Explaining Teacher Effects on Achievement

Using Measures from Multiple Research Traditions

Andrew Bacher-Hicks, Mark 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

Researchers have identified many characteristics of teachers and teaching that contribute to student outcomes. However, most studies investigate only a small number of these characteristics, likely underestimating the overall contribution. In this paper, we use a set of 28 teacher-level predictors drawn from multiple research traditions to explain teacher-level variation in student outcomes. These predictors collectively explain 28% of teacher-level variability in state standardized math test scores and 40% in a predictor-aligned math test. In addition, each individual predictor explains only a small, relatively unique portion of the total teacher-level variability. This first finding highlights the importance of choosing predictors and outcomes that are well aligned, and the second suggests that the phenomena underlying teacher effects is multidimensional.

Keywords: teacher effectiveness; hierarchical modeling; teacher characteristics/traits; instructional practices; teacher knowledge

Explaining Teacher Effects on Achievement

Using Measures from Multiple Research Traditions

Research on teachers and teaching has generally proceeded along two separate tracks. In one, economists have identified substantial differences among teachers, with a one standard deviation increase in teacher effects typically associated with a 0.10 to 0.20 standard deviation difference in student test score outcomes (Aaronson, Barrow, & Sander, 2007; Jacob & Lefgren, 2008; Kane, Rockoff, & Staiger, 2008; Kane & Staiger, 2008; Nye, Konstantopoulos, & Hedges, 2004; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). Yet despite wide consensus around the substantive significance and approximate size of these effects, the indicators conventionally used by economists, such as teacher experience, degree completion, and entry route, seldom explain more than a small fraction of teacher-level variation in scores.

A second track has for decades described, measured, and correlated hundreds of variables representing characteristics of teachers and teaching other than those explored by economists. Several subfields exist within the literature of teachers and teaching. For example, scholars focusing on teacher characteristics have examined teachers' locus of control (Rose & Medway, 1981), knowledge of students' thinking (Sadler, Sonnert, Coyle, Cook-Smith, & Miller, 2013), mathematical knowledge and mathematical knowledge for teaching (Begle, 1972; Harbison & Hanushek, 1992; Hill, Rowan, & Ball, 2005), and efficacy (Tschannen-Moran, Hoy, & Hoy, 1998). Scholars focusing on teaching have examined factors such as classroom organization and behavior management (Brophy & Good, 1986; Stronge, Ward, & Grant, 2011), content-specific instructional practices (Grossman et al., 2010; Stein & Lane, 1996), formative assessment (Wiliam, Lee, Harrison, & Black, 2004), and classroom climate (Pianta, LaParo, & Hamre, 2007). By comparing these characteristics and factors with student gains on standardized tests,

scholars have identified characteristics of effective teachers and classrooms. Yet similar to the econometric studies, each characteristic typically explains only a small amount of the variability in teacher contributions to student outcomes (e.g., Hill et al., 2005).

Thus these two separate research traditions have arrived at a similar position: Some teacher-level predictors of student outcomes have been identified, yet the amount of explained teacher-level variance in student outcome models is small. One reason might be that, despite the extensive literature in both fields, methods and findings remain specialized within each. Another reason might be that research on teachers and teaching often focuses on developing and refining individual constructs rather than synthesizing and comparing many across these subfields. In fact, on only a few occasions have scholars tested multiple distinct explanations for the role teacher and teaching quality plays in student outcomes (e.g., Boonen, Van Damme, & Onghena, 2013; Palardy & Rumberger, 2008; Stronge et al., 2011). This means that while we know that multiple teacher and teaching characteristics relate to student outcomes, we know little about which matter most, or how much these characteristics *together* contribute to explaining variation in teacher effects.

To address this issue, we measured multiple components of teacher and teaching quality over two academic years, and in this paper relate those to student outcomes on both a state standardized and a low-stakes alternative assessment that was aligned to several predictor variables. We use hierarchical linear modeling to identify the amount of variance in student outcomes due to differences between teachers, before and after the inclusion of different teacher-level predictors. We find that predictors from both the economics and education literature explain, in conjunction, up to 28% of teacher effects on state test outcomes and 40% of teacher effects on the outcomes for the alternative assessment. Further, no specific set of predictors

(e.g., teacher preparation, instruction, knowledge) appears to alone explain significant amounts of the observed teacher-level variation of either outcome.

In what follows, we review the literature on explaining teacher effects and then describe our methods and results. Two notes on the language used throughout the paper: First, although there is considerable disagreement over whether differences among teachers in average student outputs are causal or the result of sorting of students to teachers (Kane & Staiger, 2008; Kane, McCafrey, Miller, & Staiger, 2013; Rothstein, 2009), for convenience, we refer throughout this paper to teacher *effects*. Second, we and other studies can differentiate between *classroom-level effects*, which refer to classroom-years within teachers, and teacher effects, and adopt this nomenclature to signify differences between the two.

Background

Recent studies consistently show that student achievement outcomes differ meaningfully by teacher assignment. Using mainly non-experimental methods, researchers have shown that a 1-standard-deviation difference in teacher effects is associated with a 0.08- to 0.12-standard-deviation unit difference in reading outcomes and 0.11- to 0.26-standard-deviation unit difference in math outcomes (Aaronson et al., 2007; Jacob & Lefgren, 2008; Kane et al., 2008; Rivkin et al., 2005; Rockoff, 2004). A study that took advantage of random assignment of students to teachers (Nye et al., 2004) found similar effects of between 0.06 and 0.10 for reading, and 0.10 and 0.14 for mathematics, while another (Kane & Staiger, 2008) found slightly larger effects of 0.18 for reading and 0.20 for mathematics. Although economists, statisticians, and policy-makers have debated the validity of using student test scores for teacher accountability, few disagree that student outcomes on standardized tests vary systematically according to teacher assignment.

Given that differences among teachers have consistently been identified, analysts have additionally asked whether this variation can be explained by teacher or classroom characteristics. For this paper, we focus on two aspects of this literature: the teacher and classroom characteristics associated with differences in student test score outcomes, and the amount of variation in teacher effects explained by prior analyses.

What Teacher-Level Measures Explain Student Achievement Growth?

The last three decades have seen considerable effort around identifying teacher and teaching characteristics that explain variability in student outcomes. In a long line of research, economists have attempted to explain such variability using indicators of teacher preparation found in administrative datasets, such as educational attainment, coursework, certification, and experience. Results have varied. For example, Hanushek (1986) analyzed 147 education production function studies and concluded that "the results are startlingly consistent in finding no strong evidence that teacher-student ratios, teacher education, or teacher experience have an expected positive effect on student achievement" (p. 1162). With a more stringently selected sample of only 60 studies, however, Greenwald, Hedges, and Laine (1996) argued that factors such as teacher education, experience, and smaller classes are positively related to student achievement. Additionally, Wayne and Youngs (2003) conducted a narrative review of 21 studies, concluding that "in the case of degrees, coursework, and certification, findings have been inconclusive except in mathematics, where high school students clearly learn more from teachers with certification in mathematics, degrees related to mathematics, and coursework related to mathematics" (p. 107). More recently, several studies have also found that early-career teachers typically have weaker student outcomes than others (Chetty et al., 2011; Kane & Staiger, 2008; Rockoff, 2004). As these disparate results suggest, with the possible exception of

teacher experience, little consensus exists around the specific teacher background characteristics related to student outcomes.

It is similarly difficult to summarize studies that use surveys and assessments to measure teacher and teaching characteristics. For instance, recent investigations into teachers' subject matter knowledge have found generally positive results regarding the relationship between this knowledge and student outcomes (Baumert et al., 2010; Carlisle, Kelcey, Rowan, & Phelps, 2011; Carpenter, Fennema, Peterson, & Carey, 1988; Helmke & Schrader, 1987; Hill et al., 2005; Hoge & Coladarci, 1989; Metzler & Woessmann, 2012; Rockoff, Jacob, Kane, & Staiger, 2011; Shechtman, Roschelle, Haertel, & Knudsen, 2010). However, the definition and operationalization of teacher knowledge varies markedly across projects (Depaepe, Verschaffel, & Kelchtermans, 2013), and a number of divergent findings exist (Kane & Cantrell, 2013; Kersting, Givvin, Thompson, Santagata, & Stigler, 2012).

Other survey-based studies have offered an array of teacher-level variables thought to be related to student outcomes. For example, a recent study measuring teachers' knowledge of student misconceptions found that stronger knowledge of such misconceptions was associated with stronger student outcomes (Sadler et al., 2013). Scholars interested in teachers' locus of control and self-efficacy have identified associations between these variables, which summarize teachers' sense of success in teaching and ability to reach students, and student outcomes (Armor et al., 1976; Ross, 1992). Alignment of the material taught in class to the material being tested is also important: In a classic study, Cooley and Leinhardt (1980) found students' opportunity to learn what was on the test—literally, the match between the test and teachers' enacted curriculum—explained the largest portion of student outcomes for all grade-test combinations studied. Gamoran, Porter, Smithson, & White (1997) similarly found that content coverage

accounts for much of the differences in student achievement across different levels of high school mathematics instruction. Increased teacher effort is thought to at least partially explain the success of teacher incentive programs (Muralidharan & Sundararaman, 2011) on improving student outcomes. Teachers' use of formative assessment has also been identified in the literature as being related to increased achievement (Black & Wiliam, 1998; for dissenting views, see Briggs, Araceli Ruiz-Primo, Furtak, Shepard, & Yin, 2012; Kingston & Nash, 2011). Common to most of these studies, however, is the identification and testing of only one feature of teaching or teachers for comparison with student outcomes.

By contrast, research that uses observational metrics has often made such explicit comparisons, but only within the subfield itself. Beginning with the process-product work of the 1970s (Brophy & Good, 1986), and continuing through the recent Measures of Effective Teaching study (Kane & Staiger, 2012), results consistently suggest that classroom climate and organization have stronger relationships to student outcomes than other instructional features, such as student inquiry or discipline-specific features. For example, Stronge and colleagues (2011) observed that variables focused on classroom climate and management best differentiate between teachers with high and low value-added scores. Bell and colleagues (2012) found that the classroom organization scale of the Classroom Assessment Scoring System (CLASS) best predicts student gains in high school algebra classrooms. Tyler, Taylor, Kane, & Wooten (2010) concluded that, in Cincinnati, having relatively better scores on the classroom management dimension than the instruction dimension of Framework for Teaching predicted student outcomes in mathematics and reading; the contrast between inquiry-oriented practices and routine instruction is only significant for reading. In English language arts, Grossman and colleagues (2010) found teachers' explicit strategy instruction and student engagement to

differentiate between teachers with high and low value-added scores. These findings are striking, for while many items of these observational instruments describe inquiry-oriented instruction, very few appear significant.

Single studies rarely combine measures from these different sources—administrative databases, survey-based measures of teachers and teaching, and observation-based metrics—to understand the educational production function. With the use of a single method for data collection, most studies are limited by what cannot be measured; observational instruments, for instance, rarely include metrics capturing formative assessment techniques, for these are difficult to observe in only a few lessons. This has several implications for the field. Without multiple distinct measures, it is impossible to estimate the extent of overlap between important variables—for instance, how much teachers' subject-matter knowledge coursework and their subject-matter knowledge per se correlate. In such cases, researchers also cannot identify how much each measure predicts student outcomes after accounting for other important measures. These limitations of the existing research provide one motivation for our study.

How Much Variation in Teacher Effects Can Be Explained?

Few studies that attempt to predict teacher effects also provide estimates of the proportion of explained variance in teacher-level outcomes. However, among those that do, the proportion is typically small. Among studies that use preparation, experience, and knowledge to explain teacher effects, for instance, the proportion of explained variance ranges between 2% and 22%, with more comprehensive studies, in terms of variables deployed, explaining more variability in student outcomes. Studies with fewer variables include Goldhaber, Brewer, and Anderson (1999), who found that variables such as teacher race, certification, and degree explained around 2% of the teacher-level variance in National Education Longitudinal Study of

1988 data. Nye and colleagues (2004) observed that teacher experience and teacher education both independently explained less than 5% of the variance in teacher effects in the Tennessee class size experiment data. Palardy and Rumberger (2008) used the Early Childhood Longitudinal Study (ECLS) dataset and found that teacher certification status accounted for 2.4% of the classroom-level variance in first-grade reading achievement gains; Boonen et al. (2013) found that teacher experience and amount of in-service training accounted for 8.4% of the classroom-level variance in Belgian first graders' mathematics achievement gains. Studies using more variables include Hill et al. (2005), who, using data from a study of comprehensive school reform, explained 17% and 19% of first- and third-grade teacher-level variation, respectively, in student mathematics outcomes with variables representing teacher mathematics preparation, experience, mathematical knowledge for teaching, and average lesson length. In a study of Reading First sites, Carlisle, Correnti, Phelps, & Zeng (2009) found that teacher race, certification, and knowledge of reading explained between 5% and 22% of teacher-level variance in their models.

The addition of information about teachers' beliefs and instructional practices to such models appears to leave the overall picture unchanged; even after considering these indicators, such studies still fail to explain a large proportion of between-teacher differences in student outcomes. Using ECLS, Palardy and Rumberger (2008) found that measures such as teacher expectations of the impact of teaching and reported time spent on instruction and specific practices (i.e., silent reading, journal writing) accounted for 14.1% of the classroom-level variance in reading achievement gains. In mathematics, measures of teacher efficacy and instructional practice explained 8.9% of the classroom-level variance. Boonen et al. (2013) found that measures of teacher attitudes and instruction accounted for 10.6%, 19.4%, and 17% of

the Belgian classroom-level variance in math, reading, and spelling achievement gains, respectively.

Research Questions

Our review suggests that though the literature has identified many teacher measures that relate to teacher effects on student achievement, only about one fifth of the teacher-level variation in student outcomes, at best, has been modeled by such measures. One reason may be that, within specific studies, authors typically use a limited set of predictors; studies that focused on teacher knowledge, for instance, seldom included more than a few teacher background variables (e.g., Hill et al., 2005). Further, as noted above, most studies in this genre construct teacher variables from a single method of data collection, such as administrative records or a teacher survey, limiting potential explanations for student outcomes and thus limiting our knowledge of how a variety of different factors may contribute individually or jointly to student outcomes. Finally, most studies use instruments—including teacher and teaching metrics—that are not selected based on their alignment with the outcome variable.

These issues lead us to ask:

1. How much teacher-level variability in student outcomes can we explain using a rich set of predictors?
2. Does the amount of explained variation differ when student outcomes are constructed from state standardized assessments versus a low-stakes assessment designed to align with key teacher-level predictor variables?
3. What is the nature of the "production function" between predictors and student outcomes? For instance, is there a small set of teacher characteristics that explains

differences between teachers, or do different factors (e.g., preparation, teacher characteristics) play equal roles?

To answer these questions, we draw on a dataset that captures both student achievement and demographic data as well as information on 28 teacher-level predictors. The latter includes conventional predictors, such as experience and teaching preparation, teacher knowledge, teacher work habits, teaching behaviors, and variables representing instruction as observed by independent raters on two different observation instruments.

Because we measure theoretically related constructs (e.g., teachers' mathematical knowledge and the quality of their mathematics teaching) and place these in the same model, we cannot definitively identify the pathways between teacher quality, teaching quality, and student outcomes; this analysis is also not causal. With these two cautions in mind, however, we also ask:

4. Are there characteristics of teachers or teaching that consistently predict student outcomes across assessments?

With these goals of explaining differences in teacher effects and identifying qualities characteristic of effective teachers, we hope to contribute to the literature on how teachers influence student outcomes.

Data and Methods

Teacher Data

The teacher data reported here result from [identifying citation omitted for blind review], an effort to marry research on teacher effects with research on teachers and teaching. We collected teacher background, observational, and survey data from two academic years (2010–11 to 2011–12) in fourth- and fifth-grade mathematics classrooms across four large East Coast

public school districts. Theoretical guidance, primarily drawn from writings about the "instructional triangle" (Ball & Forzani, 2009; Cohen, Raudenbush, & Ball, 2003), and existing empirical explorations in the economic and teaching literature, referenced above, helped direct measure selection for the study.

Once measures were selected, data collection proceeded using four primary avenues: (a) we administered general teacher surveys in fall 2010 and fall 2011, with questions capturing teacher's mathematical knowledge,[1] beliefs, and behaviors; (b) we administered teacher grade-specific surveys in spring 2011 and spring 2012, with questions capturing teachers' knowledge of their students and teachers' coverage of the mathematical topics at their grade level; (c) we administered a survey once at the beginning of each teacher's participation in the study, with questions capturing background characteristics; and (d) we collected videos of classroom practice scored using the Mathematical Quality of Instruction (MQI) (Hill, Blunk, et al., 2008), and CLASS (Pianta et al., 2007) observation instruments. In Table 1, we categorize and describe the teacher and teaching measures and indicators used in our analyses. When appropriate, we report intraclass correlation coefficients (ICCs), or the amount of variance in teacher scores on the measures that is attributable to the teacher and not to other identified construct irrelevant sources of variation (e.g., lessons, survey items). These coefficients are adjusted for the modal number of lessons or survey items used to generate each measure, such that they reflect the teacher-level variance of the measure score for the typical teacher in our sample, as opposed to the teacher-level variance of the score for a single observation.

Insert Table 1.

Below we describe these measures and our methods for creating teacher scores on each. To facilitate interpretation and also to preserve differences in the distribution of teachers between

districts, for each metric we subtract the district mean from each individual teacher's score, and then we divide the result by the standard deviation of scores across the entire sample of teachers. This transformation preserves differences in variation across districts but sets the mean for each construct to be zero and the standard deviation of each construct to one across all districts (but not within districts).

Teacher Instruction Measures

We videotaped teachers' instruction on up to three different occasions each year across two school years.[2] Teachers selected the dates for taping, under the restrictions that they would choose lessons typical of their teaching and not choose lessons during which student testing would occur. Each taped lesson lasted approximately one hour and was scored by a set of trained CLASS and MQI raters.

CLASS. The CLASS is a subject-matter-independent observation tool organized to capture three primary domains of student-teacher interactions: emotional support, classroom organization, and instructional support. Each 15-minute segment of instruction was scored by a single rater; each code of the CLASS rubric is scored on a scale from one to seven. For a more complete description of what CLASS captures, see Pianta et al. (2007).

Exploratory factor analysis suggested that scores on the individual codes of the CLASS instrument formed two primary dimensions: classroom organization (Class Organization) and teacher emotional and instructional support (Support). To generate teacher scores for each of these two dimensions, we estimate the following multilevel lesson-level model, where lessons are nested within teachers:

$$CLASS_{j,k} = \beta_0 + \mu_k + \epsilon_{j,k} \quad (1)$$

The outcome of interest, $CLASS_{j,k}$, represents teacher k's Class Organization or Support score for lesson j. The model parameter μ_k represents teacher k's shrunken random effect on $CLASS_{j,k}$, adjusted for differences in the reliability of estimates from teacher to teacher due to differences in total number of lessons scored.[3]

MQI. The MQI observation instrument was developed to capture the quality of instruction on a set of mathematic-specific dimensions, including the meaning orientation of the mathematics presented to students, the teacher's ability to work with students and mathematics, teacher errors and imprecisions, and the extent to which students engaged in mathematical thinking and reasoning. Each 7.5-minute segment of instruction was scored on every code of the MQI rubric by two raters. For a more complete description of the MQI, see Hill, Blunk, et al. (2008).

Exploratory factor analysis suggested that scores on the individual codes of the MQI instrument formed three primary dimensions: whether the instruction in the classroom was connected to mathematics (Classroom Work Connected to Math), the errors and imprecision present in a teacher's instruction (Errors), and the extent to which instruction features mathematical meaning and student thinking and reasoning (Ambitious Instruction). To generate teacher scores for each of these three dimensions, we estimate the following multilevel lesson-level model, where lessons are nested within teachers:

$$MQI_{j,k} = \beta_0 + \mu_k + \epsilon_{j,k} \quad (2)$$

The outcome of interest, $MQI_{j,k}$, represents teacher k's Classroom Work Connected to Math, Errors, or Ambitious Instruction score for lesson j. From the equation, parameter μ_k represents teacher k's shrunken random effect on $MQI_{j,k}$. This teacher level MQI score has been adjusted

for differences in the reliability of estimates from teacher to teacher due to differences in total number of lessons scored.

Teacher Knowledge Measures

Biannual surveys contained questions designed to test several dimensions of teacher knowledge, including teacher mathematical knowledge for teaching (heretofore MKT; Hill, Rowan, & Ball, 2005), teacher general mathematical knowledge (derived from performance on released items from the Massachusetts Test for Educator Licensure, or MTEL), and teacher knowledge of their students' ability and misconceptions.

MKT and MTEL. Each fall, teachers completed a survey that included a teacher mathematical knowledge section. That section contained a mix of MKT items and items from the MTEL. We pooled teacher responses across years, giving us a total of 72 MKT items and 33 MTEL items.

A factor analysis of these items was ambiguous, in that there was no clear structure related to item origin (MKT versus MTEL) or item content (math knowledge specific to teaching versus common math knowledge, as judged by a blind panel of experts). Thus we pooled all items and treated them as a single, unidimensional test of teacher mathematical knowledge. Missing responses were scored as incorrect unless the teacher skipped six or more contiguous items nearby, in which case we scored items as "not presented" on the theory that responses were more reflective of other issues (e.g., time constraints, lack of interest in the assessment) than mathematical knowledge. Because MKT items contain testlets (a common stem producing several related items), we used a one-parameter graded response model in IRTPRO to estimate one overall score (Knowledge—Teaching and Content) for teacher mathematical knowledge.

Knowledge of students. Teachers in our sample were scored on two scales measuring knowledge of students. These measures were inspired by theories of teacher pedagogical content knowledge (Shulman, 1986), and one was adapted from an existing instrument measuring knowledge of student misconceptions (see Sadler et al., 2013). To generate these measures, teachers were presented with an item from the alternative math test administered by the project, and then were asked (a) what percent of their students would answer the item correctly (Student Ability Knowledge), and (b) from a set of incorrect answers, which answer would be most frequently chosen by their students (Student Misconception Knowledge). In 2010–11, we asked teachers of both Grades 4 and 5 these questions for 14 and 15 different student test items, respectively. In 2011–12, we asked both Grade 4 and Grade 5 teachers these questions for eight different student test items.

To generate a Student Ability Knowledge score for a teacher, we calculate the absolute difference between the teacher estimate and actual percentage correct within the teacher's classroom. We then estimated the following multilevel model, where items are nested within teachers and weighted by the number of students in each classroom:

$$p_{j,k} = \beta_0 + \mu_k + \epsilon_{j,k} \quad (3)$$

The outcome of interest, $p_{j,k}$, is the absolute difference between teacher k's estimated percentage of his or her students answering item j on the low-stakes assessment correctly and the actual percentage of students answering item j correctly, populated for each student taught by teacher k answering item j. The model parameter μ_k is teacher k's shrunken Student Ability Knowledge score, adjusted for differences in the reliability of estimates from teacher to teacher due to differences in the total number of students answering each item j.

To generate a Student Misconception Knowledge score for a teacher, we estimate the following multilevel model, where items are nested within teachers and weighted by the number of students answering each item:

$$\log\left[\frac{\pi_{j,k}}{1-\pi_{j,k}}\right] = \beta_0 + \mu_k \quad (4)$$

The outcome of interest, $\log\left[\frac{\pi_{j,k}}{1-\pi_{j,k}}\right]$, is the log-odds of teacher *k* correctly predicting the most common incorrect answer among his or her students answering item *j* on the low-stakes assessment, populated for each student taught by teacher *k* answering item *j*. The model parameter μ_k is teacher *k*'s shrunken Student Misconception Knowledge score, adjusted for differences in the reliability of the estimates from teacher to teacher due to differences in the total number of students answering each item *j*, and the number of Student Misconception Knowledge questions answered by the teacher *k*.

Other Teacher Measures

From the fall questionnaires distributed to teachers, we also recovered a set of measures investigating a range of teacher beliefs and behaviors, including teacher self-efficacy (Self-Efficacy), teacher use of formative assessment (Formative Assessment), teacher time and effort in preparation for instruction (Effort), teacher use of test preparation practices (Test Prep—Activities), and perception that testing has led to undesirable changes in their mathematics instruction (Test Prep—Instruction). These variables were included on the teacher questionnaire because they proved difficult or impossible to observe from videotapes of lessons.

To generate teacher scores on these metrics, we estimated the following model:

$$T_{j,k} = \beta_0 + \mu_k + \nu_k + \epsilon_{j,k} \quad (5)$$

The outcome of interest, $T_{j,k}$, represents teacher k's response to item j related to the measure in consideration. The model parameter μ_k is teacher k's shrunken measure score, adjusted for differences in the reliability of the estimates from teacher to teacher due to differences in the total number of items on the measure responded to.

From the spring questionnaires distributed to teachers, we recovered measures capturing the mathematics content covered by teachers in the classroom for either numbers and operations (Content Coverage—Numbers and Operations) or elementary algebra (Content Coverage—Algebra). This measure was presented as a list of grade-level-specific topics, with instructions to the teacher to indicate whether the topic was covered in class. These lists were developed from a survey of content typically taught at the two grade levels, and the intent was to assess the alignment between teachers' content coverage and topics likely to be on the state and actually on the alternative assessment.

To generate teacher scores on these metrics, we estimated the following model:

$$\log\left[\frac{\pi_{j,k}}{1-\pi_{j,k}}\right] = \beta_0 + \mu_k + \nu_j \quad (6)$$

The outcome of interest, $\log\left[\frac{\pi_{j,k}}{1-\pi_{j,k}}\right]$, is the log-odds of teacher k covering mathematical subtopic j for either numbers and operations or algebra. The model parameter μ_k is teacher k's shrunken content coverage score, adjusted for differences in the reliability of the estimates from teacher to teacher due to differences in number of subtopics for which teachers provided a response. Higher scores indicate more coverage of the topics on the project-developed alternative assessment.

Student Data

For each student in participating study classrooms, we collected the following from the school years 2010–11, and 2011–12: (a) student-teacher links, (b) student demographic information, (c) student performance on state standardized mathematics and reading exams, and, (d) student performance on a study-designed-and-administered alternative low-stakes mathematics exam aligned with study-developed teacher measures. We also collected (a), (b), and (c) for all nonparticipating students in fourth and fifth grade in the study districts. The state tests used in our analysis had a range of reliability estimates from 0.90 to 0.93.

The low-stakes alternative fourth- and fifth-grade mathematics exams were jointly developed by [test author institutions omitted for blind review] between 2009 and 2012. These were designed to measure student learning gains resulting from teacher professional development and to be sensitive to variation in teachers' mathematical knowledge for teaching and instructional quality. The tests focused on three mathematical domains—numbers and operations, algebra, and geometry and measurement—in order to align with the fourth- and fifth-grade Common Core mathematics standards and the MKT items. To further align the assessments with the MKT and MQI, items were focused on the meaning of the mathematics (e.g., matching concrete representations to computations) and on students' knowledge of alternative procedures. In addition to standard multiple choice items, gridded responses and "nested-sets" of items were used to provide a better assessment of student understanding. Six different forms were used, and the reliability estimates of each form ranged from 0.82 to 0.89.[4]

Sample. We include students in our analysis sample who met the following restrictions:

(1) The student was reliably linked to a single primary mathematics teacher in the given school year.

(2) The student had current- and prior-year state standardized mathematics test scores and beginning- and end-of-year alternative mathematics test scores.[5]

(3) The student had demographic information.

(4) The student did not skip or repeat a tested mathematics grade.

(5) The student was in a classroom with (a) at least five students, (b) less than 50% special education students, and (c) less than 50% students missing either mathematics achievement score.

We include teachers in our analysis sample when they taught students meeting the restrictions listed above and had at least one score from any collected measure. In cases where a teacher was missing some but not all 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, with five or fewer teachers ultimately having imputed scores for any given measure. Our final sample consisted of 283 teachers and 7,843 students.

Insert Table 2.

In Table 2, we provide summary statistics of these 283 teachers and their students in our analysis sample and those outside of it. The particular sample of teachers and students used in our analyses does not differ substantially from the rest of the teachers and students in these four districts, based on characteristics such as teacher value-added scores and the student demographics. Using a difference in mean *t* test, we find that only the percentage of students in a classroom who are African American differs between our analysis sample and the rest of the

classrooms in these districts. Thus, we find no evidence that sample selection limits the ability to generalize the findings of our analyses to the larger sample of teachers.

Analysis Strategy

To answer our primary research question, we investigate how much teacher-level variance in student outcomes is explained by different features of teachers and teaching, both in isolation and in conjunction with one another. We use two different student outcomes: scores on state standardized math tests and scores on the alternative, low-stakes math test. In modeling the relationship between the tests and predictors, we include only direct linear effects; though theory does suggest the possibility of both, modeling nonlinearities and interaction terms would risk overfitting the model to potential idiosyncrasies in our sample, especially given the large number of predictors.

We estimate the amount of teacher-level variation in these student outcomes by fitting the following multilevel equation for each outcome:

$$a_{i,k,t} = A_{i,t-1}\alpha + S_{i,t}\beta + P_{k,t}\delta + \eta + \nu_{i,k,t}, \quad (7)$$

$$\text{where } \nu_{i,k,t} = \mu_k + \theta_{k,t} + \varepsilon_{i,k,t}$$

and $a_{i,k,t}$ is the outcome score for student i taught by teacher k during school year t. In addition to grade-by-year and district fixed effects,[6] η, we include the following control variables: $A_{i,t-1}$, a cubic polynomial of student i's prior 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 $P_{k,t}$, a vector of average characteristics of student i's peers in the same class and school, including average test scores and averages of $S_{i,t}$. For a full description of student-level variables included in these models, please see Tables A1 and A2 in the Appendix.

In Equation 7 we estimate teacher random effects, μ_k. The variation in μ_k represents the amount of variation in student outcomes explained by differences between teachers in our sample. To determine how much of this variation in student outcomes is explained by different observable features of teachers or teaching, we estimate a taxonomy of multilevel models similar to Equation 7:

$$a_{i,k,t} = A_{i,t-1}\alpha + S_{i,t}\beta + P_{k,t}\delta + \eta + \nu_{i,k,t}, \quad (8)$$

$$\text{where } \nu_{i,k,t} = \varphi_k + \theta_{k,t} + \varepsilon_{i,k,t}$$

$$\text{and } \varphi_k = T_k\gamma + \tau_k.$$

Equation 8 uses the same specification as in Equation 7, except that in Equation 8, we include various teacher level variables, depending on the specific specification. In Equation 8, φ_k is a function of a vector of m teacher-level variables, T_k, representing features of teacher and teaching, and τ_k, the teacher random effects after controlling for these features. The variance in τ_k represents the teacher-level variance in student outcomes due to differences between teachers after taking these features into account.

By comparing an adjusted ratio of teacher-level variance components for parameters μ_k and τ_k from Equation 7 and Equation 8, we generate 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_k. We define the "adjusted teacher-level R^2" statistic as follows:

$$R^2 = 1 - \frac{n-1}{n-1-m} \times \frac{\mathrm{Var}(\tau_k)}{\mathrm{Var}(\mu_k)}, \quad (9)$$

where n represents the number of teachers in the sample, m represents the number of teacher-level variables used in Equation 8, $\mathrm{Var}(\tau_k)$ represents the teacher-level variation in student outcomes in Equation 8 after controlling for the vector of teacher-level variables T_k, and $\mathrm{Var}(\mu_k)$ represents the teacher-level variation in student outcomes in Equation 7 without this vector. The

ratio ($\frac{n-1}{n-1-m}$) adjusts for the mechanical reduction in $\mathrm{Var}(\tau_k)$ that tends to occur when more teacher-level variables are added to the model. We use this statistic to estimate how much teacher-level variation in student outcomes is due to different features of teachers or teaching. Along with our estimate of adjusted teacher-level R^2, we also provide a bootstrapped 95% confidence interval for this estimate.[7]

Results

First, we present a set of bivariate analyses, providing an overview of how the characteristics of teachers and teaching are related to each other. For a selection of univariate descriptives of these indicators, please see Table A3, and Figures A1, A2, and A3 in the Appendix.

Insert Table 3.

Insert Table 4.

In Table 3, we present correlation coefficients between our measures of teachers and teaching. In Table 4, we present correlation coefficients between our measures of teacher background characteristics and with measures of teachers and teaching. In Table 3, the correlations between these variables suggest that they measure distinct traits, rather than an underlying dimension measuring "good teachers." In only a few cases are there correlations above 0.30 or below -0.30, and among the higher correlations, several are observed across different dimensions captured by the same instrument (e.g., CLASS, MQI, teacher surveys). Across-instrument correlations above 0.30 occurred between the MQI and Content and Teaching knowledge measures, a relationship also observed in prior research (Hill, Blunk, et al., 2008). For the most part, however, the data appear to suggest that the variables chosen measure distinct aspects of teacher and teaching quality.

Insert Table 5.

In Table 5 we present the reduction in teacher-level variance (adjusted teacher-level R^2) as we add teacher-level variables to our base model explaining students' state test performance (Equation 7). In the base model (Equation 7), the standard deviation of the teacher effect estimates on state test outcomes is 0.16, comparable to other estimates from the teacher effects literature. When adding teacher race and gender, we explain about 1% of the teacher-level variance in student outcomes. Teachers' preparation routes and teaching experience—indicators conventionally explored by economists and typically found in administrative datasets—explain about 8% of this variance, and instruction as measured by the classroom observation instruments explains about 7%. Teachers' knowledge of content, of teaching, and of their students also explains about 7% of the teacher-level variability while another 8% is explained by math-specific practices reported by teachers on the survey (e.g., formative assessment, content coverage). Together, teacher knowledge, instruction, and survey-reported practices explain 20% of the variability in outcome, roughly twice the proportion explained by preparation routes and experience. In total, 28% of the teacher effect on state test performance is explained in this model.[8]

Insert Table 6.

Table 6 reports the reduction in teacher-level variance for the alternative assessment. Here again, teacher demographics explain a relatively lower proportion of the variance (2%) while teacher preparation routes and teaching experience explain a surprising 13% of the variance. Instruction as measured on the classroom observation instruments explains about 7% of the variance while the teacher knowledge measures explain 21%. Survey-reported mathematics practices explain another 6% of the variability. Together, teacher knowledge,

instruction, and survey-reported practices explain 30% of the variability in outcomes, and in total, 40% of the teacher-level variability is explained.

Predictor variables explain 12 more percentage points of the teacher-level variance on the alternative assessment than on the state assessments. One reason may be that the alternative 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 such as special needs or limited English proficient, or preparing students for specific high-stakes test questions; see Jacob, 2005) that may subsequently contribute to score distortion (Koretz, McCaffrey, & Hamilton, 2001). Further, the alternative student assessment was developed to match the teacher knowledge measures; it is not surprising to see that the knowledge measures explained far more teacher-level variance in alternative test outcomes as opposed to state test outcomes. However, this finding does suggest the importance of alignment between teacher-level predictors and outcome variables in efforts to explain teacher effects on student achievement.

In general, both the distribution of variance explained across the five groups of variables and the small coefficients on each variable suggest that there is no one single variable or group of variables that explains a large portion of teacher effects. This, in combination with the correlation matrices of Tables 3 and 4, suggests that the phenomena underlying teacher effects may be multidimensional. Teaching may also be contingent, in the sense that the effectiveness of a specific dimensions depends upon reaching a satisfactory threshold on another; because of concerns about model over-fitting, we leave this investigation for future work.

Across Tables 5 and 6, several variables stood out as related to student outcomes. Among teacher preparation and pathway variables, math content courses consistently related to

student outcomes; possession of a bachelor's degree in education also predicts state test scores as well. Among the elements of instruction scored on observational instruments, the two that captured classroom management and time on task—MQI's Classroom Work Connected to Mathematics and CLASS' Classroom Organization—appeared more related to both the state and alternative assessments than factors pertaining to inquiry-based instruction from model to model. Teachers' Content and Teaching knowledge and the accuracy of teacher predictions of students' performance on specific alternative test items both predicted outcomes, though inconsistently across models for the state tests. Finally, the teacher-reported coverage of elementary algebra concepts (e.g., meaning of the equals sign, using a symbol to stand for an unknown) predicted outcomes on the state test.

Insert Table 7.

Finally, in Table 7, we present parameter estimates for individual teacher-level variables from 31 separate regressions, each with only one teacher-level variable included. By doing this, we can shed light on past studies of teacher and teaching, which tend to examine the effects of these variables separately. Similar to past studies, we find that some of the variables are statistically significantly related to student test score gains: At the 10% significant level, six of the 31 variables tested were significantly related to the student outcomes on both tests. For either outcome variable, no more than one third of the predictor variables are statistically significantly related to the student outcome. In addition, we find that any individual variable explains, at most, 6% of the teacher-level variance when using the high-stakes state exam and 14% of the teacher-level variance when using the alternative project-developed exam as the outcome. Thus, while we find some variables that are individually statistically significantly related to teacher effectiveness, these variables only independently explain a small amount of

teacher-level variance, especially for the state outcome variable. In the case of the project-developed exam, we find that the variables that the test was constructed to measure are, indeed, most related (e.g., Content and Teaching knowledge). This suggests that analysts attempting to explain teacher effects should include predictor variables from a range of research traditions and consider the alignment of such measures to the outcome variable in order to more fully understand differences in teacher effectiveness.

Conclusion

Our analysis suggests that we can explain a modest to moderate amount of teacher-level variance in student test outcomes on two different mathematics assessments using a range of predictor variables. The proportion of explained variability was less for the state test (28%) than for a low-stakes, well-aligned assessment designed specifically for this project (40%). Though these estimates are greater than those reported in Palardy and Rumberger (2008) and Boonen et al. (2013), much teacher-level variability clearly remains unexplained by these models. One possibility is that some of the unexplained variability owes to measurement error or random disturbances in the data (e.g., the barking dog). Measurement error in our predictor variables no doubt contributes to the lower fraction of explained variance as well.[9] Another possibility might be that there are as-yet undiscovered markers of teaching and teacher quality, a question we take up below.

When considering the alternative test as the outcome, teacher knowledge held substantially more explanatory power than the other categories of variables. When considering state tests as the outcome variable, however, teachers' preparation and background held the most explanatory power. Though this result stands in contrast to earlier production function research that found teachers' preparations pathways explained little of the variability in teacher-level

outcomes, other categories of variables in our analyses explained comparable amounts of teacher-level variation in student state test outcomes as well. For both outcomes, the predictor categories of knowledge assessments, teacher self-reports of practice, and background independently explained a similar or larger proportion of variation than the indicators derived from observations of classroom. This finding might suggest to practitioners that other more cost-efficient data sources may provide the same or more information about differences in teacher effects.

Interesting too was the small coefficient on most variables, and their relative independence from one another in explaining teacher effects. This fits with views that see instruction as complex, resulting from interactions between teachers, students, students' peers, and material (Cohen, 2011); teacher effects on student achievement are not likely the simple average of teachers' position on the metrics we have used, but instead some interaction of dimensions such as classroom management, teacher content expertise, knowledge of students, and alignment between the curriculum and the test. Future work with this dataset will test for such interactions.

Finally, we comment on 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 that instruction. 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. These features are harder to gauge from observations or 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. This, as well as the low amount of variability explained by the observational metrics generally, suggests that the search for predictors of teacher effects may proceed more fruitfully along other pathways.

Notes

[1] We used a third available year of survey items capturing teacher general and teaching-specific content knowledge in our analyses.

[2] We chose to observe three lessons per year because of results from a prior decision study (citation omitted for blind review) and because three is likely similar to the number of observations enacted in many teacher evaluation systems.

[3] Teachers can have a different number of lessons scored for a variety of reasons. One of the most common reasons is that some teachers are in the study for two years while others are only in the study for one year. In these cases, the teachers in the study for only one year have fewer scored lessons, which produces a less reliable estimate than the teachers with scored lessons in two years. We have applied a Bayesian shrinkage procedure similar to this whenever we note that we estimated a shrunken effect.

[4] For more information, please see the following technical report: [omitted for blind review].

[5] For one district in one school year, beginning-of-year alternative mathematics test scores were unavailable. In our analyses, we used prior-year state mathematics test scores as a proxy.

[6] Inclusion of district fixed effects in the multilevel model predicting student achievement on the cross-district-distributed, low-stakes, alternative assessment may mask systematic variation in student achievement across districts attributable to teachers. As such, we conducted a sensitivity analysis investigating this possibility by excluding district fixed effects from our models and found no significant differences in our results.

[7] We calculate the bootstrapped 95% confidence interval by fitting 100 iterations of each model, each with a different random sample of $n_{t,E1}$ teachers. These confidence intervals are bias-corrected and account for potential overfitting in the original model.

[8] In a similar analysis on a larger student sample in which students were included regardless of whether we collected their alternative test achievement data, we recovered similar values for the adjusted teacher-level R^2 statistics.

[9] Because many of our predictor variables are no doubt measured with error, they explain less variance than an analogous error-free measure. We do not, however, attempt to correct for this, since the error is a property of the variables. In other words, we are interested in estimating the explained teacher-level variance using the data that we have, not using some underlying error-free measure.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*, 95–135.

Armor, D., Conroy-Oseguera, P., Cox, M., King, N., McDonnell, L., Pascal, A., Pauly, E., & Zellman, G. (1976). *Analysis of the School Preferred Reading Programs in Selected Los Angeles Minority Schools, REPORT NO. R*-2007-*LAUSD*. Santa Monica, CA: Rand Corporation (ERIC Document Reproduction Service No. 130 243).

Ball, D. L., & Forzani, F. M. (2009). The work of teaching and the challenge for teacher education. *Journal of Teacher Education*, *60*, 497–511.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusman, U., Krauss, S. Neubrand, M. & Tsai, Y. M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*, 133–180.

Begle, E. G. (1972). *Teacher knowledge and student achievement in algebra* (Vol. 9). Palo Alto, CA: School Mathematics Study Group.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*, 62–87.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice*, *5*, 7–74.

Boonen, T., Van Damme, J., & Onghena, P. (2013). Teacher effects on student achievement in first grade: Which aspects matter most? *School Effectiveness and School Improvement*, *25*, 1–27.

Briggs, D. C., Araceli Ruiz-Primo, M., Furtak, E., Shepard, L., & Yin, Y. (2012). Meta-analytic methodology and inferences about the efficacy of formative assessment. *Educational Measurement: Issues and Practice, 31,* 13-17.

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.), Macmillan, New York (1986), pp. 328–375.

Carlisle, J. F., Correnti, R., Phelps, G., & Zeng, J. (2009). Exploration of the contribution of teachers' knowledge about reading to their students' improvement in reading. *Reading and Writing, 22*, 457–486.

Carlisle, J. F., Kelcey, B., Rowan, B., & Phelps, G. (2011). Teachers' knowledge about early reading: Effects on students' gains in reading achievement. *Journal of Research on Educational Effectiveness, 4*, 289–321.

Carpenter, T. P., Fennema, E., Peterson, P. L., & Carey, D. A. (1988). Teachers' pedagogical content knowledge of students' problem solving in elementary arithmetic. *Journal for Research in Mathematics Education, 19,* 385–401.

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

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

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, Instruction, and Research. *Educational Evaluation and Policy Analysis, 25*, 119–142.

Cooley, W. W., & Leinhardt, G. (1980). The instructional dimensions study. *Educational Evaluation and Policy Analysis, 2*, 7–25.

Depaepe, F., Verschaffel, L., & Kelchtermans, G. (2013). Pedagogical content knowledge: A systematic review of the way in which the concept has pervaded mathematics educational research. *Teaching and Teacher Education, 34,* 12–25.

Gamoran, A., Porter, A. C., Smithson, J., & White, P. A. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis, 19,* 325–338.

Goldhaber, D. D., Brewer, D. J., & Anderson, D. J. (1999). A three-way error components analysis of educational productivity. *Education Economics, 7,* 199–208.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research, 66,* 361–396.

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2010). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. NBER Working Paper No. 16015.

Harbison, R. W., & Hanushek, E. A. (1992). *Educational performance of the poor: Lessons from rural northeast Brazil.* Oxford University Press.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature, 24,* 1141–1177.

Helmke, A., & Schrader, F. W. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. *Teaching and Teacher Education, 3,* 91–98.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39,* 372–400.

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

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*, 371–406.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, *59*, 297–313.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*, 761–796.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*, 101–136.

Kane, T. J., & Cantrell, S. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study*. Seattle, WA: Bill and Melinda Gates Foundation.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D.O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle, WA: Bill and Melinda Gates Foundation.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, *27*, 615–631.

Kane, T. J., & Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. NBER Working Paper No. 14607.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill and Melinda Gates Foundation.

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal, 49*, 568–589.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30,* 28–37.

Koretz, D. (2008). A measured approach. *American Educator, 32*, 18–39.

Koretz, D. M., McCaffrey, D. F., & Hamilton, L. S. (2001). *Toward a framework for validating gains under high-stakes conditions*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education and Information Studies, University of California, Los Angeles.

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

Lavy, V. (2004). Performance pay and teachers' effort, productivity and grading ethics. NBER Working Paper No. 10622.

Metzler, J., & Woessmann, L. (2012). The impact of teacher subject knowledge on student achievement: Evidence from within-teacher within-student variation. *Journal of Development Economics, 99*, 486–496.

Muralidharan K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy, 119*, 39–77.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*, 237–257.

Palardy, G. J., & Rumberger, R. W. (2008). Teacher effectiveness in first grade: The importance of background qualifications, attitudes, and instructional practices for student learning. *Educational Evaluation and Policy Analysis, 30*, 111–140.

Pianta, R. C., LaParo, K. M., & Hamre, B. K. (2007). *Classroom Assessment Scoring System (CLASS) manual*. Baltimore, MD: Brookes Publishing.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*, 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from Panel Data. *American Economic Review, 94*, 247–252.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy, 6*, 43–74.

Rose, J. S., & Medway, F. J. (1981). Measurement of teachers' beliefs in their control over student outcome. *The Journal of Educational Research, 74*, 185–190.

Ross, J. A. (1992). Teacher efficacy and the effects of coaching on student achievement. *Canadian Journal of Education/Revue canadienne de l'education, 17*, 51–65.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*, 537–571.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*, 473–489.

Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle school physical science classrooms. *American Educational Research Journal, 50*, 1020–1049.

Shechtman, N., Roschelle, J., Haertel, G., & Knudsen, J. (2010). Investigating links from teacher knowledge, to classroom practice, to student learning in the instructional system of the middle-school mathematics classroom. *Cognition and Instruction, 28*, 317–359.

Shulman, L. 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.

Stein, M. K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation, 2*, 50–80.

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education, 62*, 339–355.

Tschannen-Moran, M., Hoy, A. W., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research, 68*, 202–248.

Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2010). Using student performance data to identify effective classroom practices. *American Economic Review, 100*, 256–260.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*, 89–122.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education, 11*, 49–65.

Tables and Figures

Table 1. *Summary of Teacher-Level Predictors*

Demographic measures
- *Male* indicates whether the teacher is male.
- *White* indicates whether the teacher is white.
- *Other Race* indicates whether the teacher is a race other than Black or White.
Background measures derived from teacher responses to surveys
- *Master's Degree* indicates whether the teacher obtained any master's degree
- *# Math Courses* indicates the number of undergraduate or graduate math courses taken by the teacher
- *Math Content* indicates the number of undergraduate or graduate math content courses for teachers taken by the teacher
- *Math Major* indicates whether the teacher obtained an undergraduate major or minor or graduate degree in mathematics
- *Ed Bachelors* indicates whether the teacher obtained a bachelor's degree in education
- *1-3 Yrs.; 4-10 Yrs.; 10+ Yrs Experience* Indicates the teacher's total number of years teaching mathematics (including current year)
- *Trad.; Alt.; No Certification* indicates the certification pathway taken by the teacher prior to his or her first teaching position (traditional, alternative, none)
- *Elementary Math Certification* indicates whether the teacher self-reported a specific certification for teaching elementary mathematics
Instruction measures derived from recorded lessons of math instruction
Mathematical Quality of Instruction (MQI) measures (Hill, Blunk, et al., 2008)
- *Classroom Work Connected to Math* captures the connection of classroom work during instruction to mathematics (ICC = .36)
- *Ambitious Instruction* captures the depth of the mathematics offered to students by the teacher, in addition to the extent to which students participate in mathematical meaning-making and reasoning (ICC = .69)
- *Errors* captures the mathematical mistakes or imprecisions in the teacher's instruction (ICC = .52)
Classroom Assessment Scoring System (CLASS) measures (Pianta et al., 2007)
- *Classroom Organization* captures the negative climate in the classroom and the productivity and behavior management demonstrated in the teacher's instruction (ICC = .65)
- *Support* captures both the emotional and instructional support provided by the teacher to the students during instruction (ICC = .51)
Teacher knowledge measures derived from teacher responses to surveys
- *Knowledge—Teaching and Content* captures both the teacher's mathematical knowledge for teaching and the teacher's general mathematics knowledge (marginal test reliability = .85) (Hill, Ball, & Schilling, 2008)
- *Knowledge—Student Ability* captures the teacher's knowledge of his or her students' mathematical abilities (ICC = .89) (Hoge & Coladarci, 1989)
- *Knowledge—Student Misconceptions* captures the teacher's knowledge of his or her students' mathematical misconceptions (ICC = .58) (Sadler et al., 2013)
Other teacher measures derived from teacher responses to surveys
- *Self-Efficacy* captures teachers' efficacy in providing strong instruction to students (ICC = .73) (Tschannen-Moran et al., 1998)
- *Formative Assessment* captures teachers' use of formative assessment (ICC = .54) (Wiliam et al., 2004)
- *Effort* captures additional time spent by teachers preparing for class, organizing materials, grading homework, etc. (ICC = .82) (Lavy, 2004)
- *Test Prep—Activities* captures teacher engagement in test preparation activities in the classroom (ICC = .81) (Koretz, 2008)
- *Test Prep—Instruction* captures changes to teacher instruction due to preparation for standardized testing (ICC = .87) (Koretz, 2008)
- *Content—Numbers and Operations* captures the range of mathematical subtopics taught by teachers concerning numbers and operations (ICC = .85)
- *Content—Algebra* captures the range of mathematical subtopics taught by teachers concerning algebra (ICC = .81)

Note: The ICCs reported are for teacher scores on each measure and not for individual observations.

Table 2. *Summary Statistics for in-Sample and out-of-Sample Teachers*

	In sample (*n*=283)	Out of sample (*n*=1,784)	Overall (*N*=2,067)
Characteristics of teachers			
Teacher's value-added (state test)	0.00	0.00	0.00
	(0.16)	(0.16)	(0.16)
Characteristics of teachers' students			
Class size	20.07	19.28	19.39
	(6.44)	(7.51)	(7.37)
Prior-year average state math score	0.00	0.01	0.00
	(0.50)	(0.55)	(0.55)
Male (%)	49.81	49.05	49.15
	(8.88)	(9.40)	(9.33)
Black (%)	43.61	37.73	38.53
	(27.11)	(28.27)	(28.18)
Asian (%)	6.93	7.90	7.77
	(11.69)	(11.65)	(11.66)
Hispanic (%)	25.87	28.61	28.23
	(23.60)	(24.82)	(24.67)
White (%)	19.53	21.75	21.45
	(20.79)	(22.82)	(22.56)
Other race (%)	4.05	3.82	3.85
	(4.04)	(4.50)	(4.44)
Subsidized-lunch eligible (%)	68.68	66.11	66.46
	(24.12)	(27.64)	(27.19)
Special education status (%)	12.01	10.88	11.03
	(9.22)	(9.55)	(9.51)
Limited English proficiency (%)	23.34	22.45	22.58
	(24.18)	(23.55)	(23.63)

Note: Standard deviations are reported in parentheses.

Table 3. *Teacher-Level Correlation Coefficients Among Measures of Teachers and Teaching*

	MQI—Errors	MQI—Ambitious Instruction	MQI—Classroom Work Connected to Math	CLASS—Class Organization	CLASS—Support	Knowledge—Student Ability	Knowledge—Student Misconceptions	Knowledge—Teaching and Content	Effort	Test Prep—Instruction	Test Prep—Activities	Self-Efficacy	Formative Assessment	Content Coverage—Numbers and Operations	Content Coverage—Algebra
MQI—Errors	1.00														
MQI—Ambitious Instruction	-.31	1.00													
MQI—Classroom Work Connected to Math	.08	.18	1.00												
CLASS—Class Organization	.05	.19	.16	1.00											
CLASS—Support	-.07	.29	.12	.42	1.00										
Knowledge—Student Ability	-.17	.20	.02	-.02	.03	1.00									
Knowledge—Student Misconceptions	-.12	.09	.00	-.06	-.03	.14	1.00								
Knowledge—Teaching and Content	-.41	.34	.07	.02	.08	.25	.14	1.00							
Effort	.20	-.05	.09	.13	.07	-.11	-.08	-.14	1.00						
Test Prep—Instruction	-.03	-.14	-.04	-.04	-.08	.16	.06	-.05	-.16	1.00					
Test Prep—Activities	.25	-.17	.03	.20	.07	-.25	-.10	-.25	.25	.08	1.00				
Self-Efficacy	.02	.05	.00	.05	.09	-.10	-.07	.02	.07	-.26	.10	1.00			
Formative Assessment	.16	-.07	-.07	.10	.08	-.09	-.07	-.11	.33	-.10	.31	.25	1.00		
Content Coverage—Numbers and Operations	.05	-.01	.04	.01	.01	.05	.17	.05	.21	-.06	.10	.13	.21	1.00	
Content Coverage—Algebra	-.15	.03	-.05	.06	.10	.08	.08	.18	.16	-.05	.02	.15	.20	.49	1.00

Note: Sample includes 283 teachers.

Table 4. *Teacher-Level Correlations of Teacher Background Characteristics and Measures of Teachers and Teaching*

		Maste's Degree	# Math Courses	Math Content	Math Major	Ed. Bachelor's	1–3 Yrs Experience	4–10 Yrs Experience	10+ Yrs Experience	Male	White	Black	Other Race	Trad. Certification	Alt. Certification	No Certification
Teacher Background Characteristics	Master's Degree	1.00														
	# Math Courses	-.01	1.00													
	Math Content	.00	.48	1.00												
	Math Major	.03	.19	.06	1.00											
	Ed. Bachelor's	-.23	-.01	.09	.05	1.00										
	1–3 Yrs Experience	-.27	-.14	-.21	-.06	-.05	1.00									
	4–10 Yrs Experience	.16	-.05	-.09	.09	-.02	-.37	1.00								
	10+ Yrs Experience	.07	.12	.21	-.05	.08	-.34	-.72	1.00							
	Male	.01	.04	-.05	.02	-.17	-.02	-.03	.04	1.00						
	White	-.01	-.07	.09	.06	.19	-.02	.09	-.04	.00	1.00					
	Black	.03	.08	-.09	-.06	-.12	-.05	-.03	.05	-.04	-.71	1.00				
	Other Race	.03	-.05	-.05	-.03	-.06	.07	-.04	-.02	.12	-.43	-.17	1.00			
	Trad. Certification	.06	-.11	.08	-.17	.36	-.07	-.01	.12	-.05	.35	-.32	-.01	1.00		
	Alt. Certification	-.01	.04	-.20	.05	-.27	.11	.00	-.10	.06	-.15	.18	.00	-.61	1.00	
	No Certification	-.01	.05	.01	.16	-.17	.03	.00	-.04	.02	-.24	.24	.00	-.61	-.08	1.00
Teaching Indficators	MQI—Errors	.00	-.05	.00	-.07	-.06	-.03	-.05	.04	-.01	-.22	.17	.12	.01	-.09	.01
	MQI—Ambitious Instruction	-.03	.02	.07	.10	.06	-.07	.09	-.03	.04	.17	-.13	-.07	.05	-.03	.01
	MQI—Classroom Work Connected to Math	-.02	.02	-.02	.04	-.09	.00	.06	-.07	.06	-.09	.09	-.02	-.03	.02	.04
	CLASS—Class Organization	-.10	.08	.11	.08	.13	-.08	.06	.00	-.02	-.11	.08	.12	.09	-.06	-.04
	CLASS—Support	-.07	.09	.06	.02	.07	.01	.02	-.04	-.09	-.05	.05	-.06	.02	.01	-.02
	Knowledge—Student Ability	.10	-.07	-.04	.05	-.01	-.01	.04	.00	-.03	.19	-.16	-.05	.14	-.05	-.08
	Knowledge—Student Misconceptions	.12	.11	-.04	-.03	-.09	-.05	.12	-.06	-.06	-.03	.11	-.11	-.02	.05	.04
	Knowledge—Teaching and Content	.03	.04	.00	.02	-.05	.05	.07	-.10	.12	.29	-.24	-.10	.09	.07	-.11
	Effort	-.07	.22	.13	-.01	.00	.05	-.15	.11	-.05	-.27	.21	.10	-.09	.05	.04
	Test Prep—Instruction	.02	-.08	-.08	-.10	.02	.05	-.06	.06	-.03	-.02	.03	-.03	.07	-.05	-.07
	Test Prep—Activities	-.03	.23	.15	.13	.03	-.07	.01	.01	-.18	-.22	.21	-.02	-.04	.00	.01
	Self-Efficacy	-.08	.10	.03	.07	-.02	-.11	.08	-.03	-.05	.02	.07	-.11	-.14	.11	.03
	Formative Assessment	-.07	.24	.19	.04	.09	-.05	.02	.01	-.05	-.10	.16	-.02	-.04	.04	.00
	Content Coverage—Numbers and Operations	.13	.09	.02	.03	-.07	-.01	.06	-.03	.08	-.12	.12	.09	-.10	.17	.03
	Content Coverage—Algebra	.09	.08	.06	.03	.01	.04	-.12	.13	.14	.00	-.01	.02	-.05	.13	-.02

Note: Sample includes 283 teachers.

Table 5. *Estimates of Teacher-Level Parameters and Adjusted Teacher-level R^2 from Hierarchical Models of State-Administered Assessments*

	Model 0: Baseline	Model 1: Demographics	Model 2: Background	Model 3: Observations	Model 4: Knowledge	Model 5: Other Characteristics	Predictors from Models 3–5	All Predictors
Male		-.012						-.004
White		-.029						-.068
Other Race		.090						.074
Master's Degree			.031					.022
# Math Courses			-.001					-.008
Math Content			.046*					.039*
Math Major			.030					.013
Ed. Bachelor's			.077*					.079**
4–10 Yrs. Experience			.050					.065
10+ Yrs. Experience			-.006					.003
Elem. Math Certification			-.041					-.055
Alt. Certification			.050					.037
No Certification			.002					.017
MQI—Errors				-.022			-.017	-.013
MQI—Ambitious Instruction				.006			.001	-.004
MQI—Classroom Work Connected to Math				.023			.025	.030*
CLASS—Class Organization				.033~			.028	.010
CLASS—Support				.007			.001	.005
Knowledge—Student Ability					.031*		.037*	.038*
Knowledge—Student Misconceptions					-.017		-.012	-.013
Knowledge—Teaching and Content					.029*		.018	.027~
Effort						.020	.020	.017
Test Prep—Instruction						-.012	-.016	-.011
Test Prep—Activities						-.001	.007	.004
Self-Efficacy						-.016	-.014	-.009
Formative Assessment						.019	.023	.018
Content Coverage—Numbers and Operations						-.006	.004	-.010
Content Coverage—Algebra						.042*	.027~	.035*
Intercept	0.943~	0.993~	0.624	0.892~	0.959~	0.963~	0.972~	0.847
Teacher Variance	0.030***	0.029***	0.026***	0.027***	0.028***	0.027***	0.023***	0.019***
Classroom Variance	0.014***	0.014***	0.015***	0.015***	0.015***	0.015***	0.015***	0.015***
Residual Variance	0.245***	0.245***	0.245***	0.245***	0.245***	0.245***	0.245***	0.245***
Adjusted Teacher-level R^2		.014	.080	.070	.067	.082	.196	.281
R^2 95% CI Lower Bound		-.024	-.017	-.018	.013	.016	.031	.177
R^2 95% CI Upper Bound		.070	.223	.178	.171	.198	.336	.361

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 includes student-, class-, and cohort-level controls for test scores and demographic characteristics (these parameter estimates are shown in Table A1). 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 (see the analysis section of the paper for details). A bootstrapped 95% confidence interval for this for the adjusted teacher-level R^2 is estimated by fitting 100 iterations of each model, each with a different random sample of teachers. Confidence intervals are bias-corrected and account for potential over-fitting in the original model.
*$p < .10$; **$p < .05$; ***$p < .01$.

Table 6. *Estimates of Teacher-Level Parameters and Adjusted Teacher-level R^2 from Hierarchical Models of the Project-Administered Alternative Assessment*

	Model 0: Baseline	Model 1: Demographics	Model 2: Background	Model 3: Observations	Model 4: Knowledge	Model 5: Other Characteristics	Predictors from Models 3-5	All Predictors
Male		.037						.022
White		.015						-.046
Other Race		.133*						.094~
Master's Degree			.020					.012
# Math Courses			.006					.012
Math Content			.044*					.037*
Math Major			-.004					-.024
Ed. Bachelor's			.039					.048~
4–10 Yrs. Experience			-.038					-.046
10+ Yrs. Experience			-.000					.005
Elem. Math Certification			-.033					-.019
Alt. Certification			.075					.069
No Certification			-.026					.002
MQI—Errors				-.017			-.002	-.000
MQI—Ambitious Instruction				.009			-.000	-.002
MQI—Classroom Work Connected to Math				.015			.012	.016
CLASS—Class Organization				.028~			.030~	.019
CLASS—Support				-.012			-.016	-.011
Knowledge—Student Ability					.025~		.032*	.036*
Knowledge—Student Misconceptions					-.021		-.020	-.021
Knowledge—Teaching and Content					.039**		.032*	.036*
Effort						-.003	-.003	-.011
Test Prep—Instruction						-.011	-.016	-.011
Test Prep—Activities						-.015	-.006	-.007
Self-Efficacy						-.009	-.006	.000
Formative Assessment						.008	.010	.005
Content Coverage—Numbers and Operations						.024	.034~	.021
Content Coverage—Algebra						.015	.001	.002
Intercept	1.397**	1.248*	1.137*	1.337**	1.449**	1.321*	1.436**	1.221*
Teacher Variance	0.015***	0.014***	0.012***	0.013***	0.011***	0.013***	0.010***	0.008***
Classroom Variance	0.016***	0.016***	0.017***	0.017***	0.017***	0.017***	0.017***	0.017***
Residual Variance	0.337***	0.337***	0.337***	0.337***	0.337***	0.337***	0.337***	0.337***
Adjusted Teacher-level R^2		.023	.129	.068	.213	.062	.295	.403
R^2 95% CI Lower Bound		-.075	.012	-.066	.039	-.028	.140	.246
R^2 95% CI Upper Bound		.242	.401	.359	.483	.179	.489	.634

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 includes student-, class-, and cohort-level controls for test scores and demographic characteristics (these parameter estimates are shown in Table A1). 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 (see the analysis section of the paper for details). A bootstrapped 95% confidence interval for this for the adjusted teacher-level R^2 is estimated by fitting 100 iterations of each model, each with a different random sample of teachers. Confidence intervals are bias-corrected and account for potential over-fitting in the original model.
*p <.10; **p <.05; ***p <.01.

Table 7. *Estimates of Parameters and Adjusted Teacher-level R^2 from Individual Regressions*

	Outcome: State Test		Outcome: Project-Administered Alternative Assessment	
	Parameter Estimate	Adjusted Teacher-level R^2	Parameter Estimate	Adjusted Teacher-level R^2
Male	-.009	-.004	.041	.005
White	-.059~	.012	-.028	-.010
Black	-.000	-.004	-.049	.023
Other Race	.109*	.019	.124**	.011
Master's Degree	.022	-.002	.010	-.003
# Math Courses	.018	.011	.022	.026
Math Content	.043*	.046	.040**	.079
Math Major	.043	.006	.001	-.003
Ed. Bachelor's	.072*	.013	.032	-.019
1–3 Yrs. Experience	-.074~	.007	-.029	-.007
4–10 Yrs. Experience	.053~	.022	.019	.016
10+ Yrs. Experience	-.024	.006	-.010	.008
Elem. Math Certification	-.022	-.007	-.030	.000
Trad. Certification	.067	-.008	.053	-.006
Alt. Certification	-.016	-.008	.035	.013
No Certification	-.028	-.001	-.057	.006
MQI—Errors	-.023	.017	-.019	.022
MQI—Ambitious Instruction	.026~	.009	.019	.020
MQI—Classroom Work Connected to Math	.030*	.042	.019	.038
CLASS—Class Organization	.042**	.031	.027~	.021
CLASS—Support	.026~	.008	.004	-.003
Knowledge—Student Ability	.035*	.040	.031*	.099
Knowledge—Student Misconceptions	-.011	-.005	-.014	.000
Knowledge—Teaching and Content	.032*	.043	.040***	.141
Effort	.032*	.051	.004	.000
Test Prep—Instruction	-.015	.009	-.013	.007
Test Prep—Activities	.010	-.001	-.010	-.003
Self-Efficacy	.000	-.003	.002	-.001
Formative Assessment	.032*	.029	.010	.012
Content Coverage—Numbers and Operations	.025	.020	.030*	.055
Content Coverage—Algebra	.045**	.063	.024~	.035

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 Tables 5 and 6) 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). Each row corresponds to a different regression model with only one teacher-level variable.

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

Appendix

Table A1. *Estimates of Student-Level Parameters from Hierarchical Models of State-Administered Assessments*

		Model 0: Baseline	Model 1: Demographics	Model 2: Background	Model 3: Observations	Model 4: Knowledge	Model 5: Other Characteristics	Predictors from Models 3-5	All Predictors
Student-Level Controls	Prior State Math Test Score	0.476***	0.476***	0.475***	0.476***	0.477***	0.476***	0.476***	0.475***
	Prior State Score, Squared	-0.026***	-0.026***	-0.026***	-0.026***	-0.026***	-0.026***	-0.026***	-0.026***
	Prior State Score, Cubed	-0.019***	-0.019***	-0.019***	-0.019***	-0.019***	-0.019***	-0.019***	-0.019***
	Grade 5 X Prior State Score Interaction	-0.031	-0.031	-0.029	-0.032	-0.032	-0.031	-0.034	-0.032
	Prior Alt. Math Test Score	0.321***	0.321***	0.321***	0.322***	0.322***	0.320***	0.321***	0.320***
	Prior Alt. Score, Squared	0.021**	0.021***	0.021***	0.021**	0.021***	0.020**	0.021**	0.021**
	Prior Alt. Score, Cubed	-0.005~	-0.005~	-0.005~	-0.005~	-0.005~	-0.005~	-0.005~	-0.005~
	Grade 5 X Prior Alt. Score Interaction	-0.044~	-0.044~	-0.045~	-0.044~	-0.045~	-0.042~	-0.042~	-0.041~
	Prior State ELA Test Score	0.128***	0.128***	0.128***	0.128***	0.128***	0.128***	0.128***	0.128***
	Mi Prior State ELA Test Score Ind	-0.018	-0.017	-0.018	-0.017	-0.018	-0.017	-0.015	-0.013
	Male	0.017	0.016	0.017	0.016	0.017	0.017	0.017	0.017
	Black	-0.043*	-0.043*	-0.043*	-0.043*	-0.043*	-0.043*	-0.043*	-0.043*
	Asian	0.130***	0.130***	0.130***	0.130***	0.130***	0.130***	0.130***	0.130***
	Hispanic	-0.007	-0.006	-0.006	-0.006	-0.007	-0.007	-0.007	-0.006
	English Language Learner	-0.047**	-0.047**	-0.047**	-0.047**	-0.047**	-0.047**	-0.047**	-0.048**
	Subsidized-Lunch Eligibility	-0.050***	-0.050***	-0.050***	-0.050***	-0.050***	-0.050***	-0.050***	-0.050***
	Special Education Status	-0.104***	-0.104***	-0.104***	-0.104***	-0.104***	-0.104***	-0.104***	-0.105***
	"Other" Race	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.004
Cohort-Level Controls	Cohort Size	-0.001*	-0.001*	-0.001*	-0.001*	-0.001*	-0.001*	-0.001*	-0.001
	Co. % Male	0.071	0.044	0.071	0.019	0.027	0.089	-0.005	-0.018
	Co. % Black	-0.338	-0.408	-0.160	-0.210	-0.277	-0.364	-0.244	-0.261
	Co. % Asian	0.100	-0.032	0.323	0.189	0.155	0.100	0.140	0.107
	Co. % Hispanic	-0.048	-0.116	0.095	0.039	-0.018	-0.085	-0.030	-0.055
	Co. % White	-0.162	-0.223	0.013	-0.073	-0.094	-0.176	-0.059	-0.032
	Co. % English Language Learner	-0.221	-0.227	-0.209	-0.232	-0.201	-0.183	-0.156	-0.167
	Co. % Subsidized-Lunch Eligible	-0.072	-0.073	-0.121	-0.069	-0.091	-0.075	-0.098	-0.135
	Co. % Special Education Status	-0.645**	-0.615**	-0.614**	-0.744**	-0.659**	-0.679**	-0.745**	-0.686**
	Co. Average State Math Prior	-0.402***	-0.402***	-0.407***	-0.412***	-0.373***	-0.392***	-0.379***	-0.393***
	Co. Avg. State ELA Prior	0.261*	0.267*	0.236*	0.258*	0.242*	0.248*	0.217~	0.213~
	Co. % Missing State Math Prior	0.370	0.340	0.305	0.500	0.539	0.290	0.524	0.453
	Co. % Missing State ELA Prior	-0.108	-0.063	-0.022	-0.300	-0.373	-0.085	-0.445	-0.324
	Co. Average Alt. Math Test Prior	-0.058	-0.062	-0.050	-0.061	-0.064	-0.064	-0.062	-0.061
	Co. % Missing Alt. Math Prior	-0.433~	-0.405~	-0.459*	-0.472*	-0.412~	-0.425~	-0.410~	-0.415~
Class-Level Controls	Class Size	-0.004	-0.004	-0.005	-0.004	-0.004	-0.006~	-0.005	-0.006~
	Cl. % Male	0.074	0.064	0.088	0.104	0.069	0.050	0.081	0.064
	Cl. % Black	-0.276	-0.252	-0.308	-0.300	-0.293	-0.233	-0.282	-0.275
	Cl. % Asian	-0.100	-0.103	-0.159	-0.168	-0.155	-0.045	-0.153	-0.183
	Cl. % Hispanic	-0.257	-0.226	-0.303	-0.270	-0.275	-0.205	-0.248	-0.243
	Cl. % White	-0.302	-0.257	-0.320	-0.320	-0.367	-0.255	-0.328	-0.306
	Co. % English Language Learner	0.009	-0.012	0.036	0.022	0.009	-0.020	-0.005	-0.003
	Cl. % Subsidized-Lunch Eligible	-0.038	-0.022	-0.028	-0.054	-0.015	-0.048	-0.032	-0.003
	Cl. % Special Education Status	0.104	0.123	0.106	0.115	0.078	0.113	0.097	0.131
	Cl. Average State Math Prior	0.142~	0.150~	0.160*	0.150~	0.125	0.136~	0.148~	0.167*
	Cl. Avg. State ELA Prior	-0.097	-0.091	-0.106	-0.092	-0.093	-0.096	-0.101	-0.097
	Cl. % Missing State Math Prior	-0.919*	-0.909*	-0.881*	-0.981*	-0.987*	-0.946*	-1.066*	-1.008*
	Cl. % Missing State ELA Prior	0.900*	0.851*	0.916*	0.983**	0.985**	0.961*	1.104**	1.047**
	Cl. Average Alt. Math Test Prior	-0.029	-0.034	-0.033	-0.030	-0.027	-0.025	-0.034	-0.041
	Cl. % Missing Alt. Math Prior	-0.280	-0.269	-0.273	-0.254	-0.330~	-0.230	-0.255	-0.253

Sample Indicator Variables	Grade-Year Indicator 1	0.065*	0.063*	0.073*	0.066*	0.069*	0.061~	0.067*	0.073*
	Grade-Year Indicator 2	-0.093**	-0.088*	-0.087*	-0.094**	-0.099**	-0.094*	-0.113**	-0.095*
	Grade-Year Indicator 3	-0.077*	-0.073~	-0.070~	-0.077*	-0.076*	-0.086*	-0.099*	-0.082~
	District Fixed Effect 1	0.051	0.046	0.042	0.036	0.042	0.057	0.026	0.020
	District Fixed Effect 2	0.079	0.065	0.086	0.049	0.062	0.075	0.043	0.035
	District Fixed Effect 3	-0.101~	-0.112~	-0.113~	-0.117*	-0.101~	-0.104~	-0.124*	-0.148**
	Imputed Teacher Fall TQ Var.	0.078	0.063	0.100	0.072	0.104	0.113	0.118	0.134
	Imputed Teacher Spring TQ Var.	-0.258	-0.246	-0.276	-0.273	-0.299	-0.246	-0.324~	-0.310~
	Imputed Teacher Background Var.	-0.001	-0.015	0.008	0.014	0.012	-0.006	0.019	-0.005

Note: Sample includes 283 teachers and 7,843 students. *p <.10; **p <.05; ***p <.01.

Table A2. *Estimates of Student-Level Parameters from Hierarchical Models of the Project-Administered Alternative Assessment*

		Model 0: Baseline	Model 1: Demographics	Model 2: Background	Model 3: Observations	Model 4: Knowledge	Model 5: Other Characteristics	Predictors from Models 3-5	All Predictors
Student-Level Controls	Prior State Math Test Score	0.432***	0.431***	0.430***	0.432***	0.433***	0.431***	0.432***	0.431***
	Prior State Score, Squared	-0.007	-0.007	-0.006	-0.007	-0.007	-0.007	-0.007	-0.007
	Prior State Score, Cubed	-0.026***	-0.025***	-0.026***	-0.025***	-0.026***	-0.026***	-0.025***	-0.025***
	Grade 5 X Prior State Score Interaction	-0.007	-0.006	-0.004	-0.008	-0.009	-0.007	-0.010	-0.007
	Prior Alt. Math Test Score	0.416***	0.416***	0.415***	0.416***	0.417***	0.416***	0.416***	0.415***
	Prior Alt. Score, Squared	0.040***	0.040***	0.040***	0.040***	0.040***	0.040***	0.041***	0.041***
	Prior Alt. Score, Cubed	-0.019***	-0.019***	-0.019***	-0.019***	-0.019***	-0.019***	-0.019***	-0.019***
	Grade 5 X Prior Alt. Score Interaction	0.132***	0.132***	0.132***	0.133***	0.132***	0.133***	0.134***	0.135***
	Prior State ELA Test Score	0.096***	0.096***	0.096***	0.096***	0.096***	0.096***	0.096***	0.096***
	Mi Prior State ELA Test Score Ind	0.040	0.044	0.040	0.041	0.042	0.040	0.044	0.046
	Male	0.027*	0.027*	0.027*	0.027*	0.027*	0.027*	0.027*	0.027*
	Black	-0.091***	-0.091***	-0.091***	-0.092***	-0.092***	-0.091***	-0.092***	-0.091***
	Asian	0.082**	0.082**	0.082**	0.082**	0.082**	0.082**	0.082**	0.083**
	Hispanic	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005
	English Language Learner	-0.033	-0.032	-0.033	-0.033	-0.033	-0.033	-0.033	-0.033
	Subsidized-Lunch Eligibility	-0.029~	-0.029~	-0.029~	-0.029~	-0.029~	-0.029~	-0.029~	-0.029~
	Special Education Status	-0.124***	-0.125***	-0.125***	-0.125***	-0.125***	-0.125***	-0.125***	-0.125***
	"Other" Race	-0.041	-0.040	-0.040	-0.041	-0.041	-0.040	-0.040	-0.040
Cohort-Level Controls	Cohort Size	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
	Co. % Male	0.432~	0.465~	0.440~	0.418	0.390	0.451~	0.352	0.357
	Co. % Black	-0.712	-0.650	-0.536	-0.614	-0.649	-0.647	-0.649	-0.583
	Co. % Asian	-0.227	-0.219	-0.064	-0.142	-0.143	-0.155	-0.204	-0.188
	Co. % Hispanic	-0.134	-0.096	0.019	-0.058	-0.077	-0.101	-0.107	-0.065
	Co. % White	-0.434	-0.409	-0.296	-0.351	-0.369	-0.407	-0.384	-0.334
	Co. % English Language Learner	-0.375*	-0.393*	-0.372*	-0.389*	-0.351*	-0.328~	-0.310~	-0.326~
	Co. % Subsidized-Lunch Eligible	-0.254	-0.255	-0.332~	-0.260	-0.301	-0.268	-0.336~	-0.397*
	Co. % Special Education Status	-0.535*	-0.505*	-0.523*	-0.618**	-0.575**	-0.544*	-0.591**	-0.546*
	Co. Average State Math Prior	-0.388***	-0.371***	-0.380***	-0.393***	-0.355**	-0.386***	-0.365***	-0.355**
	Co. Avg. State ELA Prior	0.099	0.107	0.069	0.089	0.071	0.098	0.044	0.041
	Co. % Missing State Math Prior	0.024	-0.004	0.120	0.088	0.261	0.048	0.251	0.330
	Co. % Missing State ELA Prior	0.427	0.488	0.331	0.341	0.073	0.397	0.084	0.048
	Co. Average Alt. Math Test Prior	-0.033	-0.046	-0.030	-0.040	-0.038	-0.038	-0.038	-0.051
	Co. % Missing Alt. Math Prior	-0.471*	-0.441~	-0.501*	-0.521*	-0.461~	-0.442~	-0.443~	-0.462~
Class-Level Controls	Class Size	-0.007*	-0.007*	-0.007*	-0.007*	-0.007*	-0.008*	-0.008*	-0.008*
	Cl. % Male	0.010	-0.028	0.022	0.033	0.002	0.002	0.031	-0.004
	Cl. % Black	-0.166	-0.133	-0.174	-0.170	-0.187	-0.128	-0.126	-0.104
	Cl. % Asian	-0.040	-0.051	-0.041	-0.095	-0.119	-0.012	-0.089	-0.111
	Cl. % Hispanic	-0.415	-0.366	-0.450	-0.414	-0.455	-0.350	-0.380	-0.362
	Cl. % White	-0.133	-0.060	-0.124	-0.139	-0.220	-0.065	-0.109	-0.071
	Co. % English Language Learner	0.111	0.086	0.134	0.127	0.103	0.083	0.090	0.098
	Cl. % Subsidized-Lunch Eligible	-0.026	0.015	-0.007	-0.039	0.004	-0.016	0.017	0.055
	Cl. % Special Education Status	0.175	0.188	0.177	0.188	0.141	0.165	0.142	0.186
	Cl. Average State Math Prior	0.003	0.007	0.008	0.009	-0.028	-0.002	0.003	0.005
	Cl. Avg. State ELA Prior	0.026	0.035	0.022	0.025	0.032	0.023	0.009	0.024
	Cl. % Missing State Math Prior	-0.584	-0.506	-0.624	-0.604	-0.661	-0.578	-0.623	-0.617
	Cl. % Missing State ELA Prior	0.308	0.190	0.386	0.349	0.406	0.304	0.387	0.347
	Cl. Average Alt. Math Test Prior	0.004	0.002	0.009	0.008	0.013	0.001	0.003	0.010
	Cl. % Missing Alt. Math Prior	-0.149	-0.133	-0.162	-0.123	-0.214	-0.125	-0.163	-0.175

Sample Indicator Variables	Grade-Year Indicator 1	-0.217***	-0.218***	-0.206***	-0.215***	-0.209***	-0.224***	-0.211***	-0.202***
	Grade-Year Indicator 2	0.650***	0.658***	0.654***	0.646***	0.645***	0.635***	0.613***	0.635***
	Grade-Year Indicator 3	0.257***	0.267***	0.262***	0.258***	0.260***	0.233***	0.221***	0.245***
	District Fixed Effect 1	-0.072	-0.079	-0.079	-0.087	-0.076	-0.072	-0.095~	-0.099~
	District Fixed Effect 2	-0.344***	-0.352***	-0.359***	-0.368***	-0.361***	-0.360***	-0.385***	-0.411***
	District Fixed Effect 3	0.081	0.082	0.070	0.069	0.085	0.084	0.066	0.049
	Imputed Teacher Fall TQ Var.	0.073	0.080	0.069	0.073	0.098	0.098	0.104	0.097
	Imputed Teacher Spring TQ Var.	-0.071	-0.064	-0.079	-0.066	-0.117	-0.063	-0.153	-0.123
	Imputed Teacher Background Var.	0.095*	0.103*	0.092~	0.111*	0.112*	0.100*	0.127**	0.103*

Note: Sample includes 283 teachers and 7,843 students. *p <.10; **p <.05; ***p <.01.

Table A3. *Summary Statistics of Teacher Demographic and Background Variables*

	Teachers (*n*)	**Mean**	**Min**	**Max**
Male	279	0.16	0	1
White	270	0.67	0	1
Black	270	0.23	0	1
Other Race	270	0.10	0	1
Master's Degree	280	0.77	0	1
# Math Courses	279	2.88	1	4
Math Content	281	2.48	1	4
Math Major	283	0.06	0	1
Ed. Bachelor's	283	0.53	0	1
Experience (Years)	279	10.29	0	37
Elem. Math. Certification	283	0.15	0	1
Trad. Certification	275	0.85	0	1
Alt. Certification	275	0.08	0	1
No Certification	275	0.08	0	1

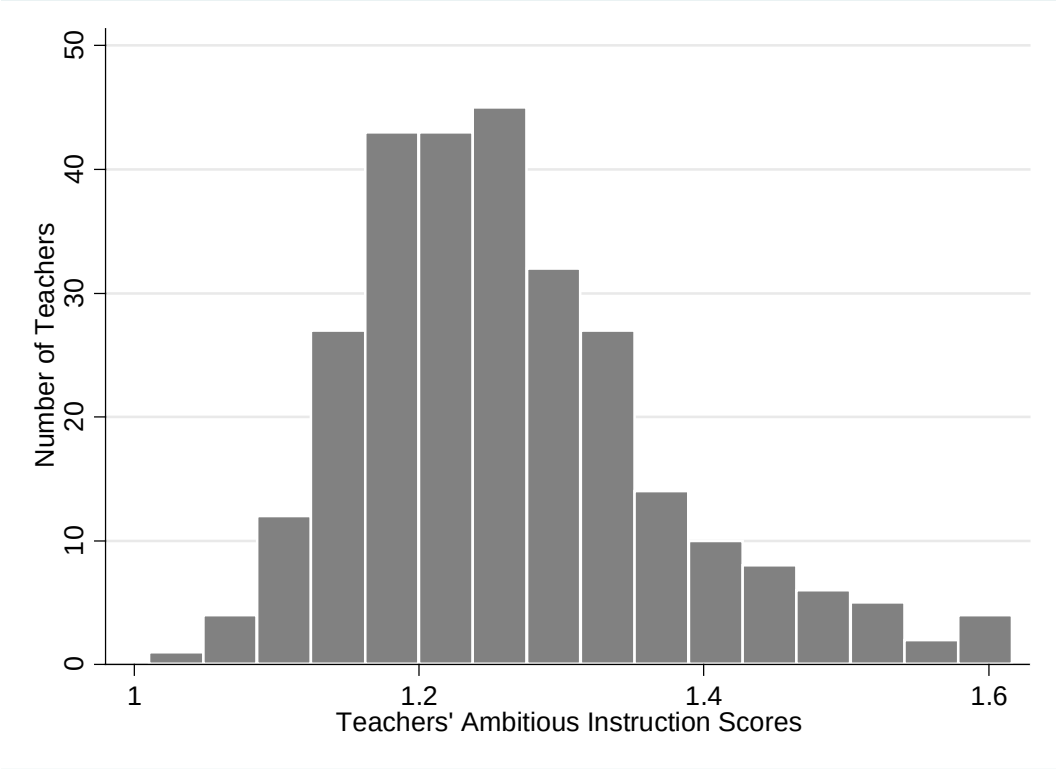

Figure A1. Raw Distribution of Teachers' Ambitious Instruction Scores.
Note: Sample includes 283 teachers. Possible range of Ambitious Instruction scores is [1,3].

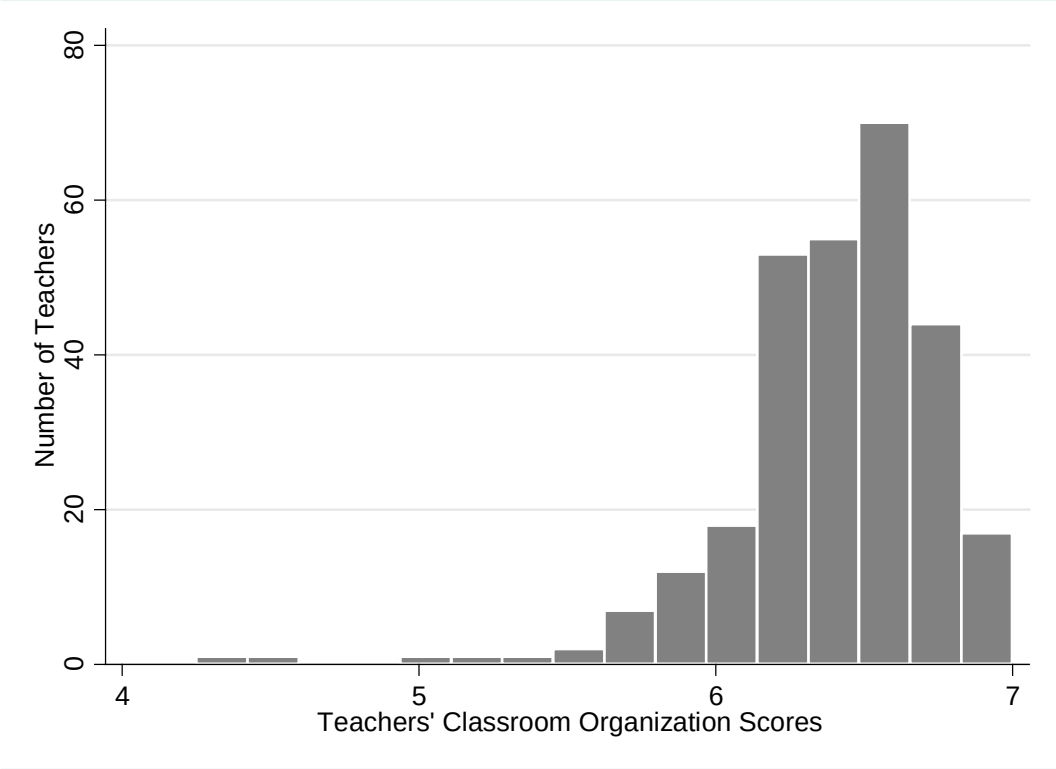

Figure A2. Raw Distribution of Teachers' Classroom Organization Scores.
Note: Sample includes 283 teachers. Possible range of classroom organization scores is [1,7].

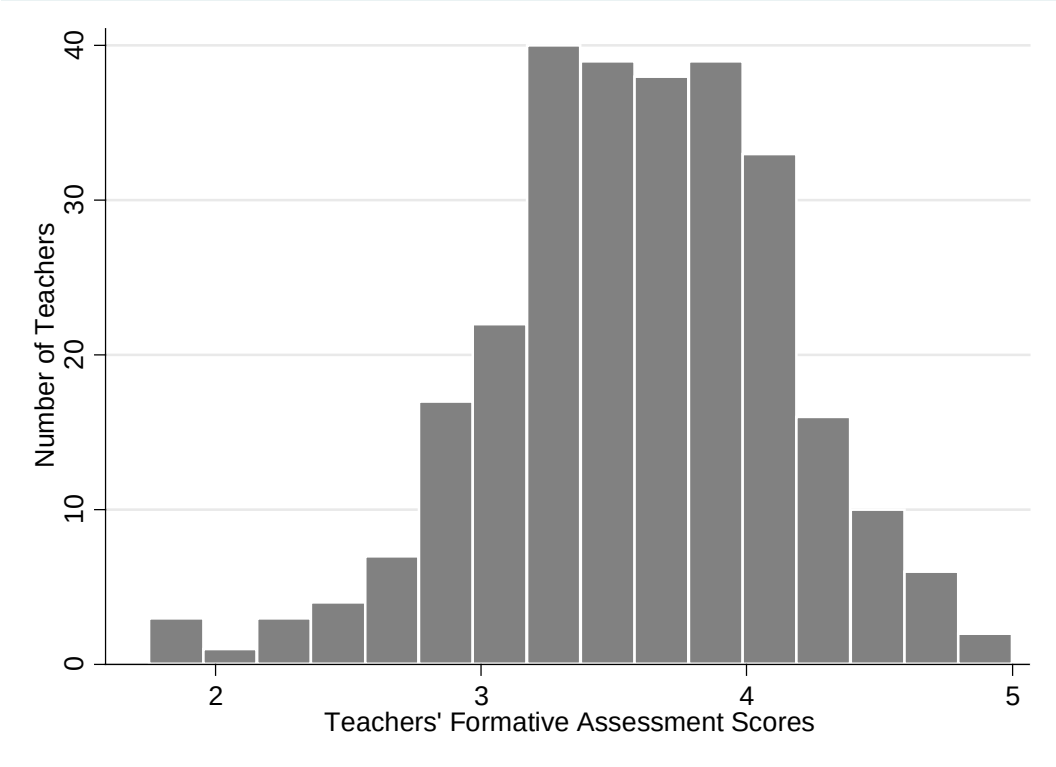

Figure A3. Raw Distribution of Teachers' Formative Assessment.
Note: Sample includes 281 teachers. Possible range of formative assessment scores is [1,5].