher-Level Measures

We derived 22 teacher-level measures from teacher surveys and classroom observations, each believed to be predictors of teacher effects on student achievement. We provide an overview of these measures in Table 1 and briefly discuss each measure below (see Appendix A for more detail).

[Insert Table 1 here.]

Measures of instruction. We derived observation-based measures of instruction by evaluating video recordings using two established instruments: CLASS (Pianta et al. 2007) and Mathematical Quality of Instruction (MQI; Hill et al. 2008). CLASS is a subject-matter-independent observation tool designed to capture content-general domains of student-teacher interactions, such as classroom organization. MQI, on the other hand, captures mathematics-specific features of instruction, such as teachers' mathematical errors and imprecisions. We followed each instrument's specific protocol to generate scores. For CLASS, one rater scored

each 15-minute segment within each recorded lesson on a scale from 1 to 7 for each of the 12 CLASS items. For MQI, two raters scored each 7.5-minute segment on a scale of 1 to 3 for each of the 13 MQI items.

Based on prior factor analyses from the same study (Authors 2016), we consolidated the 13 MQI items into two mathematics-specific factors (*Ambitious Instruction*; *Mathematical Errors*) and the 12 CLASS items into two content-general factors (*Classroom Organization*; *Support*). We generated teacher-level composite scores for each factor by first averaging across the relevant items and then adjusting for reliability (see Appendix B for more detail).

We supplemented the four video-based instructional practices with four survey-based measures of instructional content and its alignment to tested material. The first two measures focused on the extent to which classroom content and problem formats matched those on the project-administered assessments. Specifically, we measured the extent to which teachers reported covering a list of nine tested algebra topics (*Algebra Content*) and 16 number and operations topics (*Number and Operations Content*). The other two survey-based measures focused on instructional activities related to preparation for state tests. Specifically, we measured the extent to which teachers engaged in instructional behaviors designed to improve student performance on state standardized tests, such as using released test items or practice test materials (*Test Prep Activities*). We also measured the extent to which teachers changed their instructional practices in response to state-imposed testing and accountability systems, such as changing the sequencing of topics so that content more likely to appear on the state test is covered before the test is administered (*Test Prep Instructional Changes*). As with the observation-based measures of instruction, we estimated an average score for each teacher across all relevant items and then adjusted for reliability.

Measures of teacher personal characteristics. We collected four measures of teacher personal characteristics using surveys. To measure teachers' mathematical content knowledge (*Mathematical Knowledge*), we used a one-parameter graded response model to score teachers' performance on 72 items from the Mathematical Knowledge for Teaching measure (Hill et al. 2005) and 33 total released items from the mathematics component of a state test for educator licensure. The Mathematical Knowledge for Teaching assesses teachers' facility in using mathematical knowledge in the context of classroom teaching (e.g., ability to select appropriate representations and examples of mathematical concepts such as fractions) and the state test for educator licensure measures subject matter knowledge in the upper elementary and middle grades.

To generate a second measure of teacher knowledge (*Knowledge of Student Performance),* representing teachers' knowledge of their students' mathematics performance, we presented teachers with a subset of items from the project-administered mathematics assessment and then asked what percent of their students would answer the item correctly. We then calculated the absolute difference between teachers' estimated percentage of students correctly answering the item and the actual percentage of students correctly answering the item. We averaged this difference across items for each teacher, and then adjusted this composite score for reliability.

In addition to the two measures of teacher knowledge, we derived measures of teachers' self-efficacy beliefs and their teaching effort. *Self-efficacy* was based on 14 survey questions regarding teachers' beliefs about how much they can control classroom behavior, motivate students, and craft good instruction. To measure *teacher effort, we* asked teachers to indicate the

number of hours per week spent on four non-instructional activities: preparing for class, organizing materials, grading homework, and reviewing the content of lessons.

Measures of background. We generated 10 variables describing teachers' educational background and preparation, based on teacher surveys. Of these measures, three were indicator variables (0 = *no*, 1 = *yes*) that captured teachers' course-taking and educational attainment. *Master's degree* indicated any earned master's degree; *math major* indicated an undergraduate or graduate degree in mathematics; and *education bachelor's* indicated a bachelor's degree in education. We also asked teachers to provide both the number of undergraduate- or graduate-level *math courses* and *math content courses* they had taken, using a 4-point scale from 1 = *No Classes* to 4 = *Six+ Classes.*

In addition, we created five indicators of teaching preparation and experience. *Traditional certification* indicated that the teacher currently holds a traditional teaching certification; *alternative certification* indicated that the teacher currently holds an alternative teaching certification; *elementary math certification* indicated that the teacher possesses a specific certification for teaching elementary mathematics; *4–10 years experience* indicated that the teacher has between 4 and 10 years of teaching experience (including the year surveyed); and *10+ years experience* indicated that the teacher has more than 10 years of teaching experience (including the year surveyed).

3.4 Analysis Strategy

To address our research questions, we estimated how much teacher-level variance in student outcomes was explained by different features of teachers and teaching, both in isolation and in conjunction with one another. We used two student outcomes: scores on state

standardized mathematics tests and scores on the project-administered mathematics test. To estimate the amount of teacher-level variation in these student outcomes, we fit the following multilevel model to estimate teacher effects:

$$a_{i,j,k,t} = A_{i,t-1}\alpha + S_{i,t}\beta + P_{j,k,t}\delta + \eta + \nu_{i,j,k,t}, \quad (1)$$

$$\text{where } \nu_{i,j,k,t} = \omega_k + \mu_j + \theta_{j,k,t} + \varepsilon_{i,j,k,t}$$

The outcome variable, $a_{i,j,k,t}$, represents the test score for student *i* taught by teacher *j* in school *k* during school year *t*. Based on previous studies of value-added models (Chetty et al. 2014a; Kane and Staiger 2008), we included the following control variables: $A_{i,t-1}$, a cubic polynomial of student *i*'s prior-year achievement; $S_{i,t}$, a vector of indicators for gender, race and ethnicity, subsidized-priced lunch eligibility, English language learner status, and special education status; and $P_{j,k,t}$, a vector of average characteristics of student *i*'s peers in the same class and school, including average prior-year test scores and averages of $S_{i,t}$, and grade-by-year and district fixed effects, η.

In Equation 1, we specified four levels of nested random effects: school random effects, ω_k, teacher random effects, μ_j, classroom (or teacher-year) random effects, $\theta_{j,k,t}$, and student-level error, $\varepsilon_{i,j,k,t}$. In this study, we focused on the teacher-level random effects, μ_j, which represent teachers' contributions to student outcomes. To determine how much of the variation in μ_j was explained by different observable features of teachers or teaching, we estimated a taxonomy of multilevel models similar to Equation 1:

$$a_{i,j,k,t} = A_{i,t-1}\alpha + S_{i,t}\beta + P_{j,k,t}\delta + \eta + \nu_{i,j,k,t}, \quad (2)$$

$$\text{where } \nu_{i,j,k,t} = \omega_k + \varphi_j + \theta_{j,k,t} + \varepsilon_{i,j,k,t}$$

$$\text{and } \varphi_j = T_j\gamma + \tau_j.$$

Equation 2 uses the same specification as in Equation 1, except that in Equation 2, we included teacher-level predictors, T_j, to explain variation in teacher random effects, φ_j. Specifically, φ_j is a function of a vector of m teacher-level variables, T_j, representing features of teacher and teaching, and τ_j, the teacher random effects after controlling for these features. The variance in τ_j represents the teacher-level variance in student outcomes due to differences between teachers after taking these features into account. For example, in specifications where T_j includes the 10 teacher background measures, τ_j represents the teacher-level variance in student outcomes that is not explained by these 10 measures.

By comparing an adjusted ratio of teacher-level variance components for parameters μ_j and τ_j from Equation 1 and Equation 2, we generated a statistic, analogous to an adjusted R^2, to measure the percentage of teacher-level variation that was explained by the m teacher-level variables, T_j. We define this "adjusted teacher-level R^2" statistic as follows:

$$R^2 = 1 - \frac{n-1}{n-1-m} \times \frac{\mathrm{Var}(\tau_j)}{\mathrm{Var}(\mu_j)}, \quad (3)$$

where n represents the number of teachers in the sample, m represents the number of teacher-level variables used in Equation 2, $\mathrm{Var}(\tau_j)$ represents the teacher-level variation in student outcomes in Equation 2 after controlling for the vector of teacher-level variables T_j, and $\mathrm{Var}(\mu_j)$ represents the teacher-level variation in student outcomes in Equation 1 without this vector. The ratio $(\frac{n-1}{n-1-m})$ adjusts for the mechanical reduction in $\mathrm{Var}(\tau_j)$ that tends to occur when more teacher-level variables are added to the model. We used this statistic to estimate how much teacher-level variation in student outcomes was due to different features of teachers or teaching.

3.5 Descriptive Statistics and Correlations of Teacher-Level Measures

Of the 283 fourth- and fifth-grade teachers in this study, 84% were female and 67% were White. They averaged approximately 10.2 years of experience, typically arrived in teaching via a traditional certification route (85%), and held a master's degree (75%). Of the 7,843 students in participating classrooms, nearly half were Black (44%), while most of the remaining half were either Hispanic (25%) or White (20%). More than two-thirds were eligible for subsidized lunch (69%), nearly one-quarter had limited English language proficiency (23%), and 12% had special education status. These characteristics are largely reflective of the students and teachers across these four districts and sample selection does not appear to limit our ability to generalize our findings to the population within the study districts.

[Insert Table 2 here.]

In Table 2, we present descriptive statistics and correlations for our set of 22 teacher-level measures. In the first two columns, we present the mean and standard deviation of all 22 variables, followed by the correlation coefficients between our measures.[2] Nearly all of the correlations were modest, falling between -0.30 and 0.30. Among the higher correlations, most were observed across different dimensions captured by the same instrument (e.g., the two CLASS measures: classroom organization and support). Across-instrument correlations above 0.30 occurred between the mathematical knowledge and the two mathematics-specific measures of instructional practices (ambitious instruction and errors), a relationship observed in prior research (Hill et al. 2008). Overall, these modest correlations suggest that the chosen variables are related, but also measure somewhat distinct aspects of teacher and teaching quality.

4. Results

In Table 3, we present our main results for the amount of explained variance in teacher effects on students' state test scores. As a reference point, in the first column, M0, we present the results of the base model before adding any teacher-level predictors. In this base model, the variance of the teacher effect estimates for the state test outcomes was 0.032, which is comparable to other estimates from the teacher effects literature (e.g., Chetty et al. 2014a). In the next column, M1, we included all 22 teacher-level predictor variables, which reduced the variance of the teacher effects estimates to 0.022. We adjusted for the addition of these 22 predictors—using the formula described in Equation 3—to estimate the adjusted teacher-level R^2 of 25%. In other words, a full model of all 22 predictors collectively explains 25% of the teacher-level variance in student outcomes on state tests.

[Insert Table 3 here.]

In the remaining columns of Table 3, we examine the performance of select groups of predictors in explaining teacher-level variance. In model M2, we included only the eight measures of instructional practices and content. These eight measures explained 9% of the teacher-level variation on state tests. In model M3, we included only the four measures of teacher personal characteristics, which explained 12% of the teacher-level variance. Finally, in model M4, we included only the remaining 10 measures of teacher background, which explained 10% of the teacher-level variance. A model including measures of personal characteristics and instruction (M5) explained 18% of the variation, while a model including background and instruction (M6) explained 16% of the variation.

[Insert Table 4 here.]

In addition to estimating the variance explained collectively, we present the explanatory power of each individual teacher-level variable in Table 4. To do so, we fit 22 separate

regressions, each with only one teacher-level predictor included. These estimates provide a comparison to past studies of teacher and teaching, which tend to examine the effects of these variables separately. We found that any individual variable explained, at most, 6% of the teacher-level variance. The sum of the 22 individual explained variances was 40%, substantially larger than the 25% of the variance explained collectively in Table 3, which highlights the problem of simply adding the percentage of variance explained by individual predictors that are related. In other words, because the variance explained by each measure is not purely unique, adding the variance explained in separate models will overstate their collective explanatory power.

In Table 4 we also examine the extent to which measurement error attenuates the variance explained by each predictor variable. We did not correct for measurement error in our main analyses in Table 3, since measurement error is a property of the measures that are used in practice. As such, the results in Table 3 accurately reflect the predictive power of teacher measures as they are typically collected (i.e., with some amount of measurement error). To evaluate the extent to which these results are attenuated by measurement error, we compare them to estimates of the variance that would be explained by the same predictors if they contained no measurement error. To do so, we divided the adjusted teacher-level R^2 for each individual predictor by its reliability, which produces an estimate of the teacher-level R^2 for each predictor if it contained no measurement error.[3] We present these results in the third column of Table 4. We find the error-free adjusted R^2 for any individual predictor was, at most, 8% and that the increases in explained variance for each predictor were modest.

Using the above analysis of measurement error in individual predictors, we can also provide an upper bound on the effect of error in models that include all predictors

simultaneously. If the underlying error-free components and measurement errors were uncorrelated across predictors, the R^2 in a combined model would be the same as the sum of the R^2 across the individual models; otherwise, the combined model will explain less variance than the sum of the individual models. Recognizing this, we used the sum of the increases in individual predictors' R^2 after accounting for measurement error to serve as an upper bound estimate of the increase we would achieve in the combined model.[4] In Table 4, the sum of the individual predictors' R^2 increased by approximately 9 percentage points when accounting for measurement error (from 39.5% to 48.3%). Applying this 9 percentage point gain to the original estimates from the combined model, we find that the R^2 for the state test would increase, at most, from 25% to 34% if all of the predictors were measured without error.

The results in Tables 3 and 4 are based on teacher effects using state standardized assessments. In Tables 5 and 6, we present the analogous set of results using teacher effects generating from the project-administered, low-stakes test. In Table 5, we find that the 22 predictors collectively explained 38% of the teacher-level variance in student outcomes on the project-administered assessment, which is substantially more than was explained on the state test. Each of the subgroups also explained more variance than on the state test: the eight measures of instruction explained 11%; the four measures of teacher characteristics explained 24%; and the 10 teacher background variables explained 14% of the teacher level-variation on the project-administered assessment. A model including measures of personal characteristics and instruction explained 29% of the variation, while a model including background and instruction explained 23% of the teacher-level variation.

[Insert Table 5 here.]

[Insert Table 6 here.]

In Table 6, we present parameter estimates for individual teacher-level variables and examine the extent to which measurement error attenuates the variance explained by each predictor variable on the project-administered assessment. We found that each individual predictor variable explained, at most, 14% of the teacher-level variability and that the sum of the individual explained variances was 53%. In the last column of Table 6, we present the variance explained by using the theoretical error-free constructs for each of our measures. Similar to the results for the state test, the sum of the individual predictors' R^2 increased by approximately 9 percentage points when accounting for measurement error. Applying this 9 percentage point gain to the original estimates from the combined model, we find that the R^2 for the project-administered test would increase, at most, from 38% to 47% if all of the predictors were measured without error.

5. Discussion

Research Question 1: How much variance in teacher effects does a diverse set of predictors collectively explain, and how well do specific types of predictors explain that variance? In addition, how much overlap exists between measures?

Collectively, all 22 teacher-level predictors explained a modest amount of the variation in teacher effects on both assessments: 25% on the state assessment and 38% on the project-administered assessment. Although no single group of predictors was the key to explaining variation in teacher effects, the set of teacher personal characteristics (i.e., mathematical knowledge, knowledge of student performance, self-efficacy, and effort) explained the most variation in teacher effects on both assessments. On the one hand, it was surprising that these predictors explained more variation than the measures of instruction, since these measures

theoretically influence student outcomes indirectly (i.e., through teachers' instruction). On the other hand, the measures of instruction had lower reliability than other measures, which reduced their relative explanatory power. A second explanation is that the teacher characteristics may capture qualities that correlate with effective instruction, but that are not measured by the classroom observation instruments used our study. Thus, while theory suggests that measures of instruction should best account for differences in teacher and teaching quality, in practice we find that contemporary measures of instruction are either not sufficiently reliable or not sufficiently comprehensive to account for more variation than other, less proximal measures.

While teacher personal characteristics explained the most variation in teacher effects, teacher background and measures of instruction continued to explain additional variance, even after controlling for these measures of teacher personal characteristics. For example, when adding measures of instruction to these teacher characteristics, the explained variance increased from 12% to 18% for the state test and from 24% to 29% for the project-administered assessment. When additionally including measures of teachers' background (i.e., including all 22 measures), the explained variance increased yet again in both models. This suggests that each group of measures explained some unique portion of the variation in teacher effects for both tests. Of course, each group of measures did not explain *only* a unique portion of the variation. If that were the case, the total variation explained by the combined model would have been the sum of the explained variation for each group. For example, for the state test, the sum of the background, characteristics, and instruction explained variance from M2, M3, and M4 was 32% (i.e., 9.0 + 12.4 + 10.3), which was greater than the variation explained collectively in M1 (25%). Thus, although each group of measures did not explain a purely distinct portion of the variance in teacher effects, each group held some unique explanatory power.

In practice, these results suggest that school leaders and policymakers cannot expect to find a single measure that will meaningfully differentiate prospective teachers when hiring, or that can serve as the primary focus of professional development for the current workforce. Instead, differences in teacher effects are best explained using a diverse range of measures. While the full model provides the best explanatory power, it may be impractical or too costly for states and districts to collect such a broad range of measures. A more cost-effective starting point might be for school and district leaders to consider measures of personal characteristics (e.g., content knowledge), which have the benefit of being relatively inexpensive to collect while explaining a relatively large amount of teacher-level variance. For example, assessments of teachers' content knowledge could be more heavily incorporated into the hiring process or used to identify teachers who would benefit from professional development that addresses gaps in content knowledge.

Research Question 2: To what extent does measurement error in the predictor variables attenuate the variance explained?

When we adjusted our estimates of explained variance to account for measurement error in the predictor variables, we found that, although higher than the original estimates, each error-free predictor still only independently explained a relatively small amount of teacher-level variance, especially for the state outcome variable. Thus, although measurement error attenuated our original results, the potential to explain additional variance by improving the reliability of these measures is limited. Notably, the observation-based measures of classroom instruction—which contain the most measurement error of our predictors—still explained relatively little variance in teacher effects after adjusting for measurement error. Although the low reliability of

these estimates accords with other studies (e.g., Cohen and Goldhaber 2016; Ho and Kane 2013), our results suggest that even if districts and states invested in collecting more reliable versions of these measures (e.g., by increasing the number of observations per teacher), they would still fail to explain more than a modest percentage of the variation in teacher effects. Thus, districts with policies that place an emphasis on using observation measures as sole arbiters of key human capital decisions may be off target; instead, our analyses highlight the importance of including a range of predictor variables.

Research Question 3: Do predictors account for similar amounts of variance across the state tests and the predictor-aligned low-stakes test?

Comparing our results across the state tests and the project-administered assessment, we found that collectively the 22 predictor variables explained 13 percentage points more of the teacher-level variance on the project-administered assessment than on the state assessments. This highlights the importance of alignment between predictors and outcome variables, and suggests that the high-stakes nature of state tests may lead to activities not well captured by contemporary instruments. However, for both student assessments we found that measures of teacher background and characteristics explained more variance than measures of instruction. This consistency across student outcomes further supports the recommendation that school leaders and policymakers might focus on these measures to most cost-efficiently differentiate teacher effectiveness. One likely reason for the increase in variation explained is that the project-administered student assessment was developed to match the teacher knowledge measures; it is thus not surprising that the knowledge measures explained far more teacher-level variance in such test outcomes as opposed to state test outcomes. A second reason for this difference in

teacher-level variance explained across student assessments may be that the project-administered assessment had no stakes attached; by contrast, the high-stakes nature of state tests might lead to student, teacher, or school activities not captured by our instruments (e.g., classifying low-performing students into untested categories) that may subsequently contribute to score distortion (Jacob 2005; Koretz et al. 2001). This second explanation likely also plays some role given that, relative to the state test, more variance was explained in all three categories of measures, even ones that were not specifically aligned with either test (e.g., teacher background). These results highlight that the explanatory capacity of teacher measures is outcome-specific. Thus, as states begin to collect multiple student outcomes for evaluative use under the Every Student Succeeds Act (ESSA), they may consider increasing their focus on teacher measures that explain variability across outcomes most consistently.

6. Conclusions and Future Directions

The fact that there remains so much unexplained variance in teacher effects, even in a study that went to great lengths to capture many of the theoretical and empirical predictors of teacher effects, presents challenges for practice and policy. If, after decades of studying teachers and teaching, academics have not devised a clear list of background variables, teacher characteristics, or instructional practices that strongly and substantively differentiate teacher impacts on student outcomes, it may be hard for school leaders to hire the most effective teachers or for existing teachers to efficiently develop important professional skills related to student learning. It also creates a problem for those designing policy around effective teachers. For example, ESSA requires that states create plans ensuring that all students have equitable access to excellent teachers; yet, with little consensus on which features predict excellent teachers and

teaching, ESSA has allowed states to include a variety of measures in their plans, some of which have little relationship to teacher effectiveness.

However, there are some bright spots. In particular, the categories of measures that explained the most teacher-level variance in student outcomes (i.e., teacher personal characteristics and teacher background) were those that are typically the easiest and least costly to collect in practice. For informing hiring decisions, these measures could be used to provide low-cost information, with the added benefit that they explain more variance than measures of classroom instruction. For example, teachers' mathematical knowledge is typically assessed during state certification examinations; the detailed information from these assessments could be incorporated in the hiring process. Several measures of teachers' background are already collected and presumably used during hiring (e.g., certification status, years of experience, degree information), yet other measures from this study—such as the detailed information on mathematics coursework—are less commonly used. The results of this study suggest that increased focus on a set of key measures during teacher selection has the potential to improve the quality of hired teachers.

In addition to improving the quality of new hires, these measures could be used to improve the quality of the existing workforce by identifying teachers in need of professional development or for informing early-career retention decisions. As many schools across the country face shortages in the supply of new teachers, especially in special education and STEM subjects (Cowan et al. 2016; Sutcher et al. 2016), focusing on the improvement of the current workforce may be a particularly appealing strategy. Although some measures used in this study are static or otherwise poorly suited for professional development (e.g., teacher background), many of the features do evolve over the course of teachers' careers (e.g., knowledge,

instruction). Identifying teachers with room for improvement on these more malleable measures and then providing targeted professional development could thus be an effective strategy for improving overall teacher effectiveness, especially in districts and subjects with a limited supply of new teachers. For example, teachers' mathematical knowledge could be assessed at key points throughout a teachers' career and low-performing teachers could be targeted for professional development to address gaps in their knowledge.

Although the current study suggests that improvements to these malleable factors have the potential to increase the effectiveness of the existing teacher workforce, we note that our study is correlational. Because of this, our results highlight the relationship between these measures and teacher effectiveness, but do not explicitly identify causal pathways. A second limitation of this study is that we base our estimates of teacher effects on student test scores. While it is important to understand the predictors of teacher effects on student achievement, teachers have measurable impacts on more than just students' test scores (Blazar and Kraft 2017; Jackson 2016). It could be that the predictor variables which explain the least variance in these test-based measures of teacher quality explain variance in orthogonal dimensions of teacher quality, such as a teachers' ability to promote students' social skills, resilience, or grit. Adoption of new accountability systems that focus on a variety of student outcomes (e.g., graduation rates, chronic absenteeism) under ESSA underscore the importance of understanding which measures explain variation in teacher effects on these outcomes. A third limitation of the current study is that many of our measures are specific to mathematics. These content-specific measures likely will not have the same predictive power for teachers of other subjects and it is unclear if measures designed specifically for other subjects will have similar properties to the mathematics-specific measures included in this study.

Finally, we comment on directions for future research. Our results show that even error-free versions of the 22 measures in this study still only explain at most half of the variation in teacher effects. This suggests that there are as-yet undiscovered dimensions of teacher quality, which leaves open the possibility of identifying additional, measurable factors that contribute to student learning success. We and others have watched hundreds of hours of video over many years and tried to design instruments to capture salient aspects of instruction. Although these instruments capture teacher practices that, in theory, should directly influence student learning, what we have produced thus far might be thought of as the low-hanging fruits from this endeavor—clearly visible, easy-to-record aspects of instruction such as classroom climate, behavior management, teacher content errors, and ambitious instruction. What is clear from watching the video—and arguing about it with colleagues—is the existence of many other salient features of instruction. However, these features are difficult to gauge from observations and even more difficult from teacher self-reports. The pacing of instruction, for instance, must be neither too fast nor too slow for learners; lacking knowledge of learners, however, it is impossible to assess this from video, and teachers are not likely to self-report this accurately. Teachers' strategic involvement of students—for instance by calling on specific children to engage them at critical moments in the learning process (Lampert 2001)—cannot be captured via video. Because of this, as well as the low amount of variability explained by the observational metrics and the relatively high costs associated with their collection, the search for predictors of teacher effects may proceed more fruitfully along other pathways, even if they rely on measures that are less proximal to student learning than instruction.

Notes

1. Because no rewards or sanctions were contingent on their performance in the videos, teachers had no incentive to strategically select lessons. Ho and Kane (2013) find that teachers are ranked similarly based on observation scores from their self-selected lessons compared to all other lessons.

2. All teachers in the sample had at least one score from any collected teacher-level measure. In cases where a teacher was missing some but not all such measure scores, we imputed the missing scores for the measure using chained multiple imputation (Rubin 1996). Dummy indicators were included in subsequent analyses denoting teachers with imputed scores from sources (i.e., background survey, fall teacher survey, spring teacher survey). In practice, we imputed very few scores, as 95% of teachers had scores for all 22 measures.

3. In the case of classical measurement error in a single predictor variable, the estimated R^2 is attenuated by the predictor's reliability. To illustrate, imagine an unobserved true predictor x^* is measured with classical error, ν, such that the observed value, x, is equal to $x^* + \nu$. Let the variance of the underlying predictor be σ^2_{x*} and the variance of the error be σ^2_{ν}. Then the variance of the observed predictor is $\sigma^2_{x*} + \sigma^2_{\nu}$ and the reliability of the observed predictor is $\lambda = \sigma^2_{x*}/(\sigma^2_{x*} + \sigma^2_{\nu})$. Fitting the regression model $y = \beta x + \varepsilon$ using OLS yields an estimate of β and an estimate of the R^2 that are asymptotically biased by λ. Specifically, in expectation, $\hat{\beta} = \lambda\beta$ and $R^2_{observed} = \lambda R^2_{error-free}$. Thus, we divide the observed R^2 by the ICC (or marginal test reliability for Mathematical Knowledge) of the observed predictor, λ, to recover an estimate of the error-free R^2.

4. To generate other estimates of the effect of measurement error in the combined models requires detailed assumptions about the covariance across predictors in both the underlying error-free components and the measurement errors, which is beyond the scope of this paper.

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.

Authors. 2016.

Armor, David, Patricia Conry-Oseguera, Millicent Cox, Nicelma King, Lorraine McDonnell, Anthony Pascal, Edward Pauly, and Gail Zellman. 1976. *Analysis of the School Preferred Reading Programs in Selected Los Angeles Minority Schools, REPORT NO. R*-2007-*LAUSD*. Santa Monica, CA: Rand Corporation.

Ball, Deborah Loewenberg, Mark Hoover Thames, and Geoffrey Phelps. 2008. "Content Knowledge for Teaching: What Makes it Special?" *Journal of Teacher Education* 59(5): 389-407.

Bandura, Albert. 1977. "Self-Efficacy: Toward a Unifying Theory of Behavioral Change." *Psychological Review* 84(2): 191-215.

Bell, Courtney A., Drew H. Gitomer, Daniel F. McCaffrey, Bridget K. Hamre, Robert C. Pianta, and Yi Qi. 2012. "An Argument Approach to Observation Protocol Validity." *Educational Assessment* 17(2-3): 62-87.

Blazar, David, and Matthew A. Kraft. 2017. "Teacher and Teaching Effects on Students' Attitudes and Behaviors." *Educational Evaluation and Policy Analysis* 39(1): 146-170.

Boonen, Tinneke, Jan Van Damme, and Patrick Onghena. 2014. "Teacher Effects on Student Achievement in First Grade: Which Aspects Matter Most?" *School Effectiveness and School Improvement* 25(1): 126-152.

Carlisle, Joanne F., Richard Correnti, Geoffrey Phelps, and Ji Zeng. 2009. "Exploration of the Contribution of Teachers' Knowledge About Reading to Their Students' Improvement in Reading." *Reading and Writing* 22(4): 457-486.

Carlisle, Joanne F., Ben Kelcey, Brian Rowan, and Geoffrey Phelps. 2011. "Teachers' knowledge about early reading: Effects on students' gains in reading achievement." *Journal of Research on Educational Effectiveness* 4(4): 289-321.

Carpenter, Thomas P., Elizabeth Fennema, Penelope L. Peterson, and Deborah A. Carey. 1988. "Teachers' pedagogical content knowledge of students' problem solving in elementary arithmetic." *Journal for Research in Mathematics Education*: 385-401.

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.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 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. 2007. "Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects." *Economics of Education Review* 26(6): 673-682.

Cochran-Smith, Marilyn, Matthew Cannady, Kirstin P. McEachern, Kara Viesca, Peter Piazza, Christine Power, and Amy Ryan. 2012. "Teachers' education and outcomes: Mapping the research terrain." *Teachers College Record* 114(10): 1-49.

Cohen, Julie, and Dan Goldhaber. 2016. "Observations on Evaluating Teacher Performance: Assessing the Strengths and Weaknesses of Classroom Observations and Value-Added Measures." In *Improving Teacher Evaluation Systems: Making the Most of Multiple Measures,* ed. Jason A. Grissom and Peter Youngs. New York, NY: Teachers College Press.

Conover, William J. 1999. *Practical Nonparametric Statistics. Third Edition*. New York, NY: John Wiley & Sons, Inc.

Cooley, William W., and Gaea Leinhardt. 1980. "The Instructional Dimensions Study." *Educational Evaluation and Policy Analysis* 2(1): 7-25.

Cowan, James, Dan Goldhaber, Kyle Hayes, and Roddy Theobald. 2016. "Missing Elements in the Discussion of Teacher Shortages." *Educational Researcher* 45(8): 460-462.

Gamoran, Adam, Andrew C. Porter, John Smithson, and Paula A. White. 1997. "Upgrading High School Mathematics Instruction: Improving Learning Opportunities for Low-Achieving, Low-Income Youth." *Educational Evaluation and Policy Analysis* 19(4): 325-338.

Garet, Michael S., Jessica B. Heppen, Kirk Walters, Julia Parkinson, Toni M. Smith, Mengli Song, Rachel Garrett, Rui Yang, and Geoffrey D. Borman. 2016. "Focusing on Mathematical Knowledge: The Impact of Content-Intensive Teacher Professional Development. NCEE 2016-4010." Washington, DC: National Center for Education Evaluation and Regional Assistance.

Goldhaber, Dan D., and Dominic J. Brewer. 2000. "Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement." *Educational Evaluation and Policy Analysis* 22(2): 129-145.

Goldhaber, Dan D., Dominic J. Brewer, and Deborah J. Anderson. 1999. "A Three-Way Error Components Analysis of Educational Productivity." *Education Economics* 7(3): 199-208.

Good, Thomas L., and Douglas A. Grouws. 1977. "Teaching Effects: A Process-Product Study in Fourth-Grade Mathematics Classrooms." *Journal of Teacher Education* 28(3): 49-54.

Grossman, Pam, Susanna Loeb, Julie Cohen, and James Wyckoff. 2013. "Measure for Measure: The Relationship Between Measures of Instructional Practice in Middle School English

Language Arts and Teachers' Value-Added Scores." *American Journal of Education* 119(3): 445-470.

Hamilton, Laura. 2003. "Assessment as a Policy Tool." *Review of Research in Education* 27(1): 25-68.

Hanushek, Eric A. 1986. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature* 24(3): 1141-1177.

Harris, Douglas N., and Tim R. Sass. 2011. "Teacher Training, Teacher Quality and Student Achievement." *Journal of Public Economics* 95(7): 798–812.

Hill, Heather C., Merrie L. Blunk, Charalambos Y. Charalambous, Jennifer M. Lewis, Geoffrey C. Phelps, Laurie Sleep, and Deborah Loewenberg Ball. 2008. "Mathematical Knowledge for Teaching and the Mathematical Quality of Instruction: An Exploratory Study." *Cognition and Instruction* 26(4): 430–511.

Hill, Heather C., Brian Rowan, and Deborah Loewenberg Ball. 2005. "Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement." *American Educational Research Journal* 42(2): 371–406.

Ho, Andrew D., and Thomas J. Kane. 2013. "The Reliability of Classroom Observations by School Personnel." Seattle, WA: Bill and Melinda Gates Foundation.

Hoge, Robert D., and Theodore Coladarci. 1989. "Teacher-Based Judgments of Academic Achievement: A Review of Literature." *Review of Educational Research* 59(3): 297-313.

Jackson, C. Kirabo. 2016. "What Do Test Scores Miss? The Importance of Teacher Effects on Non-test Score Outcomes." Working Paper 22226, National Bureau of Economic Research.

Jacob, Brian A. 2005. "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics* 89(5-6): 761-796.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City." *Economics of Education Review* 27(6): 615-631.

Kane, Thomas J. and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Working Paper 14607, National Bureau of Economic Research.

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." Seattle, WA: Bill and Melinda Gates Foundation.

Koretz, Daniel. 2008. *Measuring Up. What Educational Testing Really Tells Us*. Cambridge, MA: Harvard University Press.

Koretz, Daniel M., Daniel F. McCaffrey, and Laura S. Hamilton. 2001. "Toward a Framework for Validating Gains Under High-Stakes Conditions." Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing.

Kunter, Mareike, Uta Klusmann, Jürgen Baumert, Dirk Richter, Thamar Voss, and Axinja Hachfeld. 2013. "Professional Competence of Teachers: Effects on Instructional Quality and Student Development." *Journal of Educational Psychology* 105(3): 805.

Lampert, Magdalene. 2001. *Teaching Problems and the Problems of Teaching*. New Haven, CT: Yale University Press.

Lavy, Victor. 2009. "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics." *American Economic Review* 99(5): 1979-2011.

Monk, David H. 1994. "Subject Area Preparation of Secondary Mathematics and Science Teachers and Student Achievement." *Economics of Education Review* 13(2): 125-145.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119(1): 39-77.

Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges. 2004. "How Large are Teacher Effects?" *Educational Evaluation and Policy Analysis* 26(3): 237-257.

Palardy, Gregory J., and Russell W. Rumberger. 2008. "Teacher Effectiveness in First Grade: The Importance of Background Qualifications, Attitudes, and Instructional Practices for Student Learning." *Educational Evaluation and Policy Analysis* 30(2): 111-140.

Papay, John P., and Matthew A. Kraft. 2015. "Productivity Returns to Experience in the Teacher Labor Market: Methodological Challenges and New Evidence on Long-Term Career Improvement." *Journal of Public Economics* 130(10): 105-119.

Pianta, Robert C., Jay Belsky, Nathan Vandergrift, Renate Houts, and Fred J. Morrison. 2008. "Classroom Effects on Children's Achievement Trajectories in Elementary School." *American Educational Research Journal* 45(2): 365-397.

Pianta, Robert C., Karen M. LaParo, and Bridget K. Hamre. 2007. *Classroom Assessment Scoring System (CLASS) Manual.* Baltimore, MD: Brookes Publishing.

Reyes, Maria R., Marc A. Brackett, Susan E. Rivers, Mark White, and Peter Salovey. 2012. "Classroom Emotional Climate, Student Engagement, and Academic Achievement." *Journal of Educational Psychology* 104(3): 700-712.

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. 2011. "Can You Recognize an Effective Teacher When You Recruit One?" *Education Finance and Policy* 6(1): 43-74.

Ross, John A. 1992. "Teacher Efficacy and The Effects of Coaching on Student Achievement." *Canadian Journal of Education/Revue Canadienne de l'education* 17(1): 51-65.

Rowan, Brian, Fang-Shen Chiang, and Robert J. Miller. 1997. "Using Research on Employees' Performance to Study the Effects of Teachers on Students' Achievement." *Sociology of Education 70(4)*: 256-284.

Rowan, Brian, Richard Correnti, and Robert J. Miller. 2002. "What Large-Scale, Survey Research Tells Us about Teacher Effects on Student Achievement: Insights from the Prospects Study of Elementary Schools." Consortium for Policy Research in Education, Graduate School of Education, University of Pennsylvania.

Rubin, Donald B. 1996. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association* 91(434): 473-489.

Sadler, Philip M., Gerhard Sonnert, Harold P. Coyle, Nancy Cook-Smith, and Jaimie L. Miller. 2013. "The Influence of Teachers' Knowledge on Student Learning in Middle School Physical Science Classrooms." *American Educational Research Journal* 50(5): 1020-1049.

Shulman, Lee S. 1986. Paradigms and Research Programs in the Study of Teaching: A Contemporary Perspective. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 3–36). New York, NY: Macmillan.

Stronge, James H., Thomas J. Ward, and Leslie W. Grant. 2011. "What Makes Good Teachers Good? A Cross-Case Analysis of the Connection Between Teacher Effectiveness and Student Achievement." *Journal of Teacher Education* 62(4): 339-355.

Sutcher, Leib, Linda Darling-Hammond and Desiree Carver-Thomas. 2016. "A Coming Crisis in Teaching? Teacher Supply, Demand, and Shortages." Palo Alto, CA: Learning Policy Institute.

Tyler, John H., Eric S. Taylor, Thomas J. Kane, and Amy L. Wooten. 2010. "Using Student Performance Data to Identify Effective Classroom Practices." *American Economic Review* 100(2): 256-60.

Wayne, Andrew J., and Peter Youngs. 2003. "Teacher Characteristics and Student Achievement Gains: A Review." *Review of Educational Research* 73(1): 89-122.

Table 1

Description of Teacher-Level Measures

Measure	Description of Measure	Data Source	Collection Schedule
Measures of Instructional Practices and Content			
Ambitious Instruction	Meaning orientation and cognitive demand of observed mathematics instruction	Class Observation	3x year
Errors	Observed teacher mathematical mistakes or imprecisions	Class Observation	3x year
Classroom Organization	Observed classroom climate, productivity, and behavior management	Class Observation	3x year
Support	Extent of observed emotional and instructional support provided by the teacher	Class Observation	3x year
Algebra Content	Number of subtopics taught covering algebra	Teacher Survey	Spring
Number & Operations Content	Number of subtopics taught covering number and operations	Teacher Survey	Spring
Test Prep Activities	Specific test preparation activities in the classroom	Teacher Survey	Fall
Test Prep Instructional Changes	General instructional changes to align with standardized tests	Teacher Survey	Fall
Measures of Personal Characteristics			
Mathematical Knowledge	Mathematical knowledge for teaching and general mathematics knowledge	Teacher Survey	Fall
Knowledge of Student Performance	Knowledge of students' mathematical ability and performance on tests	Teacher Survey	Spring
Self-Efficacy	Teachers' self-efficacy in providing strong instruction to students	Teacher Survey	Fall
Effort	Number of hours spent preparing for class, organizing materials, grading, etc.	Teacher Survey	Fall
Measures of Background			
Master's Degree	Teacher holds any master's degree	Teacher Survey	Fall
Math Major	Teacher holds an undergraduate or graduate degree in mathematics	Teacher Survey	Fall
Education Bachelor's	Teacher holds a bachelor's degree in education	Teacher Survey	Fall
Math Courses	The number of undergraduate or graduate math courses taken	Teacher Survey	Fall
Math Content Courses	The number of undergraduate or graduate math content courses taken	Teacher Survey	Fall
Traditional Certification	Teacher holds a traditional teaching certification	Teacher Survey	Fall
Alternative Certification	Teacher holds an alternative (e.g., TFA) teaching certification	Teacher Survey	Fall
Elementary Math Certification	Teacher holds a specialized certification for teaching elementary mathematics	Teacher Survey	Fall
4–10 Yrs. Experience	Teacher has 4 to 10 years of teaching experience	Teacher Survey	Fall
10+ Yrs. Experience	Teacher has more than 10 years of teaching experience	Teacher Survey	Fall

Table 2

Descriptive Statistics and Correlations of Teacher-Level Measures

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. Ambitious Instruction	.00	1.00	-																				
2. Errors	.00	1.00	-.34	-																			
3. Class Organization	.00	1.00	.21	.01	-																		
4. Support	.00	1.00	.32	-.09	.40	-																	
5. Algebra Content	.00	1.00	.07	-.12	.14	.15	-																
6. Number & Operations Content	.00	1.00	.04	.02	.06	.05	.45	-															
7. Test Prep Activities	.00	1.00	-.17	.24	.16	.04	.10	.16	-														
8. Test Prep Instructional Changes	.00	1.00	-.13	-.02	-.03	-.10	.00	-.05	.14	-													
9. Mathematical Knowledge	.00	1.00	.38	-.42	.09	.11	.21	.08	-.21	.01	-												
10. Knowledge of Student Performance	.00	1.00	.19	-.20	.02	.00	.08	.06	-.23	.14	.25	-											
11. Self-Efficacy	.00	1.00	.03	.01	.08	.10	.27	.27	.11	-.24	.01	-.10	-										
12. Effort	.00	1.00	-.08	.24	.11	.06	.14	.18	.26	-.09	-.13	-.07	.11	-									
13. Master's Degree	.78	.41	-.01	.00	-.11	-.09	-.01	.08	.00	.01	.03	.05	-.09	-.08	-								
14. Math Major	.07	.25	.11	-.06	.11	.04	.05	.03	.14	-.11	.07	.07	.09	-.01	.00	-							
15. Ed. Bachelor's	.57	.50	.02	-.05	.10	.05	.04	-.04	.03	.02	-.05	-.02	.01	.00	-.22	.02	-						
16. Math Courses	2.88	.86	.04	-.04	.08	.09	.07	.08	.22	-.06	.01	-.04	.09	.21	-.06	.18	-.02	-					
17. Math Content Courses	2.53	.81	.06	.01	.06	.06	.09	.02	.14	-.12	-.01	-.03	.06	.13	.01	.03	.06	.47	-				
18. Traditional Certification	.87	.34	.03	.03	.06	.03	-.01	-.10	-.07	.11	.08	.13	-.08	-.08	.04	-.22	.30	-.09	.04	-			
19. Alternative Certification	.05	.22	-.02	-.07	.00	.01	.10	.14	.04	-.02	.08	-.04	.10	.07	-.02	.07	-.24	.06	-.14	-.59	-		
20. Elementary Math Certification	.16	.37	.00	-.03	-.04	.15	.11	.05	-.02	-.04	.03	.00	.09	-.05	-.01	.29	.15	.16	.09	-.17	.02	-	
21. 4-10 Years Experience	.45	.50	.07	-.06	.10	.06	-.13	.09	-.03	-.08	.05	.04	.09	-.12	.10	.11	-.03	-.08	-.14	.01	-.01	-.03	-
22. 10+ Years Experience	.43	.50	-.03	.05	-.07	-.07	.12	-.08	.05	.06	-.08	-.01	-.03	.08	.05	-.07	.08	.11	.24	.06	-.05	.06	-.79

Table 3

Estimates of Teacher-Level Parameters and Adjusted Teacher-level R^2 of State-Administered Assessments

	M0	M1	M2	M3	M4	M5	M6
Ambitious Instruction		-0.006	0.010			0.001	0.004
Errors		-0.013	-0.021			-0.017	-0.015
Class Organization		0.022	0.031~			0.031~	0.024
Support		0.005	0.004			0.004	0.007
Algebra Content		0.035*	0.035*			0.027	0.043*
Number & Operations Content		-0.007	-0.000			0.004	-0.009
Test Prep Activities		0.006	0.007			0.013	0.001
Test Prep Instructional Changes		-0.019	-0.016			-0.023	-0.011
Mathematical Knowledge		0.018		0.030*		0.017	
Knowledge of Student Performance		0.035*		0.036*		0.040*	
Self-Efficacy		-0.011		0.001		-0.011	
Effort		0.034*		0.038**		0.027~	
Master's Degree		0.034			0.043		0.041
Math Major		0.043			0.069		0.048
Ed. Bachelor's		0.060~			0.067*		0.058~
Math Courses		-0.012			-0.002		-0.007
Math Content Courses		0.041*			0.051*		0.044*
Traditional Certification		0.054			0.078		0.071
Alternative Certification		0.068			0.104		0.073
Elementary Math Certification		-0.040			-0.037		-0.046
4-10 Years Experience		0.082			0.055		0.062
10+ Years Experience		0.010			-0.011		-0.005
Intercept	1.104~	0.864	1.059~	1.213*	0.748	1.180~	0.759
School Variance	0.007	0.006	0.006	0.006	0.006	0.006	0.006
Teacher Variance	0.032	0.022	0.029	0.028	0.028	0.025	0.025
Classroom Variance	0.017	0.018	0.017	0.018	0.017	0.018	0.017
Residual Variance	0.307	0.307	0.307	0.307	0.307	0.307	0.307
Adjusted Teacher-level R^2		25.1%	9.0%	12.4%	10.3%	17.7%	15.9%

Note: This table presents teacher-level parameters and adjusted teacher-level R^2 from hierarchical model where the outcome variable is student scores on state-administered assessments. Sample includes 7,843 students and 283 teachers. The model also includes student-, class-, and cohort-level controls for test scores and demographic characteristics. Adjusted teacher-level R^2 indicates the proportional reduction in teacher-level variance (from the baseline model) after including the additional teacher-level controls specified in each model. We adjust the teacher-level R^2 estimate to account for the number of additional teacher level controls in the model.
~ p <.10; *p <.05; **p <.01; ***p <.001.

Table 4

Estimates of Teacher-Level Parameters and Adjusted Teacher-level R^2 from Individual Regressions of State-Administered Assessments

	Parameter Estimate	Adjusted Teacher-level R^2	Error-Free Adjusted Teacher-level R^2
Ambitious Instruction	0.026~	0.9%	1.3%
Errors	-0.023	1.7%	3.3%
Class Organization	0.042**	3.1%	4.8%
Support	0.026~	0.8%	1.6%
Algebra Content	0.045**	6.3%	7.8%
Number & Operations Content	0.025	2.0%	2.4%
Test Prep Activities	0.010	-0.1%	-0.1%
Test Prep Instructional Changes	-0.015	0.9%	1.0%
Mathematical Knowledge	0.032*	4.3%	5.1%
Knowledge of Student Performance	0.035*	4.0%	4.5%
Self-Efficacy	0.000	-0.3%	-0.4%
Effort	0.032*	5.1%	6.2%
Master's Degree	0.022	-0.2%	-0.2%
Math Major	0.043	0.6%	0.6%
Ed. Bachelor's	0.072*	1.3%	1.3%
Math Courses	0.018	1.1%	1.1%
Math Content Courses	0.043*	4.6%	4.6%
Traditional Certification	0.067	-0.8%	-0.8%
Alternative Certification	-0.016	-0.8%	-0.8%
Elementary Math Certification	-0.022	-0.7%	-0.7%
4-10 Years Experience	0.053~	2.2%	2.2%
10+ Years Experience	-0.024	0.6%	0.6%
Total		39.5%	48.3%

Note: Sample includes 283 teachers and 7,843 students. Adjusted teacher-level R^2 indicates the proportional reduction in teacher-level variance (from the baseline model in Table 3) after including the additional teacher-level control variable specified on each row in the table. We adjust the teacher-level R^2 estimate to account for the number of additional teacher level controls in the model (see the analysis section of the paper for details). The error-free adjusted teacher-level R^2 weights the original adjusted teacher-level R^2 by the inverse of each predictor's reliability ratio. We set the reliability of demographic characteristics (e.g., race, gender) and credentials (e.g., degrees, experience) that come from administrative files to one. Each row corresponds to a different regression model with only one teacher-level variable.

~ p <.10; *p <.05; **p <.01; ***p <.001.

Table 5

Estimates of Teacher-Level Parameters and Adjusted Teacher-level R^2 of the Project-Administered Assessment

	M0	M1	M2	M3	M4	M5	M6
Ambitious Instruction		-0.003	0.008			-0.001	0.005
Errors		0.001	-0.010			-0.001	-0.007
Class Organization		0.024*	0.028*			0.027*	0.025*
Support		-0.008	-0.008			-0.008	-0.009
Algebra Content		0.004	0.010			0.004	0.010
Number & Operations Content		0.012	0.014			0.018	0.009
Test Prep Activities		-0.014	-0.018			-0.009	-0.023~
Test Prep Instructional Changes		-0.014	-0.010			-0.017	-0.008
Mathematical Knowledge		0.023~		0.029**		0.024*	
Knowledge of Student Performance		0.027*		0.026*		0.029*	
Self-Efficacy		-0.002		0.004		-0.003	
Effort		0.000		0.006		0.001	
Master's Degree		0.012			0.018		0.019
Math Major		0.014			0.028		0.027
Ed. Bachelor's		0.022			0.024		0.017
Math Courses		0.002			-0.000		0.002
Math Content Courses		0.035*			0.041**		0.036*
Traditional Certification		0.054			0.076~		0.079~
Alternative Certification		0.100~			0.124*		0.114~
Elementary Math Certification		-0.021			-0.020		-0.021
4-10 Years Experience		0.004			-0.003		-0.003
10+ Years Experience		-0.012			-0.026		-0.021
Intercept	0.661	0.463	0.584	0.715	0.426	0.680	0.342
School Variance	0.006	0.005	0.006	0.006	0.005	0.006	0.005
Teacher Variance	0.010	0.006	0.009	0.008	0.008	0.007	0.007
Classroom Variance	0.011	0.012	0.011	0.012	0.012	0.012	0.011
Residual Variance	0.270	0.270	0.270	0.270	0.270	0.270	0.270
Adjusted Teacher-level R^2		38.2%	11.4%	23.8%	13.9%	28.8%	22.9%

Note: This table presents teacher-level parameters and adjusted teacher-level R^2 from hierarchical model where the outcome variable is student scores on the project-administered assessment. Sample includes 7,843 students and 283 teachers. The model also includes student-, class-, and cohort-level controls for test scores and demographic characteristics. Adjusted teacher-level R^2 indicates the proportional reduction in teacher-level variance (from the baseline model) after including the additional teacher-level controls specified in each model. We adjust the teacher-level R^2 estimate to account for the number of additional teacher level controls in the model.
~ p <.10; *p <.05; **p <.01; ***p <.001

Table 6

Estimates of Teacher-Level Parameters and Adjusted Teacher-level R^2 from Individual Regressions of the Project-Administered Assessment

	Parameter Estimate	Adjusted Teacher-level R^2	Error-Free Adjusted Teacher-level R^2
Ambitious Instruction	0.019	2.0%	2.9%
Errors	-0.019	2.2%	4.2%
Class Organization	0.027~	2.1%	3.2%
Support	0.004	-0.3%	-0.6%
Algebra Content	0.024~	3.5%	4.3%
Number & Operations Content	0.030*	5.5%	6.5%
Test Prep Activities	-0.010	-0.3%	-0.4%
Test Prep Instructional Changes	-0.013	0.7%	0.8%
Mathematical Knowledge	0.040***	14.1%	16.6%
Knowledge of Student Performance	0.031*	9.9%	11.1%
Self-Efficacy	0.002	-0.1%	-0.1%
Effort	0.004	0.0%	0.0%
Master's Degree	0.010	-0.3%	-0.3%
Math Major	0.001	-0.3%	-0.3%
Ed. Bachelor's	0.032	-1.9%	-1.9%
Math Courses	0.022	2.6%	2.6%
Math Content Courses	0.040**	7.9%	7.9%
Traditional Certification	0.053	-0.6%	-0.6%
Alternative Certification	0.035	1.3%	1.3%
Elementary Math Certification	-0.030	0.0%	0.0%
4-10 Years Experience	0.019	1.6%	1.6%
10+ Years Experience	-0.010	0.8%	0.8%
Total		53.2%	62.4%

Note: Sample includes 283 teachers and 7,843 students. Adjusted teacher-level R^2 indicates the proportional reduction in teacher-level variance (from the baseline model in Table 5) after including the additional teacher-level control variable specified on each row in the table. We adjust the teacher-level R^2 estimate to account for the number of additional teacher level controls in the model (see the analysis section of the paper for details). The error-free adjusted teacher-level R^2 weights the original adjusted teacher-level R^2 by the inverse of each predictor's reliability ratio. We set the reliability of demographic characteristics (e.g., race, gender) and credentials (e.g., degrees, experience) that come from administrative files to one. Each row corresponds to a different regression model with only one teacher-level variable.

~ p <.10; *p <.05; **p <.01; ***p <.001.