The Multidimensional Impact of Teachers on Students

Nathan Petek

Federal Trade Commission

Nolan G. Pope

University of Maryland

Test score measures of teacher quality may not fully capture teachers' impact on students. We use test score and non–test score measures of student achievement and behavior to estimate multiple dimensions of teacher quality. We find that these two measures of teacher quality are only weakly correlated and that both affect students' high school performance. A teacher removal simulation that uses both measures improves most long-term student outcomes by over 50%, compared to a policy that uses test scores alone. Our results also show that for high school outcomes the effects of teachers in later grades are larger than those in earlier grades and that performance in core elementary school subjects matters more than that in other subjects.

I. Introduction

Teacher quality has garnered the attention of policy makers and researchers for many years. Researchers have primarily measured teacher quality

We would like to thank Michael Gilraine, John Eric Humphries, Steven Levitt, Jens Ludwig, Ofer Malamud, Magne Mogstad, Derek Neal, and Eric Nielsen for helpful comments and discussion. The views expressed in this article are those of the authors. They do not necessarily represent those of the Federal Trade Commission or any of its commissioners. This paper was edited by James J. Heckman.

Electronically published February 28, 2023

Journal of Political Economy, volume 131, number 4, April 2023.

https://doi.org/10.1086/722227

using a test score value-added framework.[1] Although the use of test score value-added has substantially affected education research, people have long recognized that good teachers likely affect a wide range of student outcomes. In fact, early theoretical formulations of value-added used an education production function that modeled educational output as a "multidimensional factor" (Hanushek 1971). Consequently, measures of teacher quality that rely solely on student test scores may not fully capture the impact teachers have on students.

In this paper, we are interested in whether teachers can noticeably affect measures of student achievement beyond test scores and to what extent the impact on non–test score measures is important for the future success of students. We use the value-added framework to construct separate measures of teacher ability to improve test scores, behavior, and a plausible measure of noncognitive skills. We use these multiple value-added estimates of teacher ability to explore the effect of teachers on students' long-term outcomes and the relative importance of cognitive and noncognitive skills in the production of human capital. We illustrate the benefits of using broader measures of teacher ability by investigating the extent to which using multiple measures of teacher ability increases the efficacy of teacher selection and assignment policies, improves the measurement of the cumulative return to high-quality teaching, and allows the measurement of teacher quality in untested subjects.

We gather administrative data from the Los Angeles Unified School District (LAUSD) for students in grades K–12 from 2003 to 2015. These data link over a million students to teachers and track students over time as they progress through the LAUSD system. Our three measures of student achievement are constructed from (1) student math and English CST (California Standards Test) scores, (2) measures of student behavior, including suspensions, attendance, GPA (grade point average), and grade retention, and (3) teacher assessments of student effort and 14 learning skills that are plausible measures of noncognitive ability. The learning skills include teacher assessments such as whether a student makes good use of time, exercises self-control, and resolves conflicts appropriately. We measure the long-term effects of teachers using student performance in high school, including dropping out of high school, taking the SAT, SAT scores, CAHSEE (California High School Exit Examination) scores, GPA, teacher assessments of effort and cooperation, attendance, suspensions, and grade retention.

We first document that elementary school students with better test scores, behavior, and learning skills perform better in high school. We then estimate teacher value-added measures of three dimensions of teacher

[1] An important exception is a paper by Kirabo Jackson (2018) that estimates non–test score measures of teacher quality that we discuss in more detail below.

quality: student test scores (using math and English CST), student behavior (using GPA, attendance, suspensions, and grade retention), and student learning skills (using teacher assessments of effort and 14 learning skills). To avoid bias and potential teacher manipulation when using teacher-reported non–test score variables, we modify the standard value-added framework to use student outcomes from the year after the student was in a teacher's class, instead of the contemporaneous year. Using these value-added measures, we show that teachers affect both test score and non–test score dimensions of student achievement, but we find little evidence that learning skills value-added affects high school outcomes.

We find that having a high–test score value-added teacher in elementary school improves students' high school performance. These long-term effects of test score value-added are not substantially reduced by adding teachers' behavior or learning skills value-added to the model. This result suggests that the long-term effects of test score value-added may not be biased by omitting non–test score teaching ability.

We also find that behavior value-added is only weakly correlated with test score value-added and has a similarly large effect on students' long-term outcomes. Therefore, test score value-added misses the dimensions of teacher quality captured by behavior value-added that matter for long-term outcomes. Consequently, test score value-added underestimates the total effect of teachers on students. However, we find no evidence of an interaction effect for teachers who are better or worse on both dimensions of teacher ability.

The low correlation between the two value-added measures also suggests that using behavior value-added in conjunction with test score value-added may substantially enhance the accuracy with which overall teacher quality is measured. We illustrate how behavior value-added enhances the measurement of teacher quality, using a hypothetical policy simulation that replaces teachers in the bottom 5% of the teacher quality distribution with district-average teachers. Relative to relying on test score value-added alone, a simple rule that equally weights the test score and behavior value-added of a teacher increases the efficacy of the policy by at least 50% for the likelihood of dropping out of high school, taking the SAT, high school GPA, suspensions, absences, and on-time progression. These gains are obtained with little to no decline in student test scores, are similar to the gains obtained if an optimal weighting scheme is used, and do not require administering additional tests or using data beyond what schools typically collect.

Finally, we use test score and non–test score measures of ability in two applications. First, we estimate the effect of test score and behavior value-added for each grade from 3 to 12. We find that middle school and high school teachers have a larger effect on outcomes measured in eleventh or twelfth grade than elementary school teachers. This result

suggests that teachers in later grades may play a more important role in benefiting students' high school outcomes than teachers in earlier grades, under the strong assumption that a 1–standard deviation change in teacher value-added induces the same amount of learning in each grade. Assuming constant returns to higher-quality teachers, these results imply large cumulative benefits of teacher value-added. For example, giving students a teacher with a standard deviation better test score value-added each year from grade 3 to grade 12 increases the likelihood of taking the SAT by 8.1 percentage points and reduces the likelihood of dropping out of high school by 0.5 percentage points. Giving students a teacher with a standard deviation better behavior value-added over the same period increases the likelihood of taking the SAT by 8.4 percentage points and reduces the likelihood of dropping out of high school by 5.9 percentage points. The cumulative effects of better teachers are only somewhat reduced by controls for tracking.

Second, the focus on test scores has limited the study of teacher quality to a few regularly tested subjects (i.e., math and English). We instead use subject-specific GPAs to compute value-added measures of teacher quality in 10 elementary school subjects. We find that students with higher value-added teachers in reading and health perform better in high school, whereas having a higher value-added teacher in speaking has negative effects on high school performance. Hiring teachers who are relatively better at teaching reading could potentially benefit the long-term outcomes of students.

From a policy perspective, there are potentially large benefits from adopting a measure of teacher quality that includes both test score and non–test score dimensions. For example, policy makers can use non–test score value-added to measure teacher quality for all teachers, not just math and English teachers. In addition, since focusing on only one output of the multidimensional education production function (i.e., test scores) may distort the efficient allocation of teachers' time and resources, using a broader measure of teacher quality may help alleviate this distortion. Finally, using a better measure of overall teacher quality can make school districts' hiring and tenure decisions more effective.

Our paper contributes to a literature that estimates the effect of various non–test score value-added measures on contemporaneous outcomes (Jennings and DiPrete 2010; Ruzek et al. 2015; Gershenson 2016; Blazar and Kraft 2017; Kraft 2019), on within–high school outcomes (Jackson 2018), and on outcomes of 20-year-olds (Flèche 2017).[2] The paper most

[2] The non–test score value-added measures include social and behavioral skills (Jennings and DiPrete 2010); motivation (Ruzek et al. 2015); absences (Gershenson 2016); belief in the ability to do math and happiness in math class (Blazar and Kraft 2017); grit, growth mindset, effort, and answering open-ended questions (Kraft 2019); absences, suspensions, grades,

closely related to ours is Jackson (2018). Using North Carolina data, he finds that above and beyond test scores, teachers affect proxies for noncognitive skills (behavior value-added) in ninth grade and subsequently outcomes in twelfth grade, such as high school completion, SAT taking, and intentions to attend college. He finds that including both test score and behavior value-added measures in ninth grade more than doubles the predictable variability of teacher effects on twelfth-grade outcomes.

Our paper contributes in several ways. First, we estimate measures of behavior and noncognitive value-added in addition to the standard test score value-added. We find that the non–test score value-added measures are only weakly correlated with test score value-added. Importantly, this additional dimension of teacher quality matters for students' long-term outcomes, even independent of test score value-added. Furthermore, the estimated effects of test score value-added on long-term outcomes are not biased by omitting these additional dimensions of teacher quality. Our results indicate that incorporating both of these measures of teacher quality in teacher hiring and retention decisions would substantially improve students' high school outcomes. In addition, we can estimate the benefit of having a higher-quality teacher in each grade from 3 through 11, which allows for comparisons in the effect of teacher quality across grades and the cumulative effect of increasing teacher quality.[3] Finally, the test score value-added literature has been limited to measuring teacher performance in tested subjects, primarily English and math (Jackson, Rockoff, and Staiger 2014). Our extension of the value-added framework to measure subject-specific GPA value-added provides novel estimates of the effect of teaching quality on long-term outcomes in subjects that are not tested.

We also contribute to the broader literature on the role of cognitive and noncognitive skills in the production of human capital and long-term outcomes (e.g., Heckman, Stixrud, and Urzua 2006; Cunha, Heckman, and Schennach 2010) by analyzing how teachers with varying levels of ability to increase students' cognitive and noncognitive skills affect their students' long-term outcomes. More specifically, we contribute to understanding the role that the development of cognitive and noncognitive skills plays in the long-term effects of educational interventions (Heckman, Pinto, and Savelyev 2013), using a different source of variation in cognitive and noncognitive skills.

The rest of the paper proceeds as follows. Section II provides background information on value-added scores. Section III describes the

---

and grade progression (Jackson 2018); and internalizing and externalizing behavior (Flèche 2017). Araujo et al. (2016) produce short-term estimates of classroom value-added on measures of executive function. Other studies assess multidimensional teacher effects with non–value-added approaches (Rockoff and Speroni 2010; Mihaly et al. 2013).

[3] Chetty, Friedman, and Rockoff (2014b) estimated test score value-added for grades 4–8, and Jackson (2018) estimated test score and non–test score value-added in grade 9.

LAUSD data that we use and, in particular, describes the variables used to measure test score, behavior, and learning skills value-added. Section IV outlines the empirical method for estimating teacher value-added measures and the effect of teacher value-added on long-term student outcomes. Section V presents the descriptive results of the test score, behavior, and learning skills value-added of teachers and then reports the results for how teachers affect students' concurrent and long-term outcomes. The gains from teacher removal policies that use multiple dimensions of teacher quality are also presented. Section VI presents the relative value of higher-quality teachers over the students' educational life cycle and in specific subjects. Section VII concludes.

II. Background on Test Score Value-Added

Since the early 1970s, researchers have used test score valued-added to measure teacher quality (Hanushek 1971). This research led states and school districts to use test score value-added in teacher evaluations as early as the 1990s (Horn and Sanders 1994). Since then, the use of test score value-added has expanded, and 27 states now require that teacher evaluations include "growth measures as a significant criterion" (National Council on Teacher Quality 2015). This increased use of test score value-added has largely been due to a lack of other predictors of teacher quality (Hanushek and Rivkin 2010). Much of the recent work in the value-added literature focuses on the validity of value-added models (Rockoff 2004; Kane and Staiger 2008; Rothstein 2010, 2017; Kane et al. 2013; Bacher-Hicks, Kane, and Staiger 2014; Chetty, Friedman, and Rockoff 2014a, 2017), gains from using them in personnel decisions (Gordon, Kane, and Staiger 2006; Goldhaber and Hansen 2010; Hanushek 2011), and theoretical and empirical studies of their use in pay for performance (Neal 2011; Fryer 2013; Goodman and Turner 2013). In particular, Chetty, Friedman, and Rockoff (2014b) find that students with higher test score value-added teachers earn significantly more by their late 20s, have fewer births as teenagers, and are more likely to attend college.

III. Los Angeles Student Data

The LAUSD is the second-largest school district in the United States, educating over 600,000 students each year. In 2003, the school district was 71.9% Hispanic, 12.1% Black, and 9.4% white.[4] We use a panel of student-level administrative data on all public school students in the LAUSD. The panel links students to teachers over time and includes the 2002–3 to 2014–15 school years, which we reference by year of graduation

[4] Statistics can be found at http://dq.cde.ca.gov/dataquest.

(e.g., we refer to the 2002–3 school year as 2003). Our analysis focuses on the over 110,000 third- to fifth-grade students studying in the LAUSD each year.

These data are unique in the level of detail they provide about each student's academic performance. For grades 2–11, math and English CST scores are available for each student. The testing regime is relatively consistent over this period, with the only major change being an essay section added to the fourth- and seventh-grade English tests in 2011. For all grades, these data contain the number of days a student was suspended, the number of days a student was absent, and whether a student did not progress on time to the next grade (i.e., was held back). Both elementary and high school students received progress reports with their grades by subject and a number of additional teacher assessments of student performance.

Elementary school (grades K–5) progress reports are given each trimester by the student's sole classroom teacher and contain achievement grades in 10 subjects (e.g., reading, mathematics, and art), effort grades for the same 10 subjects, grades for five "work and study habits" (e.g., "makes good use of time" and "organizes materials"), and grades for nine "learning and social skills" (e.g., "resolves conflicts appropriately" and "exercises self-control"). All grades are on a 4-point scale, with no fractional points given. We compute an annual GPA for each of the four groups listed above. Figure 1 shows a template of the progress report.

Starting in the sixth grade, middle school and high school students receive progress reports each semester from multiple classroom teachers, with three categories of grades for each of their classes: achievement (i.e., academic performance), "work habits," which we term "effort" (i.e., "effort," "responsibility," "attendance," and "evaluation"), and "cooperation" (i.e., "courtesy," "conduct," "improvement," and "class relations"). Achievement is graded on a 4-point scale, and effort and cooperation are graded on a 3-point scale, with no fractional points given. We compute annual GPAs for each of these three groups of measures. Figure A.1 shows additional details on grading criteria.

Additional data are available for middle and high school students, including whether a student dropped out of high school (i.e., the student enrolled in the LAUSD in grade 9 and did not graduate high school in the LAUSD within 5 years), graduated from the LAUSD conditional on enrolling in the LAUSD in twelfth grade, SAT scores, PSAT (preliminary SAT) scores, math and English CAHSEE scores, science CST scores (grades 5, 8, and 10), social science CST scores (grades 8 and 11 and world history), and the number of AP (advanced placement) courses taken. All test scores are normalized to be mean 0 and standard deviation 1 at the grade-year level, except both SAT and PSAT scores, which we place on a 600–2400 scale (the PSAT is normally on a 60–240 scale, and for some years, the SAT was on a 400–1600 scale). We top-code days absent at

PROGRESS REPORT

School Year:

Principal:

Teacher: Room: Grade Level:

Birth Date:

Academic Subjects	Grade Reporting Period 1 AC	1 EF	2 AC	2 EF	3 AC	3 EF
Reading						
ELD Reading						
Writing						
ELD Writing						
Listening						
ELD Listening						
Speaking						
ELD Speaking						
Mathematics						
History/Social Science						
Science						
Health Education						
Physical Education						
Arts						

ACHIEVEMENT SCORES
***Meets Standards**
4 = Advanced*
3 = Proficient*
2 = Partially Proficient
1 = Not Proficient

(ELD) ENGLISH LANGUAGE DEVELOPMENT SCORES
4 = Advanced Progress
3 = Average Progress
2 = Partial Progress
1 = Limited Progress

EFFORT SCORES
4 = Strong
3 = Consistent
2 = Inconsistent
1 = Poor

Work and Study Habits	Reporting Period 1	2	3
Makes good use of time			
Works independently			
Organizes materials			
Presents neat and careful work			
Completes homework on time			
Learning and Social Skills			
Follows directions and procedures			
Accepts and respects authority			
Cooperates well in a group situation			
Shows dependability			
Takes responsibility			
Exercises self-control			
Resolves conflicts appropriately			
Demonstrates appropriate social interaction with peers			
Demonstrates fairplay			

Student Assessment

Instructional Programs
Master Plan Program

ELD Level **Start Date** **End Date** **Grade Period**

Instructional Services
Interventions

Intervention **Date**

Office Copy

FIG. 1.—Blank copy of an LAUSD elementary school progress report. Each row labels the academic subject, work and study habits, or learning and social skill each student is graded on by their teacher. Columns 1–3 correspond to the three trimesters students receive a grade. For the academic subjects, the "AC" column stands for achievement scores and the "EF" column stands for effort scores. For all academic subjects, work and study habits, and learning and social skills, students receive a grade ranging from 1 (the poorest performing) to 4 (the best performing).

180 days per year and report log absences as the log of 1 plus the number of absences.

We compute test score value-added measures using math and English CST scores and behavior value-added using log days absent, achievement GPA, an indicator for suspensions, and an indicator for being held back; we describe the measures fully in sections IV.A and IV.B. For elementary school teachers, we compute learning skills value-added using the three additional types of elementary school GPAs. We reduce the dimensionality of both the inputs to the value-added variables and the value-added variables themselves by creating equally weighted indices.

Our main outcome variables are measures of high school performance, including an indicator for dropping out of high school, an indicator for taking the SAT, SAT scores, the three high school GPA measures averaged from grades 9–12, math and English CAHSEE scores, days suspended in grades 9–12, log absences in grades 9–12, and an indicator for being held back in grades 9–12. We treat graduation as a supplemental measure because it is conditional on enrolling in the LAUSD in twelfth grade. The unit of observation in the main analysis data set is a student–academic year, where the outcome is typically a measure of the student's performance in high school and the independent variables of interest are the three teacher value-added indices for a student in a particular academic year. We focus on the pooled effect of teachers in grades 3, 4, and 5, and thus the same student often appears in the main analysis three times.

Summary statistics for these data are shown in table 1. Panel A shows summary statistics for all students in grades 3–5 during the 2004–10 school years. Panel B shows high school summary statistics for the students in panel A who attended high school in the LAUSD. Figure A.2 shows retention rates by grade. While the dropout numbers seem high, they are consistent with LAUSD's official graduation rate, which was 64% in 2005–6 and 62% in 2009–10. The LAUSD dropout variable overestimates the dropout rate as a measure of people who never graduate from any school in CA by a factor of about 1.5 because it includes both dropouts and students who transferred to schools outside of the LAUSD. The actual LAUSD dropout rate, estimated by the LAUSD, was 37.2% in 2010. The graduated variable is an overestimate of the actual graduation rate because it is conditioned on entering twelfth grade.

IV. Empirical Method

A. Estimating Teacher Value-Added

Let S_{ijt} be a measure of student i's test scores, behavior, or learning skills in academic year t in teacher j's class. For example, S_{ijt} could be a standardized test score, an indicator for whether the student was suspended,

TABLE 1
SUMMARY STATISTICS

Variable	Mean	Standard Deviation	Observations
	A. Grades 3–5		
Math CST score	.01	1.00	891,643
English CST score	.00	1.00	891,751
GPA	2.88	.42	861,977
Effort GPA	3.14	.46	861,977
Learning skills GPA	3.10	.58	614,532
Fraction of days absent (%)	3.9		858,308
Days suspended	.05	.38	906,193
Held back (%)	.7		837,401
English learner (%)	42.0		906,193
	B. High School Outcomes		
LAUSD dropout (%)	54.6		333,513
Took SAT (%)	50.5		249,436
SAT score	1,330	298	145,265
GPA	2.25	.96	536,868
Effort GPA	2.12	.52	476,548
Cooperation GPA	2.33	.45	476,548
Math CAHSEE score	.07	1.01	331,266
English CAHSEE score	.08	.98	329,980
Days suspended	.18	.83	588,273
Fraction of days absent (%)	7.8		542,959
Held back a grade (%)	29.0		449,533
Graduated if entered 12th grade (%)	88.6		190,278
Took PSAT (%)	68.9		470,703
PSAT score	1,110	248	348,992
Math CST score	−.05	.99	124,044
English CST score	.00	.99	135,769
Grade 8 science CST score	.02	1.00	599,880
Grade 10 science CST score	.09	1.00	296,069
Grade 8 social science CST score	.03	1.00	548,439
Grade 11 social science CST score	.07	.99	160,483
World history CST score	.06	.98	270,403
Number of AP courses	.73	1.70	588,273

NOTE.—The sample includes students in grades 3–5 who attended the LAUSD. The unit of observation is a student–academic year. Panel A reports summary statistics for all LAUSD students in grades 3–5 from 2004 to 2010. Panel B reports high school summary statistics for all LAUSD students who were in grades 3–5 from 2004 to 2010 and attended high school in the LAUSD. Elementary school GPA, effort GPA, and learning skills GPA are on a 4-point scale. GPA in high school is on a 4-point scale, and effort GPA and cooperation GPA in high school are on a 3-point scale. All test scores except the SAT and PSAT are normalized at the grade-year level. Both the SAT score and the PSAT score are on a 600–2400 scale. The LAUSD dropout variable is the fraction of students who enrolled in an LAUSD school in ninth grade and did not graduate from an LAUSD school within 5 years. The "Graduated if entered 12th grade" variable shows the fraction of students who enrolled in an LAUSD school in twelfth grade and graduated from the LAUSD.

or a teacher's assessment of a particular learning skill. The goal is to estimate the effect of a teacher on several measures of students' test scores, behavior, and learning skills. Recent research estimating teacher test score value-added and its affects on long-term outcomes has used slightly

different estimation strategies (Chetty, Friedman, and Rockoff 2014b; Rothstein 2017; Jackson 2018). Our approach combines elements of each, and we show in the appendix that the main results are robust across a range of estimation strategies.

For test score value-added, we use the following estimation procedure. Although the procedures are very similar, a small but important adjustment is made when estimating behavior and learning skills value-added, which is discussed below. We construct value-added measures by first residualizing the achievement measure, S_{ijt}, by regressing it on a vector of controls, X_{ijt}, for lagged student achievement and the classroom environment, using equation (1). The baseline controls X_{ijt} include equation (1) lags of a cubic polynomial of the student's math test score, English test score, achievement GPA, effort GPA, work and study habits GPA, and learning and social skills GPA; (2) lags of log days absent, an indicator for suspensions, and an indicator for being held back; (3) current English-language-learner status; (4) a cubic polynomial of both class- and grade-level averages of lagged math test score, English test score, achievement GPA, effort GPA, work and study habits GPA, and learning and social skills GPA; (5) class- and grade-level averages of lagged log days absent, an indicator for suspensions, an indicator for being held back, and English-language-learner status; (6) current class size; and (7) grade fixed effects and year fixed effects. Each of the controls, except current English-language-learner status and current class size, are interacted with grade fixed effects. In middle and high school, our controls X_{ijt} are the same as in elementary school, except that we exclude effort GPA, work and study habits GPA, and learning and social skills GPA for lack of data.

$$S_{ijt} = \Gamma X_{ijt} + \varepsilon_{ijt}, \tag{1}$$

$$\varepsilon_{ijt} = \mu_{jt} + \alpha_c + \gamma_{it}. \tag{2}$$

We assume that the error term, ε_{ijt}, is an additively separable function of teacher quality (μ_{jt}), classroom shocks (α_c), and student-year shocks (γ_{it}), as defined in equation (2). This specification of the error term, ε_{ijt}, is more flexible than one often used in the value-added literature that requires a teacher's quality be the same in each year. This approach requires a stationarity assumption to separately identify μ_{jt} and α_c.[5]

Let ν_{ijt} be the residualized student achievement, computed as follows:

$$\nu_{ijt} = S_{ijt} - \hat{\Gamma} X_{ijt}. \tag{3}$$

The residualization purges S_{ijt} of measures of the prior achievement of each student, each student's class, and each student's grade. For middle

[5] See assumption 1 of Chetty, Friedman, and Rockoff (2014a).

and high school students, we obtain the residuals using only students who had one teacher per subject for English or math.

We then take the mean of the residuals, $\bar{\nu}_{jt} = (1/N)\Sigma_{i=1}^{N}\nu_{ijt}$, by year for each teacher j. This provides an estimate of the teacher's value-added score, which is a measure of their ability to affect student achievement in each year t. It is unbiased as long as certain teachers do not tend to receive students with relatively better- or worse-than-average unobserved achievement, specifically, if $E[\alpha_i + \gamma_{it}|j] = E[\alpha_i + \gamma_{it}]$ (Chetty, Friedman, and Rockoff 2011). Although this is a strong assumption, it is plausible in this context because the value-added model includes extensive controls for students' prior achievement and behavior in school that have been shown to account for most student sorting in test score value-added models (Bacher-Hicks, Kane, and Staiger 2014; Chetty, Friedman, and Rockoff 2014a). To help alleviate concerns about student sorting based on unobservable components of student achievement, in section V.E we check for forecast bias, examine the effect of teacher value-added on predicted outcomes as a placebo test, and perform a quasi-experimental analysis that uses teachers switching grades and schools. We also show that our estimates are robust to including controls for tracking in middle and high school.

We then predict teacher quality in year t with the teacher's estimated value-added scores in the surrounding years, using the equation $\hat{\nu}_{jt} = \Sigma_{s=t-a}^{t+a}\hat{\psi}_s\bar{\nu}_{js}1[s \neq t]$, where a equals 6 years for test scores and 5 years for behavior and learning skills. This approach measures teacher quality with data from the surrounding years to avoid biasing estimates of the long-term effects of teacher quality on student outcomes (Jacob, Lefgren, and Sims 2010). Including year t in the prediction would likely bias the long-term estimates, because unobservables in year t that affect any dimension of student performance may also affect both the estimated value-added measure in year t and the long-term outcomes. We allow the weight placed on the value-added measure, $\hat{\psi}_s$, to vary by the number of years before or after year t. We estimate the weights by minimizing the mean-squared error of the difference between $\bar{\nu}_{jt}$ and predictions of $\bar{\nu}_{jt}$ made with the teacher value-added measures the years before and after t—specifically, by solving the following minimization problem: $\boldsymbol{\psi} = \arg\min_{\{\psi_{t-a},\ldots,\psi_{t+a}\}}\Sigma_j^J(\bar{\nu}_{jt} - \Sigma_{s=t-a}^{t+a}\psi\bar{\nu}_{js}1[s \neq t])^2$.[6] This procedure produces leave-year-out jackknife value-added predictions that allow teacher quality to change over time, and it shrinks the value-added predictions to the mean through Bayesian shrinkage (Chetty, Friedman, and Rockoff 2014a).

We modify this procedure when we calculate non–test score value-added measures by using the lead of the achievement measure, $S_{ij(t+1)}$, as the outcome variable. This approach contrasts with most of the non–test score

[6] See Chetty, Friedman, and Rockoff (2014a) for additional details.

value-added literature, which uses contemporaneous student outcomes instead of outcomes in the next year, but is closely related to approaches used to calculate the value-added of professors (Carrell and West 2010; Figlio, Schapiro, and Soter 2015). This approach requires the main assumption—that teachers do not systematically receive students with relatively better or worse unobserved achievement—holds for 2 years instead of just one. This assumption is otherwise identical to the assumption used in the value-added literature.

We use $S_{ij(t+1)}$ because using S_{ijt} creates the potential for the non–test score value-added measures to capture aspects of teacher behavior unrelated to teachers' ability to affect students' behavior or learning skills. For example, grades are likely affected not only by how much a teacher helps a student learn and work diligently but also by how strictly the teacher grades. Similarly, suspensions are affected both by whether a teacher helps develop student behavior and by how harshly or leniently a teacher chooses to punish a student. These types of measurement error could lead to biased estimates of the effect of teacher value-added on student outcomes.

A related concern is that teachers could directly affect long-term outcomes without affecting a student's behavior or learning skills. For example, if a teacher is more likely than other teachers to recommend a student be held back, that student may be more likely to drop out of high school even if the teacher actually has no effect on the student's behavior or learning skills. This potential direct effect could bias the effect of teacher value-added on long-term outcomes in the direction of affecting long-term outcomes.

We remove bias from variation in teacher strictness (or leniency), and the direct effect of teachers on long-term student outcomes, by using the lead of the student achievement measure (i.e., achievement in year $t + 1$ rather than the measure of student achievement in year t):

$$S_{ij(t+1)} = \Gamma X_{ijt} + \varepsilon_{ijt} . \tag{4}$$

This approach introduces noise to our estimates, because it partially captures the effect of the teacher in year $t + 1$, but removes systematic bias from teachers evaluating their own students. In addition, using student achievement in year $t + 1$ makes it more difficult for teachers to manipulate their behavior or learning skills value-added. It does not eliminate all measurement problems, as there are a limited number of year $t + 1$ teachers evaluating each teacher's students and some measures have an inherent subjectivity that is not fixed using this approach. However, we show later in the paper that the GPA and learning skills measures, S_{ijt}, are highly correlated with students' long-term outcomes, and that GPA and learning skills value-added affect outcomes after year $t + 1$, which suggests that these variables measure meaningful skills.

In order for this approach to work, teachers must be able to exert some long-term influence on their students' behavior. One possible channel is by directly helping students develop noncognitive skills (e.g., Cunha, Heckman, and Schennach 2010; Heckman, Pinto, and Savelyev 2013; Kraft 2019). Teachers may also affect parents' behaviors and beliefs, for example, about the value of attending school or their child's performance in school, and induce parents to reduce absences or help their children with their homework.[7]

B. *Estimating the Long-Term Effects of Teacher Value-Added*

Once we have leave-year-out estimates of teacher quality, $\hat{v}_{jt}$, we explore how having either a higher- or a lower-quality teacher along some dimension of teacher quality affects a student in the long term. Let y_i be a long-term outcome of interest, such as whether a student is a high school dropout, an indicator for taking the SAT, or a score on a test required for high school graduation. Let k index the distinct leave-year-out value-added measures of test scores, behavior, or learning skills. Elementary school students have only one teacher per year, so we have only one measure of $\hat{v}_{jkt}$ per student-year. Let n_{ijt} be the number of classes student i has with a teacher j in academic year t. For a high school or middle school student i and subject $s \in \{\text{English, math}\}$, we construct $\hat{v}^s_{ikt} = (1/\Sigma^J_{j=1} n_{ijt})\Sigma^J_{j=1} n_{ij}\hat{v}^s_{ijkt}$, which is the mean value-added score weighted by the number of classes the student has with the teacher.

We regress outcome, y_i, on a number of value-added measures and our controls from equation (1), X_{ijt}, with the exception of variables that are not consistently available across years, including work and study habits GPA and learning and social skills GPA.[8] The estimates of $\widehat{\beta}_k$ for each value-added measure assess how each dimension of teacher quality affects the outcome of interest:

[7] For example, in California 42%–56% of elementary school parents were contacted by their school about attendance in a 6-month period (Ad Council 2015). Nationally, 90% of parents of third- to fifth-graders report attending parent-teacher conferences (McQuiggan and Megra 2017), and one of the topics LAUSD suggests teachers inform parents about is the benefit of better attendance (*LAUSD New Teacher Resource Guide*). There is also evidence from randomized controlled trials that communication from teachers or schools to parents substantially increases attendance (Kraft and Rogers 2015; Cook et al. 2017; Robinson et al. 2018; Rogers and Feller 2018)

[8] In particular, the controls include lags of a cubic polynomial of the student's math test score, English test score, GPA, and effort GPA; a cubic polynomial of lagged class- and grade-level means of each of those variables; current English-learner status, lagged log days absent, lagged suspensions, lagged being held back, and the class- and grade-level averages of these variables. Each of these variables, except English-learner status, is fully interacted with grade fixed effects, and a control for class size is included.

$$y_i = \sum_{k=1}^{K} \beta_k \hat{\nu}_{jkt} + \Gamma X_{ijt} + \eta_{ijt}. \quad (5)$$

We reduce the dimensionality of the estimates of teacher quality by constructing three indices of the value-added variables. The first index is computed from teacher math and English test score value-added, which we call the "test score value-added," or $\hat{\theta}^{s}_{jt}$. The second value-added index is computed from value-added for suspensions, log days absent, GPA, and not progressing to the next grade on time (i.e., being held back), which we call the "behavior value-added," or $\hat{\theta}^{b}_{jt}$. The third value-added index is computed from the value-added from effort GPA, work and study habits GPA, and learning and social skills GPA, which we call "learning skills value-added," or $\hat{\theta}^{l}_{jt}$. We chose these three groups because they separate test scores from non–test scores and because the behavior value-added includes variables that are available for all grades, whereas the learning skills value-added is available only for elementary school students. Our behavior value-added measures for middle and high school students include measures for both their math and English teachers. The indices are computed by summing the standardized value-added variables, recoded so that each has the same expected sign, and then standardizing the resulting index to be mean 0 and standard deviation 1. In the appendix, we show that the main results are robust to grouping GPA with learning skills, using factor analysis to construct the three indices, and using exploratory factor analysis to choose the factors and the factor load on each value-added measure.

We estimate the long-term effect of these value-added measures, using the following specification:

$$y_i = \beta^{s}\hat{\theta}^{s}_{jt} + \beta^{b}\hat{\theta}^{b}_{jt} + \beta^{l}\hat{\theta}^{l}_{jt} + \Gamma X_{ijt} + \eta_{ijt}, \quad (6)$$

where X_{ijt} is the vector of baseline controls used in equation (5). We also compare the estimates from equation (6) with the estimates from a model that omits non–test score value-added indices. This comparison allows us to sign the bias from omitting non–test score measures in papers that estimate the effect of teachers' test score value-added on long-term outcomes. If we find that $\widehat{\beta^{s}}$ falls when we move from a model that excludes $\hat{\theta}^{b}_{jt}$ and $\hat{\theta}^{l}_{jt}$ to one that includes them, it suggests that typical estimates of the long-term effects of test score value-added are biased upward by omitted measures of behavioral or noncognitive skills. Alternatively, if $\hat{\theta}^{b}_{jt}$ or $\hat{\theta}^{l}_{jt}$ affects long-term outcomes and the estimate of $\widehat{\beta^{s}}$ is unaffected by adding $\hat{\theta}^{b}_{jt}$ or $\hat{\theta}^{l}_{jt}$, the long-term effects of test score value-added may be unbiased, but estimates of the total effect of teachers on students is larger than the effects found when using test score value-added alone.[9] We cluster

[9] We use the modal high school that students in each elementary school progress to, so all students in a given elementary school are in the same cluster.

the standard errors at the high school level. As shown in table A.1, the standard errors not substantially affected by bootstrapping over the estimation of the value-added scores described in section IV.A and the estimation of the long-term effects using equation (6).

Let tildes denote residualized student value-added indices; for example, $\tilde{\theta}^{s}_{jt} = \hat{\theta}^{s}_{jt} - \hat{\Gamma} X_{ijt} - \widehat{\beta^{b}}\hat{\theta}^{b}_{jt} - \widehat{\beta^{l}}\hat{\theta}^{l}_{jt}$. To interpret the estimates in equation (6) as causal, we must assume $\mathrm{Cov}(\tilde{\theta}^{s}_{jt}, \eta_{ijt}) = \mathrm{Cov}(\tilde{\theta}^{b}_{jt}, \eta_{ijt}) = \mathrm{Cov}(\tilde{\theta}^{l}_{jt}, \eta_{ijt}) = 0$; the residualized leave-year-out predicted teacher value-added indices and student unobservables that affect the outcome, y_i, are uncorrelated. Although this is a strong assumption, sorting based on teacher characteristics that are uncorrelated with residualized teacher value-added (e.g., master's degree attainment) will not bias our estimates (Chetty, Friedman, and Rockoff 2011). To help alleviate some of the concerns with this assumption, in section V.E we examine the effect of teacher value-added on predicted outcomes as a placebo test and perform a quasi-experimental analysis that uses teachers switching grades and schools.

In addition, there are reasons to believe that this approach is conservative. First, we find somewhat larger, although much less precisely estimated, effects using a quasi-experimental design that uses variation in teachers switching between grades and schools. Second, we estimate smaller effects than if we use the approach taken by Chetty, Friedman, and Rockoff (2014b).

We also extend this analysis in two ways. First, we examine the dynamic effects of a teacher on student outcomes for years $\tau \in [0, 1, .., 7]$:

$$y_{i(t+\tau)} = \beta^{s}\hat{\theta}^{s}_{jt} + \beta^{b}\hat{\theta}^{b}_{jt} + \beta^{l}\hat{\theta}^{l}_{jt} + \Gamma X_{ijt} + \eta_{ijt}. \quad (7)$$

The model shows the extent to which the effect of teacher value-added on student outcomes persists or fades over a number of years. Second, we assess the effects on high school outcomes by grade to see in which grades high-quality teachers have the most impact on students in high school.

V. Results

A. *Descriptive Results*

1. Descriptive Relationships in Student Data

To assess whether multiple dimensions of teacher quality might matter for long-term outcomes, we estimate the relationship between measures of student achievement, both with each other and with long-term outcomes. Table A.2 shows bivariate correlations between each of the measures of student achievement. English and math test scores are highly correlated. The relationships between test scores and students' GPA,

learning skills GPA, and effort GPA are weaker, but the correlation still ranges from 0.45 to 0.68. The correlations of these variables with attendance, days suspended, and being held back are substantially weaker, which suggests that these variables largely capture different aspects of student achievement. These correlations suggest that test scores, behavior, and learning skills are related but that some room remains for them to have an independent effect on long-term outcomes. Reducing the dimensionality of these variables by separately computing test score, behavior (i.e., attendance, days suspended, being held back, and GPA), and learning skills (i.e., learning skills GPA and effort GPA) indices, as described in section IV.B, yields correlations between 0.46 and 0.55 (table 2).

Next, we assess whether these measures of student achievement are related to long-term outcomes, conditional on the same set of controls we use to compute the value-added measures. English and math test scores, GPA, learning skills GPA, suspensions, log days absent, and being held back in grades 3–5 typically have a statistically significant relationship with high school outcomes (table A.3). After reducing the dimensionality of these measures to three indices of student achievement—test scores, behavior, and learning skills—we find that student achievement in grades 3–5 nearly always has a statistically significant effect on high school outcomes (table 3). For many of the high school outcomes, behavior and learning skills are as predictive of the outcome as test scores.

These results are consistent with test scores, behavior, and learning skills each independently affecting long-term outcomes. However, despite the fact that we control for a wide range of measures of student achievement, these estimates may be biased because of unobservables. Consequently, these results may not hold in situations in which there is exogenous variation in students' test score, behavior, and learning skills achievement. To address this concern, we move to a teacher value-added framework in which omitted variables are less likely to bias the results.

TABLE 2
CORRELATION OF ELEMENTARY SCHOOL STUDENT ACHIEVEMENT MEASURES

Measure	Test Scores	Behavior	Learning Skills
Test scores	1		
Behavior	.463	1	
Learning skills	.532	.550	1

NOTE.—The sample includes students in grades 3–5 who attended the LAUSD. The unit of observation is a student–academic year. This table reports the correlations between the three measures of student achievement for grades 3–5. Each of the three measures of student achievement are equally weighted indices. The test score index is computed from the students' normalized math and English test scores. The behavior index is computed from students' GPA, suspensions, log days absent, and not progressing to the next grade on time (held back). The learning skills index is computed from students' effort GPA, work and study habits GPA, and learning and social skills GPA.

TABLE 3

Relationship between Elementary School Achievement and High School Outcomes

Measure	LAUSD Dropout	Took SAT	SAT Score	GPA	Effort GPA	Cooperation GPA	Math CAHSEE	English CAHSEE	Days Suspended	Log Absences	Held Back
Test scores	−.032***	.091***	143.271***	.208***	.103***	.089***	.472***	.398***	−.009***	−.112***	−.052***
	(.003)	(.004)	(2.644)	(.004)	(.002)	(.002)	(.006)	(.006)	(.002)	(.005)	(.003)
Behavior	−.020***	.028***	10.826***	.062***	.032***	.030***	.028***	.020***	−.028***	−.130***	−.017***
	(.002)	(.003)	(1.554)	(.004)	(.002)	(.002)	(.003)	(.003)	(.003)	(.006)	(.002)
Learning skills	−.040***	.068***	−2.393*	.228***	.133***	.121***	.045***	.064***	−.045***	−.062***	−.061***
	(.002)	(.003)	(1.290)	(.003)	(.002)	(.002)	(.003)	(.003)	(.003)	(.004)	(.002)
Observations	134,356	100,691	59,582	321,236	251,460	251,460	160,385	159,911	343,720	302,902	240,204
R^2	.421	.179	.686	.321	.315	.326	.572	.577	.045	.281	.134

Note.—The sample includes students in grades 3–5 who attended high school in the LAUSD. The unit of observation is a student–academic year. This table reports the predictive effect of a standard-deviation increase in each of the three measures of student achievement (see table 2) in grades 3–5 on high school student outcomes. Specifically, each column of the table reports the coefficients on each of the three achievement measures of students from an OLS regression of the high school student outcome on the students' three achievement measures in grades 3, 4, or 5, along with the baseline controls described in sec. IV.A. The baseline controls include lags of a cubic polynomial of the student's math test score, English test score, GPA, and effort GPA; a cubic polynomial of lagged class- and grade-level means of each of those variables; current English-learner status, lagged log days absent, lagged suspensions, lagged being held back, and the class- and grade-level means of these variables. Each of these variables, except English-learner status, is fully interacted with grade fixed effects, and a control for class size is included. Standard errors clustered at the modal–high school level are reported in parentheses.

* $p < .10$.

*** $p < .01$.

2. Descriptive Relationships in Teacher Value-Added Data

We compute teacher value-added as described in section IV.A. Table A.4 shows the relationship between the value-added measures. English and math test score value-added measures are highly correlated. The correlations between test score value-added and all other variables are much weaker, but test scores are positively correlated with GPA, effort GPA, and learning skills GPA, which have correlations between 0.14 and 0.20. The GPA, effort GPA, and learning skills GPA value-added are highly correlated with one another. The three value-added measures of student behavior—log absences, days suspended, and being held back—are all weakly correlated with one another, test scores, and GPA measures. These correlations suggest that math and English test score measures of teacher quality are closely related, as are GPA-based measures of teacher quality, whereas the ability to influence student behavior relates less closely. Table 4 shows similar results. The correlation between test score value-added and behavior value-added is 0.15, the correlation of test score value-added with learning skills value-added is 0.17, and the correlation of behavior value-added with learning skills value-added is 0.46.

B. Effects of Teacher Quality on Long-Term Outcomes

1. Single Value-Added Effects

Figures 2–4 show the effect of teachers' test score, behavior, and learning skills value-added individually on each high school outcome, conditional on the set of controls used to compute value-added measures (also see table A.5 for the results in this subsection reported in a regression table). The plotted points show the relationship between the mean residualized outcome and the mean residualized value-added variables (with the

TABLE 4
CORRELATION OF ELEMENTARY SCHOOL TEACHER VALUE-ADDED (VA) MEASURES

Grades 3–5 VA	Test Score VA	Behavior VA	Learning Skills VA
Test score VA	1		
Behavior VA	.145	1	
Learning skills VA	.174	.459	1

NOTE.—The sample includes students in grades 3–5 who attended the LAUSD. The unit of observation is a student–academic year. This table reports the correlations between the three measures of teacher VA for grades 3–5. Each of the three measures are equally weighted indices. The test score VA is computed from teachers' VA for math and English test scores. The behavior VA is computed from teachers' VA for GPA, suspensions, log days absent, and not progressing to the next grade on time (held back). The learning skills VA is computed from teachers' VA for effort GPA, work and study habits GPA, and learning and social skills GPA.

unconditional mean of the outcome and value-added variables added back in) for 20 equally sized bins of teacher value-added measures. The coefficients and standard errors reported in the figures are from an OLS (ordinary least squares) regression, using the micro data, of the outcome variable on the value-added variable, conditional on the same set of controls.

Figure 2 shows that students with better teachers in grades 3–5, as measured by the test score value-added, score significantly higher on the SAT, have marginally significantly higher effort GPAs and significantly higher

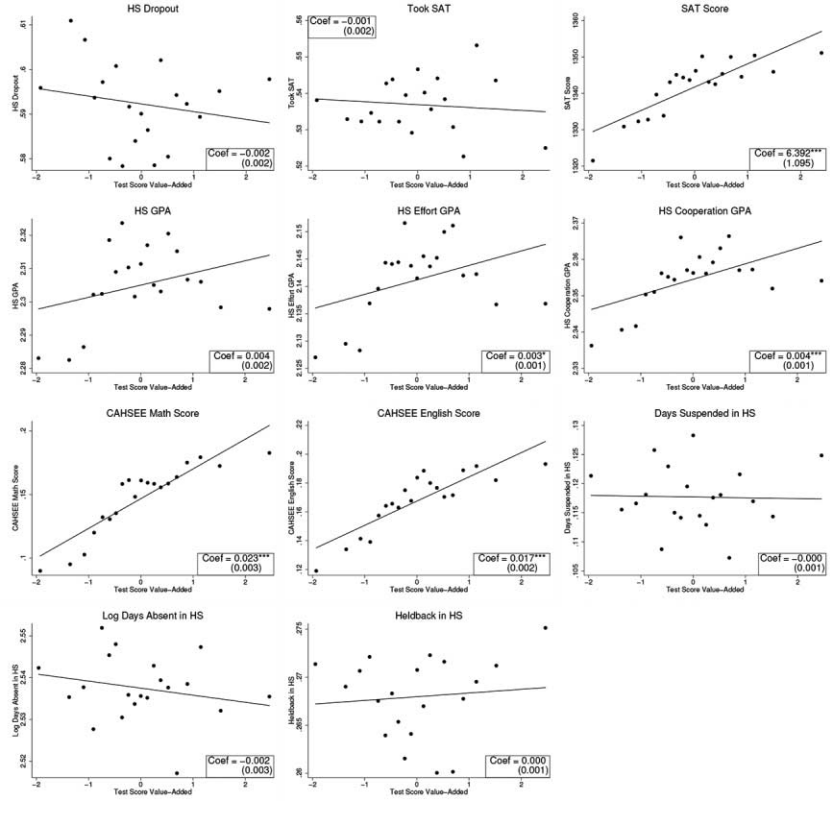

FIG. 2.—Effect of teacher test score value-added on high school (HS) outcomes. The sample includes students in grades 3–5 who attended high school in the LAUSD. The unit of observation is a student–academic year. This figure shows binned scatter plots of residualized high school outcome variables and normalized teacher test score value-added for grades 3, 4, or 5. We construct these plots by first residualizing the outcome and teacher value-added variables using the X_{ijt} controls shown in equation (6). We then plot the mean values of both variables in 20 equally sized bins. Finally, we add back the unconditional mean of both variables. We also plot the best linear fit estimated before binning the data and report its slope coefficient and standard error, clustered at the modal–high school level. $^{*}p < .10$; $^{***}p < .01$.

cooperation GPAs, and score significantly higher on the CAHSEEs. We find no significant effects on dropping out of high school, taking the SAT, GPA, being held back, log days absent, or being suspended. These results are consistent with the existing literature, which shows benefits in adulthood from higher test score value-added teachers, although research that demonstrates positive effects of elementary school teachers on high school outcomes is rare (Rothstein 2017).

Figure 3 shows the effect of teachers' behavior value-added on each outcome. We observe at least marginally statistically significant effects

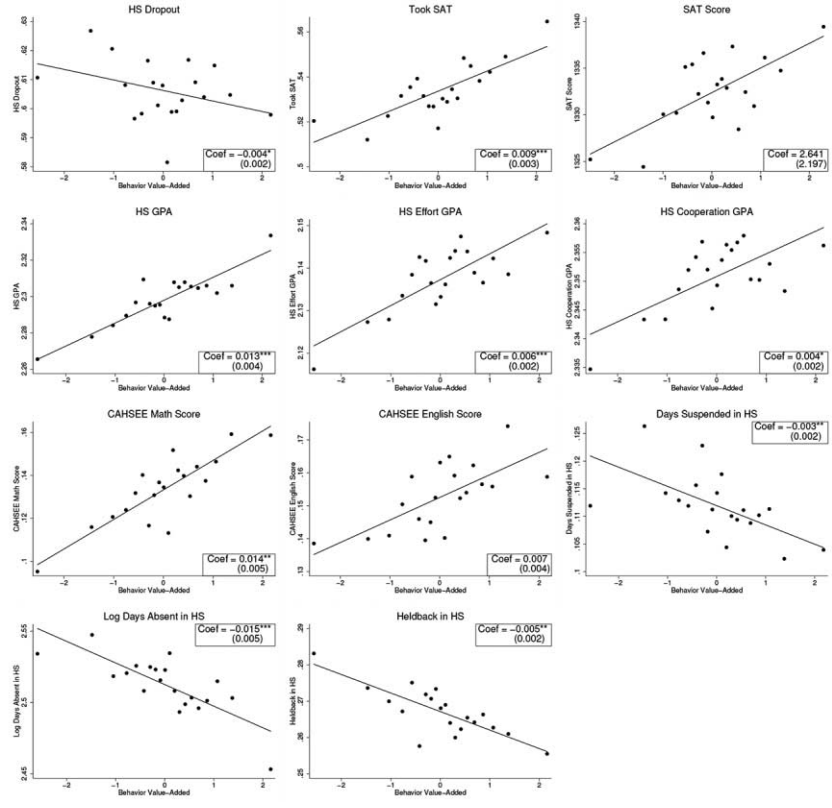

FIG. 3.—Effect of teacher behavior value-added on high school (HS) outcomes. The sample includes students in grades 3–5 who attended high school in the LAUSD. The unit of observation is a student–academic year. This figure shows binned scatter plots of residualized high school outcome variables and normalized teacher behavior value-added for grades 3, 4, or 5. We construct these plots by first residualizing the outcome and teacher value-added variables using the X_{ijt} controls shown in equation (6). We then plot the mean values of both variables in 20 equally sized bins. Finally, we add back the unconditional mean of both variables. We also plot the best linear fit estimated before binning the data and report its slope coefficient and standard error, clustered at the modal–high school level. $^{*}p < .10$; $^{**}p < .05$; $^{***}p < .01$.

in the expected direction on all the outcome variables except SAT score and English CAHSEE score. This indicates that, in the absence of test score value-added, having a teacher with a higher behavior valued-added affects the high school outcomes of students in a meaningful way. Figure 4 shows the effect of teachers' learning skills value-added on each outcome. We find less evidence of an effect than for the other two value-added measures. The coefficient on the learning skills value-added typically has the expected sign, but the only marginally significant effect is on days suspended. The confidence intervals are sufficiently small to reject moderate effects,

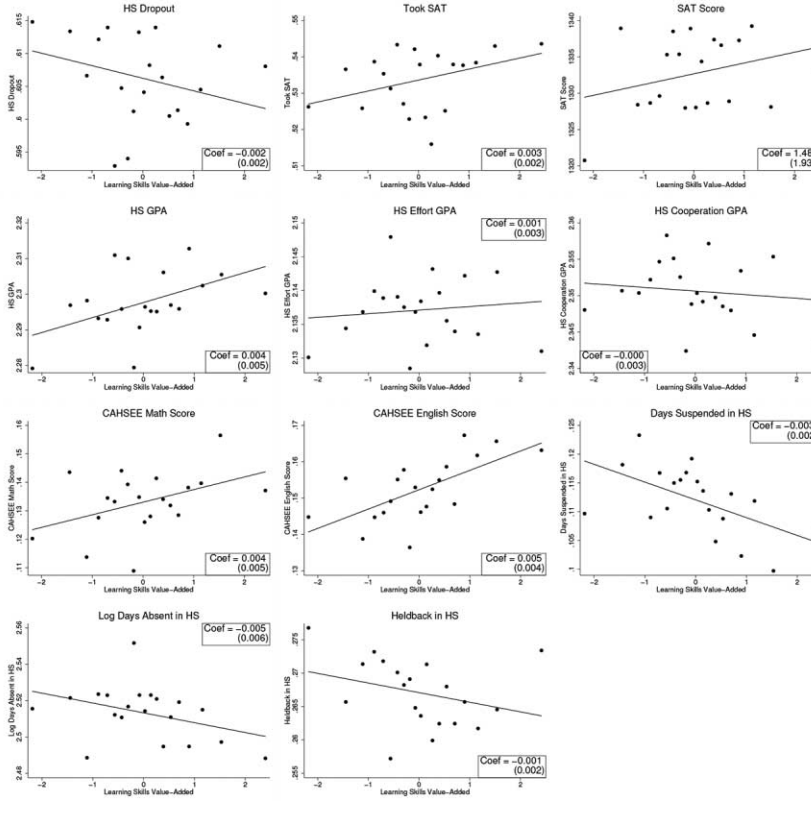

FIG. 4.—Effect of teacher learning skills value-added on high school (HS) outcomes. The sample includes students in grades 3–5 who attended high school in the LAUSD. The unit of observation is a student–academic year. This figure shows binned scatter plots of residualized high school outcome variables and normalized teacher learning skills value-added for grades 3, 4, or 5. We construct these plots by first residualizing the outcome and teacher value-added variables using the X_{ijt} controls shown in equation (6). We then plot the mean values of both variables in 20 equally sized bins. Finally, we add back the unconditional mean of both variables. We also plot the best linear fit estimated before binning the data and report its slope coefficient and standard error, clustered at the modal–high school level. $^{*}p < .10$.

but for several of the outcomes, we cannot reject effects of the same size as the behavior value-added estimates. These results suggest that elementary school teachers affect students' long-term outcomes by increasing student achievement as measured by both test score and non–test score data.

Comparing the magnitudes across the analyses, we tend to find that test score value-added has a large effect on outcomes that have more cognitive content than the behavior or learning skills value-added, whereas the pattern of results is reversed for outcomes that have more noncognitive content. For example, having a teacher with a standard deviation higher test score value-added increases the math CAHSEE scores by 0.023 standard deviations, whereas the increase for behavior value-added is 0.014 standard deviations, and the statistically insignificant increase for the learning skills value-added is 0.004 standard deviations. However, the effect of having a teacher with a standard deviation higher test score value-added on days suspended is less than 0.001 days, whereas behavior and learning skills value-added both reduce days suspended by 0.003, a 2% decrease.

The test score value-added estimates appear to have two sets of potential nonlinear effects. First, the effect of test score value-added on all three GPA measures is positive until teachers become above average, and then the relationship is, if anything, negative. Second, there is suggestive evidence that the top ventile or two of the test score value-added distribution has a smaller effect on several outcomes than would be predicted from the rest of the test score value-added distribution (fig. 2). Chetty, Friedman, and Rockoff (2014b) find a similar anomaly in their fourth- to eighth-grade test score value-added results and drop the top 1% of teachers because of evidence of "test manipulation." We leave those teachers in, although including them biases the effects of test score value-added toward zero if "test manipulation" exists. We find less evidence of nonmonotonicities for both behavior and learning skills value-added, and outliers in the top ventile are less common. One explanation for this finding is that, because non–test score value-added measures are constructed from student achievement in year $t + 1$, teachers are unable to manipulate their non–test score value-added measures unless they influence the actions of their students' teachers in the subsequent year. Supporting this explanation, we find that when figure 2 is recreated with value-added measures using test scores in year $t + 1$ instead of year t (fig. A.3), these nonlinearities no longer exist.

Taken together, these results suggest that multiple components of teacher quality affect long-term outcomes. Our findings also indicate that, in situations in which no test score data are available but other administrative data such as grades, attendance, suspensions, and being held back are available, creating estimates of teacher quality that are associated with long-term benefits to students is possible. Some evidence also suggests

that non–test score value-added measures calculated with our approach are less prone to manipulation by teachers, although they might begin to be manipulated if used in high-stakes settings.

2. Multivariate Value-Added Effects

Now that we have found that two of the dimensions of teacher quality affect high school outcomes, we can determine whether more than one value-added measure independently affect long-term outcomes. Significant effects of more than one value-added measure would suggest that teacher quality is multidimensional in a way that both matters for long-term outcomes and is measurable with a value-added approach. In addition, this analysis informs the extent to which the long-term effects of test score value-added measures are driven by teachers' effect on behavior and learning skills.

Table 5 shows the effect that each of the three elementary school value-added measures has on high school outcomes in an OLS regression in which all three value-added measures are included simultaneously, along with the baseline controls (eq. [6]). Including behavior and learning skills value-added only slightly affects the coefficients on the test score value-added measures. For example, the coefficient in the SAT score regression falls from 6.39 to 6.24 SAT points (or is a constant 0.021 standard deviations), the coefficient in the math CAHSEE regression falls from 0.023 to 0.022 standard deviations, and the coefficient in the high school GPA regression falls from 0.004 to 0.002 GPA points. These results indicate that the long-term effects of test score value-added are likely not driven by teachers' effects on students' behavior and learning skills that are correlated with test score value-added.

The effects of behavior value-added on most outcomes are also not affected substantially by conditioning on the test score and learning skills value-added. Behavior value-added picks up a dimension of teacher quality that is largely unrelated to the other two value-added measures and that matters for long-term outcomes. In addition, table A.6 shows that there is no evidence of an interaction effect of elementary school teachers' test score and behavior value-added on students' high school outcomes, with coefficient estimates suggesting that any interaction effect is very close to zero.

Adding the other value-added measures does not affect the evidence for an independent effect of teachers on long-term outcomes through learning skills. None of the coefficients on the learning skills value-added in table 5 are statistically significant. For some of these outcomes, conditioning on test score and behavior value-added results in the learning skills value-added coefficients no longer having the expected sign. The confidence intervals often, but not always, exclude effects of the magnitude of the coefficients on behavior value-added.

TABLE 5
EFFECT OF ELEMENTARY SCHOOL TEACHER VALUE-ADDED (VA) ON HIGH SCHOOL OUTCOMES

Pooled Grades 3–5 VA	LAUSD Dropout	Took SAT	SAT Score	GPA	Effort GPA	Cooperation GPA	Math CAHSEE	English CAHSEE	Days Suspended	Log Absences	Held Back
Test score VA	−.002	−.002	6.237***	.002	.003*	.005***	.022***	.016***	.001	.001	.001
	(.002)	(.002)	(1.262)	(.002)	(.001)	(.001)	(.003)	(.002)	(.001)	(.003)	(.001)
Behavior VA	−.003	.010***	1.955	.013***	.007***	.005**	.013**	.004	−.003	−.016***	−.006**
	(.002)	(.003)	(2.139)	(.004)	(.003)	(.002)	(.006)	(.005)	(.002)	(.005)	(.003)
Learning skills VA	−.000	−.001	−.547	−.002	−.003	−.003	−.005	.001	−.002	.002	.001
	(.002)	(.003)	(1.756)	(.005)	(.003)	(.003)	(.005)	(.004)	(.002)	(.006)	(.002)
Observations	135,786	102,517	60,694	293,021	233,078	233,078	152,345	151,820	316,116	277,331	221,757
R^2	.293	.145	.617	.244	.234	.239	.500	.512	.040	.267	.108

NOTE.—The sample includes students in grades 3–5 who attended high school in the LAUSD. The unit of observation is a student–academic year. This table reports the effect of a standard-deviation increase in the three measures of elementary school teacher VA (see table 4) on students' high school outcomes. Specifically, each column of the table reports the coefficients on each of the three normalized measures of teacher VA from an OLS regression of the students' high school outcome on the three measures of teacher VA for the students, teachers in grades 3, 4, or 5, along with the baseline controls described in sec. IV.A. The baseline controls include lags of a cubic polynomial of the student's math test score, English test score, GPA, and effort GPA; a cubic polynomial of lagged class- and grade-level means of each of those variables; current English-learner status, lagged log days absent, lagged suspensions, lagged being held back, and the class- and grade-level means of these variables. Each of these variables, except English-learner status, is fully interacted with grade fixed effects, and a control for class size is included. Standard errors clustered at the modal–high school level are reported in parentheses.

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

Figure 5 shows how the value-added measures affect a number of outcomes that can be tracked over time, beginning in elementary school. The effect of test score value-added on test scores shows the expected pattern of results. Having a teacher with a standard deviation higher test score value-added has a large effect on math and English test scores in year zero that largely—but not completely—fades out over the next 7 years. Behavior value-added and learning skills value-added show less evidence of fade-out, but our approach to constructing these variables should result in measures with less fade-out than test score value-added. By measuring behavior and learning skills value-added using the effect of a teacher this year on student achievement in the next year, we are effectively removing the first year of fade-out from the estimates. In addition, because some of the student achievement variables are grades and students may be graded on a curve, seeing little effect of behavior and learning skills value-added on GPA measures in year zero would not be surprising.

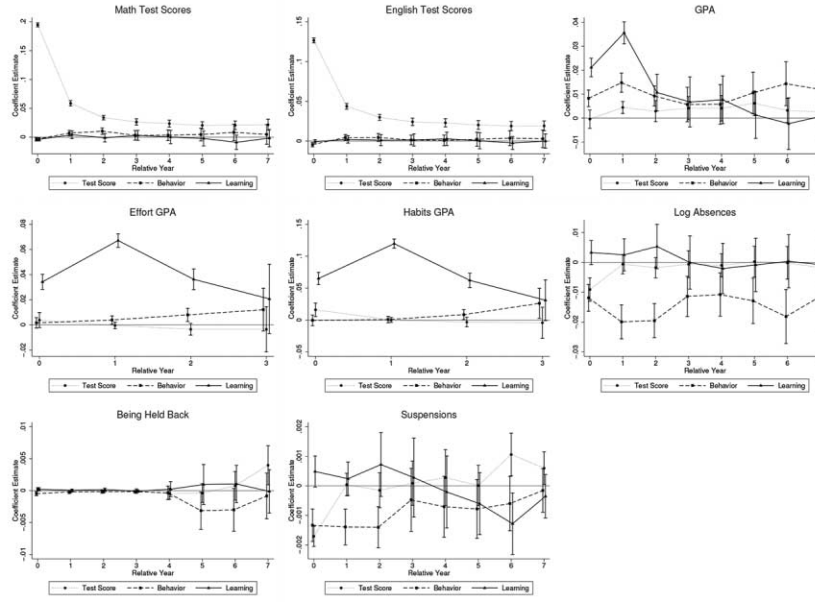

FIG. 5.—Dynamic effects of test score, behavior, and learning skills teacher value-added. The unit of observation is a student–academic year. Each plot shows the effect of test score, behavior, and learning skills value-added of teachers in grades 3, 4, or 5 on student outcomes in the concurrent year (the year a student was in a teacher's classroom) and future years (the years after a student was in a teacher's classroom). The estimated effects are obtained by regressing leads of outcome variables on teacher test score, behavior, and learning skills value-added and the baseline controls as specified in equation (7). The coefficients on test score, behavior, and learning skills value-added are plotted, along with 95% confidence intervals, with standard errors clustered at the modal–high school level.

The size of the long-term results can be interpreted using the cross-sectional relationship between test scores and earnings. Hanushek and Woessmann (2008) find consistent evidence that, in the cross section, a 1–standard deviation increase in test scores at the end of high school increases earnings by 12% (Mulligan 1999; Murnane et al. 2000; Lazear 2003). Chetty, Friedman, and Rockoff (2014b) find a similar relationship in the cross section for fourth- to eighth-graders. They also show that, with direct estimates, a 1–standard deviation increase in test score value-added increases earnings by approximately 1.3%. When using the cross-sectional relationship, they estimate that a 1–standard deviation increase in test score value-added increases earnings by a similar amount, 1.5%. Using this cross-sectional relationship and our estimated effect of teachers on contemporaneous test scores, we estimate effects on earnings of approximately the same size as or larger than those of Chetty, Friedman, and Rockoff (2014b). If we instead use the effect on high school test scores of having a teacher with a standard deviation higher test score value-added in elementary school, the estimated increase in earnings is 0.23%. This much smaller effect is likely driven by the substantial fade-out in the effect of teachers on test scores over time.

Combined, these results indicate that teacher quality is multidimensional in a way that matters for long-term outcomes. Importantly, this multidimensionality can be measured with a combination of test scores and other data that schools routinely collect.

C. Teacher Selection Policies

Policies that use teachers' test score value-added to hire, fire, or incentivize teachers have been widely criticized because making decisions using only one (potentially gameable) dimension of teacher quality is considered unfair or even counterproductive. However, the effect on long-term outcomes of having higher test score and behavior value-added teachers implies that policies that shift the distribution of teacher quality upward in these dimensions benefit students. In comparison to just using test score value-added, we show that using multiple dimensions of teacher quality in teacher removal policies substantially improves the measurement of teacher quality and students' long-term outcomes.

Figure 6 shows scatter plots of teacher quality as measured by value-added in a given year. The dashed lines show the 5th percentile of teachers for a given value-added measure. The first panel plots test score and behavior value-added and shows that, although both dimensions of teacher quality are positively correlated, the correlation is relatively weak, and some teachers who perform poorly as measured by test score value-added perform well on the behavior value-added dimension. For example, the majority of teachers who are in the bottom 5% of teachers as

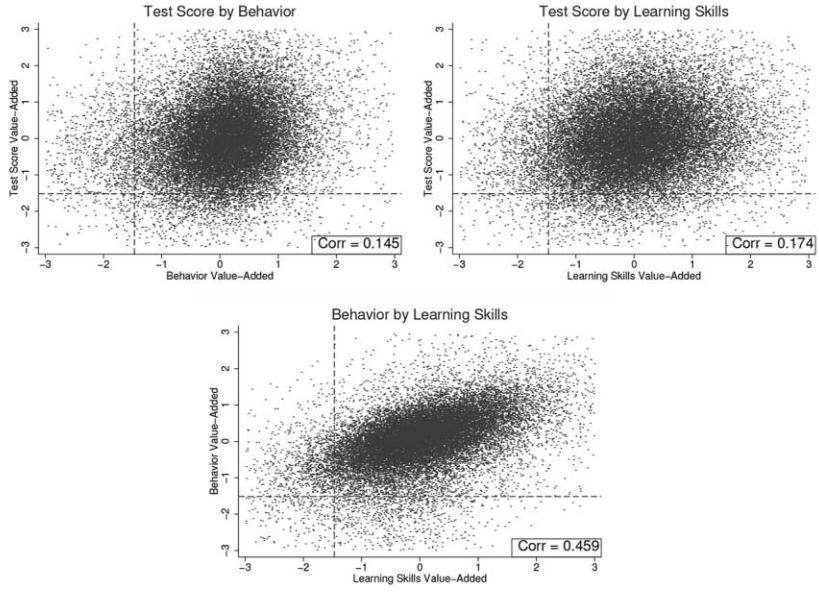

FIG. 6.—Two-dimensional cross-teacher value-added plots. The first scatter plot shows a plot of elementary school teachers' annual, normalized test score and behavior value-added within 3 standard deviations of the mean. The dashed lines show the cutoffs for the 5th percentile of the teacher test score and behavior value-added. The second and third scatter plots are constructed analogously for test score value-added versus learning skills value-added and behavior value-added versus learning skills value-added, respectively.

measured by the test score value-added are not in the bottom 5% of teachers as measured by the behavior value-added. Therefore, a linear combination of a teacher's value-added measures might be a better predictor of teacher quality and a better measure for teacher removal policies.

One way to assess the value of using multiple measures of teacher quality is to ask to what extent students' long-term outcomes could be improved under a policy that replaces a school district's bottom 5% of teachers with average teachers as measured by only test score value-added versus different linear combinations of the three value-added measures. Panel A of table 6 shows the effect on a student's high school outcomes of being assigned an average teacher instead of a teacher in the bottom 5% of teachers as measured by a teacher's true value-added (realized value-added ex post, $\bar{\nu}_{jt}$). This panel shows the upper bound on the effects of the teacher removal policy. The simulation uses estimated effects of teacher value-added on high school outcomes (figs. 2–4; table 5) and the within-teacher correlations between the three teacher value-added measures (table 4). Standard errors for the estimated forecasts are shown in parentheses.

Each cell in row 1 shows the effect on students' high school outcomes if their bottom 5% test score value-added elementary school teachers were replaced by average teachers. For example, the students whom the policy would affect (about 5%) would see their SAT scores increase by 13 points and their high school GPA increase by 0.008 points. Row 2 shows the effect on students' high school outcomes if their bottom 5% behavior value-added elementary school teachers were replaced by average teachers. The benefits from using behavior value-added are comparable to those from using test score value-added, and in some cases, the benefits are larger.

Row 3 uses the average of teachers' test score and behavior value-added. When this combined measure is used, students affected by the policy would see beneficial effects on all but one of the high school outcomes. Row 5 shows the percent change in students' outcomes if the replacement of the bottom 5% of teachers uses the average of teachers' test score and behavior value-added instead of only teachers' test score value-added. There is over a 100% increase in the beneficial effects for students measured by high school dropout rate, taking the SAT, GPA, effort GPA, days suspended, log absences, and being held back. Importantly, these gains are accompanied by only small decreases in English CAHSEE scores and SAT scores.

Row 4 uses a maximization procedure to choose the optimal weights to be placed on a linear combination of teachers' test score, behavior, and learning skills value-added to determine the bottom 5% of teachers for the indicated outcome variable. The optimal weights vary, depending on the outcome variable, so simultaneously improving all outcomes by the calculated amount would not be possible. However, for most of the outcomes, a policy that uses the optimal weights for a particular outcome only slightly outperforms a simple policy that places equal weight on test score and behavior value-added.

Panel B of table 6 shows analogous results using teachers' estimated value-added based on the three previous years of student data. These results reflect the potential student gains if the teacher removal policy were to be implemented for teachers who had taught for 3 years. Similar to panel A, student gains can be obtained if both test score and behavior value-added are used to make the teacher removal decision. Because the autocorrelation between years for the behavior measure is smaller than that for the test score measure (fig. A.4), the percent gain from using both value-added measures instead of just the test score value-added is smaller.

These results suggest that the dimensions of teacher quality captured by behavior value-added are roughly as important for long-term outcomes as test score value-added and, in combination, can benefit students. Most of these benefits do not require new tests or assessments—only a new use for data that schools already collect.

TABLE 6
EFFECT OF REPLACING THE BOTTOM 5% OF ELEMENTARY SCHOOL TEACHERS ON HIGH SCHOOL OUTCOMES

Pooled Grades 3–5 VA	LAUSD Dropout	Took SAT	SAT Score	GPA	Effort GPA	Cooperation GPA	Math CAHSEE	English CAHSEE	Days Suspended	Log Absences	Held Back
	A. Using True VA										
1. Test score VA	−.004	−.002	13.176***	.008*	.005*	.009***	.048***	.035***	−.000	−.003	.001
	(.003)	(.004)	(2.257)	(.005)	(.003)	(.003)	(.005)	(.005)	(.003)	(.005)	(.002)
2. Behavior VA	−.007*	.019***	5.446	.026***	.013***	.008*	.028**	.014	−.007**	−.031***	−.011**
	(.004)	(.006)	(4.531)	(.008)	(.005)	(.004)	(.011)	(.009)	(.003)	(.010)	(.005)
3. (1/2) (test score + behavior)	−.008*	.012**	12.280***	.023***	.012***	.012***	.050***	.033***	−.004	−.022**	−.007
	(.004)	(.005)	(3.965)	(.007)	(.004)	(.004)	(.009)	(.008)	(.003)	(.009)	(.004)
4. Optimal three VA	−.008	.019***	13.733***	.027***	.014***	.013***	.053***	.036***	−.008	−.031***	−.011*
5. Weighted average	(.006)	(.007)	(3.144)	(.009)	(.005)	(.004)	(.008)	(.007)	(.005)	(.011)	(.005)
6. % Gain (row 1–row 3)	121	200+	−7	200	129	34	4	−6	200+	200+	200+
	B: Using Previous 3 Years of Student Data										
1. Test score VA	−.002	−.001	7.688***	.004*	.003*	.005***	.028***	.020***	−.000	−.002	.000
	(.002)	(.002)	(1.317)	(.003)	(.002)	(.001)	(.003)	(.003)	(.002)	(.003)	(.001)
2. Behavior VA	−.001*	.003***	.968	.005***	.002***	.001*	.005**	.003	−.001**	−.006***	−.002**
	(.001)	(.001)	(.805)	(.001)	(.001)	(.001)	(.002)	(.002)	(.001)	(.002)	(.001)
3. (1/2) (test score + behavior)	−.003	.001	6.463***	.006**	.004***	.005***	.024***	.017***	−.000	−.004	−.001
	(.002)	(.002)	(1.371)	(.003)	(.002)	(.001)	(.003)	(.003)	(.001)	(.003)	(.002)
4. Optimal three VA	−.003	.003***	7.688***	.006**	.004***	.005***	.028***	.020***	−.002	−.006***	−.002**
5. Weighted average	(.002)	(.001)	(1.317)	(.003)	(.002)	(.001)	(.003)	(.003)	(.001)	(.002)	(.001)
6. % Gain (row 1–row 3)	29	200+	−16	36	24	−0	−15	−15	142	95	200+

NOTE.—The sample includes students in grades 3–5 who attended high school in the LAUSD. The unit of observation is a student–academic year. Panel A shows the effect on a student's high school outcomes of being assigned an average teacher instead of a teacher in the bottom 5%, as measured by the teacher value-added (VA) variable indicated in each row. The simulation uses the estimated effects of teacher VA measures on high school outcomes (from figs. 2–4 and table 5) and the within-teacher correlations between the three teacher VA measures (shown in table 4) and assumes that teacher VA measures are normally distributed. Row 1 in panel A uses a measure of teachers' test score VA to replace the bottom 5% of teachers. Therefore, each cell in row 1 shows the effect on a student's high school outcome (shown in the column) if she were to move from a teacher in the bottom 5% of test score teacher VA to an average teacher. Row 2 of panel A shows the improvement in outcomes for a move from a teacher in the bottom 5% to an average teacher, as measured by behavior VA. Row 3 of panel A uses the average of teachers' test score VA and their behavior VA. Row 4 of panel A uses a maximization procedure to choose the optimal weights to be placed on teachers' test score, behavior, and learning skills VA to determine the bottom 5% of teachers for the indicated outcome variable. Row 5 of panel A shows the percent improvement in students' outcomes if the replacement of the bottom 5% of teachers uses the average of teachers' test score and behavior VA instead of just teachers' test score VA. Panel B shows analogous results for teachers' VA based on only the three previous years of student data. These VA measures are estimated from three prior years of data on each teacher, along with the autocorrelations in teachers' VA across years shown in fig. A.4. The standard errors on the estimated forecast are shown in parentheses.

* $p < .10$.

** $p < .05$.

*** $p < .01$.

D. Which Behaviors Matter for Long-Term Outcomes?

Behavior value-added includes several weakly correlated value-added measures, some of which may matter more than others for long-term outcomes. A straightforward way to assess which variables matter most is to regress high school outcomes on the full set of value-added measures that we use to construct the lower-dimensional representation of teacher quality plus the usual set of controls.

We regress each outcome on each component of test score, behavior, and learning skills value-added, conditional on our baseline controls, so each cell in table A.7 is from a separate regression. The test score value-added measures have significant effects on SAT taking, the GPA measures, and CAHSEEs. GPA value-added has significant effects on SAT taking and the English CAHSEE and an effect similar in magnitude but insignificant on the math CAHSEE. Two other components of behavior value-added, absence and suspension value-added, have significant effects on outcomes with both cognitive and noncognitive content. There is less evidence of an effect of held-back value-added, but the held-back confidence interval generally cannot reject effects of the magnitude of the coefficients on absence or suspension value-added. We generally do not see evidence of an effect of the various components of learning skills value-added. The coefficients typically have the expected sign, but the only significant effect is of effort GPA on SAT taking. These estimates are somewhat less precise than the test score value-added results, and we often cannot reject effects of the magnitude of the test score effects, although the estimates tend to be slightly more precise than the coefficients on absences and days-suspended value-added. If we instead include all value-added measures at once, we find broadly similar effects, except that interpreting the test score and GPA value-added results is more difficult because the test score value-added measures are highly correlated, as are the GPA-based value-added measures (table A.8). The results suggest that the behavior value-added results are driven primarily by teachers' effects on suspensions and absences.

Another way to illustrate this is to move GPA value-added from behavior to learning skills value-added and conduct the main analysis again. The new behavior value-added constructed only from absences, suspension, and being held back has a significant or marginally significant effect in the expected direction on six high school outcomes (table A.9). The new GPA-based value-added has only a marginally significant effect on taking the SAT. The point estimates of the GPA-based value-added are often smaller than the significant effects of the other value-added measures, although the confidence intervals generally cannot reject effects of the magnitude of the behavior value-added estimates.

These results suggest multiple dimensions through which teachers affect long-term student outcomes, one that is closely related to increased

performance on tests and others related to reduced absences and suspensions. The abilities reflected in achievement GPA, effort GPA, and learning skills GPA matter for long-term outcomes, but the portion of these abilities that teachers are able to affect is largely captured by test scores and the ability to keep the students in the classroom.[10]

From a policy perspective, these results suggest that finding ways to keep children in the classroom may have large benefits; for example, the effect of reducing absence value-added on SAT scores is about the same size as the sum of the effects of the two test score value-added measures. There is also little evidence that a lack of on-time progression in elementary school results in worse high school outcomes, although our held-back value-added measure may be less informative than our other value-added measures because most elementary school students progress on time to the next grade.

E. Checking for Bias in Long-Term Effects

We conduct three analyses to look for evidence of bias in the estimates of the long-term effects of teachers. Consistent with the Chetty, Friedman, and Rockoff (2014a) results for fourth- to eighth-grade test score value-added, most tests show no evidence of bias, and the magnitude of the bias in the remaining tests is sufficiently small that it does not substantially affect our conclusions.

First, we show that the value-added measures are forecast unbiased. Specifically, all but one of the leave-year-out value-added variables cause an increase in the corresponding residualized achievement variable that is statistically indistinguishable from 1 (table A.11). Although this is a weak test, failing this test would be problematic. Only math test scores are statistically different from 1, for which a 1-unit increase in the math test score value-added causes a 0.99 standard deviation increase in math test scores. The confidence intervals are relatively tight for most outcomes, although absences, suspensions, and being held back are exceptions.

Second, we show that, after conditioning on the main controls, students expected to perform better in elementary school on the basis of their twice-lagged values of achievement are largely not sorting to higher value-added teachers. The estimated forecast bias from selection on student characteristics is between −1.6% and 1.3%, which is smaller than the Chetty, Friedman, and Rockoff's (2014a) point estimate of 2.2% for test score value-added (Rothstein 2007). The forecast bias is only marginally

[10] We also check how the results are affected by including GPA in test score value-added rather than in behavior value-added. The results are quite similar to those of our baseline specification, although the effect of behavior value-added on SAT and grade retention is about half as large, and the estimates are marginally significant and not statistically significant, respectively (table A.10).

significant for GPA, with a point estimate of 1.3% (table A.11). An analogous calculation using predicted high school outcomes from the twice-lagged values of the control variables shows no evidence of upward bias. The only significant point estimates are for behavior value-added, but each suggests that better students are sorted to worse teachers (table 7).

Third, we aggregate these data to the school-grade-year level and estimate long-term effects using quasi-experimental variation in teacher value-added caused by teachers switching between grades or schools. The analysis removes variation in teacher value-added caused by students sorting to teachers within a grade. Following Chetty, Friedman, and Rockoff (2014b), we regress changes in school-grade-year high school outcomes on changes in the mean teacher value-added weighted by the number of students.[11] Table A.12 shows that the signs on the estimated coefficients are generally consistent with the main results in table 5, and the point estimates tend to be larger. However, the estimates are much less precise. Despite this loss in statistical power, we observe a significant effect of test score value-added on math CAHSEEs and either significant or marginally significant effects in the expected direction of behavior value-added for the three GPA outcomes. As in the main table, learning skills value-added is often wrong-signed and for some outcomes is statistically significant. Overall, these tests show little evidence that unobservably better students sort to higher value-added teachers (i.e., that the key student sorting assumption in section IV.B is violated).

F. Robustness Checks

We conduct a number of robustness checks, in which we look at additional high school outcomes, use alternative approaches to constructing the value-added measures, estimate the long-term effects using different specifications, add controls for tracking, estimate a model that is robust to dynamic selection, and adjust for multiple hypothesis testing. The results of these robustness checks are qualitatively consistent with the main results.

1. Additional Outcomes

Table A.13 reports the effect of teacher value-added on additional high school outcomes, such as graduating from the LAUSD if enrolled in the LAUSD in twelfth grade, taking the PSAT, PSAT score, eleventh-grade English CST score (the last grade in which the CST is administered), eleventh-grade math CST score, eighth-grade science CST score, tenth-grade science

[11] The sample is limited to cases in which we have value-added measures for all teachers in a given school-grade in two consecutive years, to the subset of students for which we have both the long-term outcome variable and a teacher value-added measure, and to value-added measures that can be computed leaving out both year t and $t - 1$.

TABLE 7
EFFECT OF ELEMENTARY SCHOOL TEACHER VALUE-ADDED (VA) ON PREDICTED HIGH SCHOOL OUTCOMES

Pooled Grades 3–5 VA	LAUSD Dropout	Took SAT	SAT Score	GPA	Effort GPA	Cooperation GPA	Math CAHSEE	English CAHSEE	Days Suspended	Log Absences	Held Back
Test score VA	.000	.000	.423	.000	.000	.000	.000	.000	.000	.000	.000
	(.000)	(.001)	(.982)	(.001)	(.001)	(.001)	(.002)	(.002)	(.000)	(.001)	(.000)
Behavior VA	.001*	−.001*	−2.024*	−.003**	−.002**	−.001**	−.007**	−.006***	.000**	.002*	.001*
	(.000)	(.001)	(1.030)	(.001)	(.001)	(.001)	(.003)	(.002)	(.000)	(.001)	(.000)
Learning skills VA	.000	.000	.229	.000	.000	.000	−.001	.001	.000	.001	.000
	(.001)	(.001)	(1.128)	(.001)	(.001)	(.001)	(.003)	(.003)	(.000)	(.001)	(.000)
Observations	66,354	50,137	29,269	158,327	123,347	123,347	78,651	78,434	169,929	149,194	117,763
R^2	.962	.709	.750	.681	.665	.666	.702	.750	.764	.833	.737

NOTE.—The sample includes students in grades 3–5 who attended high school in the LAUSD. The unit of observation is a student–academic year. This table reports the effect of a standard-deviation increase in the three measures of elementary school teacher VA (see table 4) on students' predicted high school outcomes, using double-lagged student achievement. The predicted outcomes are created by estimating an OLS regression of the high school outcome indicated in each column on a cubic polynomial of double-lagged (year $t - 2$) math and English test scores, GPA, effort GPA, learning skills GPA, log absences, an indicator for suspension, an indicator for being held back, and an indicator of English-learner status. The coefficients obtained from this OLS regression are then used to predict students' high school outcomes. Each column of the table reports the coefficients on each of the three normalized measures of teacher VA from an OLS regression of the students' predicted high school outcome on the three measures of teacher VA for the students' teachers in grades 3, 4, or 5, along with the baseline controls described in sec. IV.A. The baseline controls include lags of a cubic polynomial of the student's math test score, English test score, GPA, and effort GPA; a cubic polynomial of lagged class- and grade-level means of each of those variables; current English-learner status, lagged log days absent, lagged suspensions, lagged being held back, and the class- and grade-level means of these variables. Each of these variables, except English-learner status, is fully interacted with grade fixed effects, and a control for class size is included. Standard errors clustered at the modal–high school level are reported in parentheses.

* $p < .10$.

** $p < .05$.

*** $p < .01$.

CST score, eighth-grade social studies CST score, eleventh-grade social studies CST score, world history CST score, and the number of AP courses. We see significant or marginally significant effects of test score value-added on all outcomes except LAUSD graduation and taking the PSAT. Test score value-added affects test score outcomes by between 0.013 and 0.018 standard deviations and does not vary noticeably by subject. Having a higher test score value-added teacher in elementary school increases the long-term performance across a number of subjects, not just English and math. The coefficients on behavior value-added are typically of the expected sign but are only marginally significant for one outcome, whereas the coefficients on the learning skills value-added are typically wrong-signed and are statistically significant for one outcome.

We also check whether the graduation results are robust to coding students as not graduating if they reached twelfth, eleventh, tenth, or ninth grade without having a recorded graduation status. We find slight increases in the effect of teacher value-added on graduation, but the estimates remain small and are not statistically significant (table A.14).

2. Alternative Value-Added Measures

We now show that the results are robust to a number of changes to our approach to computing the value-added indices and estimating the long-term effects. The main results are larger and more often statistically significant if we follow Chetty, Friedman, and Rockoff (2014b) when computing value-added measures by residualizing the achievement data using within-teacher variation in the controls (table A.15). The effects are even larger if we use Chetty, Friedman, and Rockoff's (2014b) approach to computing long-term effects by residualizing the outcome variables using within-teacher variation in the controls and then regressing the residualized high school outcomes on teacher value-added with no controls (table A.16). There is also more evidence of an effect of learning skills value-added on long-term outcomes, including significant effects in the expected direction on dropout rate and marginally significant effects on SAT score and English CAHSEE score.

We also estimate the effect of teachers on long-term outcomes using grade-specific measures of teacher value-added that should reduce measurement error due to grade-specific components of teacher value-added (table A.17).[12] The point estimates generally increase and are newly significant for some outcomes, although the effect of behavior value-added on taking the SAT is no longer statistically significant.

[12] We find that while the teacher value-added measured for the same teacher in different grades is closely related, the slope is well below 1 (fig. A.5).

The main results are essentially unchanged if we use factor analysis to construct the three indices for teacher quality (tables A.18, A.19) and weaker for the non–test score factor if we use exploratory factor analysis to construct two orthogonal factors (tables A.20, A.21). Alternatively, after behavior value-added is removed from the main analysis, learning skills value-added has fewer wrong-signed coefficients (table A.22). This suggests that part of the reason for the unintuitive results for learning skills value-added is that behavior value-added and learning skills value-added are moderately correlated. Additionally, we convert the test score value-added into deciles and test score outcomes into percentiles, which means that we use ordinal, rather than cardinal, measures of teacher value-added and test score outcomes. We continue to find significant effects of test score value-added on CAHSEE test scores and significant or marginally significant effects of behavior value-added on taking the SAT, math test scores, absences, and being held back (table A.23).

Finally, we test the sensitivity of our results to using year $t + 1$ measures of the behavior and learning skills to estimate behavior and learning skills value-added. First, we reestimate test score value-added using year $t + 1$ test scores, and then we replicate our main results from table 5 using test score, behavior, and learning skills value-added, all estimated using year $t + 1$ measures. Generally, we find that long-term effects of test score value-added are larger when estimated with $t + 1$ data than with year t data. There are also slight decreases in the effect of behavior value-added on some long-term outcomes (table A.24). We also reestimate behavior and learning skills value-added using year t data rather than year $t + 1$. The test score value-added measures are essentially unchanged, while the estimates of behavior value-added are generally smaller and the effects on SAT taking, math test scores, and grade retention are no longer statistically significant. However, there are still significant or marginally significant effects on the three GPA measures and log days absent. Interestingly, the learning skills estimates tend to be slightly larger in absolute value, and the results suggests that higher learning skills value-added leads to significantly or marginally significantly worse results for six high school outcomes (table A.25).

3. Additional Robustness Checks

We also check whether our results are sensitive to using GPA measured in z-scores, rather than grade points, both when computing value-added scores and as long-term outcomes (table A.26). Converting the GPA measures to z-scores before estimating teacher value-added had no effect on the relationship between the teacher value-added indices and the GPA outcomes measured in grade points. The statistical significance of GPA outcomes in standard-deviation units is the same as when they are measured

in grade points. The change in magnitudes is what would be expected, given the standard deviation of the high school outcomes reported in table 1.

Another concern with our interpretation of the results is that the effect of teacher value-added on long-term outcomes could be driven by the direct effect of teacher actions such as suspensions or retaining a student on both teacher value-added and the student's long-term outcomes. While this is likely happening to some degree, it is unlikely that this effect is large enough to influence our main results, because both types of events are rare for elementary school students. In any given year, only about 0.7% of elementary school students are held back, and the average elementary school student is suspended for 0.05 days (table 1). To check the magnitude of this effect, we replicate our main results excluding students who were held back or suspended. We find only slight changes in our results (table A.27). The magnitudes are very similar to the baseline results in table 5, and there are only minor changes in the statistical significance of the coefficients of interest. This suggests that the direct effect of suspension and grade retention plays a small role in our estimated results.

4. Tracking and Dynamic Selection

In this section, we first address the possible effect of selection due to tracking in middle and high school and then estimate a model that allows for dynamic selection. During the period we use to calculate value-added scores, students are tracked in math classes starting in eighth grade, they have the opportunity to take more advanced English classes starting in eleventh grade, and they may take elective English courses in both middle and high school. To the extent that our interest is in estimating the benefit to an individual elementary school student of being shifted into a higher value-added teacher's classroom, which is the estimated benefit a parent would want to know, our estimates are not biased by future tracking. Additionally, if a better teacher in elementary school results in a student being prepared to take rigorous courses in high school or increases the quality of the student's peers by directly increasing the peers' performance, we view that as an outcome of interest rather than a source of bias. This is particularly true in the LAUSD, where a large number of students struggle in rigorous coursework and do not progress from ninth to tenth grade on time. However, some benefits of a better elementary school teacher are zero-sum, so our estimates may be too large for certain policy experiments.

To help address this concern, we reestimate our main results from table 5 after adding controls for tracking by including fixed effects for the number of future middle and high school English electives, English helper classes, English AP classes, math classes on each of the three math

pathways, math helper or below-grade-level classes, and AP math courses. Including these controls likely results in overcontrolling for tracking. We find that the effect of a better teacher in elementary school on most high school outcomes falls slightly and that the statistical significance of the estimates is generally unaffected (table A.28). For example, the effect of test score value-added on math CAHSEE scores falls from 0.22 to 0.20. Two exceptions to this pattern are the effect of test score value-added on dropping out of the LAUSD, which becomes marginally significant, and the effect of test score value-added on effort GPA, which is no longer marginally significant. Additionally, figure 5 provides some evidence that tracking is not driving our results. Tracking starts 3–5 years after the current year for math and 5–8 years after the current year for English. There do not appear to be appreciable changes in figure 5 in the effects of a better teacher in the posttracking years. Finally, there is also little evidence that teacher value-added in grades 3–5 affects whether students leave the school district in subsequent years (fig. A.6).

Tracking and attrition are forms of a more general type of dynamic selection problem potentially present in our setting. Even if student assignment to teachers is conditionally independent of unobservables, being assigned to a good teacher may improve students' performance and result in their being assigned to better teachers in the future. This type of selection could ultimately improve their high school outcomes or reduce the probability that they leave the LAUSD. Epidemiologists face a similar dynamic selection problem because the treatments they study can affect participants' inclusion in the sample as well as subsequent treatment via the effect of the treatment on their control variables. Epidemiologists have developed a set of approaches to estimate treatment effects in the face of this problem, utilizing tools such as inverse probability weighting or parametric estimation to construct pseudopopulations where the controls do not affect future treatment and to simulate counterfactual treatment strategies (Hernán and Robins 2020). We apply this type of empirical method in our setting, following Robins (1986), to estimate the effect of teacher value-added on high school outcomes using student panel data from grades 3–5 and 3–12. We use the resulting estimates to simulate the effect of having a standard deviation better teacher from third to fifth grade on high school outcomes.[13]

The first panel of table A.29 shows the effect of having a teacher with a standard deviation better test score and behavior value-added in grades 3–5, estimated using the panel of grade 3–5 data. If there were constant returns to having better teachers in grades 3–5, we would expect the estimates in the first panel of table A.29 to be three times as large as those

[13] We use the particular application of Robins (1986) from Daniel, De Stavola, and Cousens (2011).

in table A.5. That is approximately what we find, and the statistical significance of the estimates, computed by bootstrapping, is similar to that in the main results. The second panel of table A.29 shows the same treatment effects estimated using a panel of grade 3–12 data. The results for test score value-added are similar. For behavior value-added, there is a reduction in dropping out but no statistically significant improvements in the other outcomes. The reduction in dropping out is reasonably large and will tend to reduce the effect of behavior value-added on other outcomes if the students induced not to drop out perform worse in school than other students. These results suggest that the effect of test score value-added on high school outcomes is robust to accounting for dynamic selection. The evidence that elementary school behavior value-added affects outcomes like test scores and grades is robust to accounting for dynamic selection in elementary school. Additionally, improved behavior value-added has benefits in terms of reduced dropout rate even after accounting for dynamic selection in high school.

5. Multiple-Hypothesis Testing

In this section, we use several approaches to check whether the statistical significance of the results is robust to adjusting for multiple-hypothesis testing. For the first approach, we construct indices of three families of outcomes: test scores (math and English CAHSEE),[14] GPA (GPA, effort GPA, and cooperation GPA), and behaviors (taking SAT, days suspended, log absences, and being held back). This approach greatly reduces the number of hypotheses tested, makes use of the information in all our outcomes, and is much less computationally intensive than other methods. Table A.30 shows the effect of each value-added measure on these three indices. Test score value-added significantly affects the test score and GPA indices, behavior value-added significantly affects all three indices, and learning skills value-added has no significant effect on these outcomes. These results suggest that the long-term effects of teachers are robust to reducing the number of hypothesis tests. We also apply the index approach to our main specification in table 5, where we include all three measures of teacher value-added simultaneously, and find similar results, except that the effect of behavior value-added on the test score index is not statistically significant (table A.31).

Next, we use a number of methods to adjust for multiple-hypothesis testing for each of our independent variables of interest. We recomputed the *p*-values using the Bonferroni-Holm, Sidak-Holm, and Romano-Wolf step-down methods, along with the Westfall and Young (1993)

[14] We do not include the SAT score because it would either make our sample much smaller or require us to impute an SAT score for a large fraction of students.

algorithm.[15] We grouped the outcomes into three families of hypothesis tests: test scores, grades, and all other outcomes. We then adjusted for multiple-hypothesis testing within each family and independent variable of interest. Table A.32 shows the results for each of these four approaches (using 10,000 bootstrap draws where applicable). The *p*-values for test-score value-added increase; however, the statistical significance of the results is generally robust to allowing for multiple-hypothesis testing, although effort GPA is no longer marginally significant under most adjustments. Similarly, for behavior value-added, the statistical significance of the results is generally robust to these approaches, with the exception that the effect on math test scores is now generally only marginally significant and that of being held back is generally statistically insignificant.

If instead of adjusting for multiple-hypothesis tests by family we adjust across all outcomes, the statistical significance of test score value-added is the same as that using *p*-values adjusted within family (table A.33). For behavior value-added, the effects on math test scores and cooperation GPA are no longer significant or marginally significant. The other outcomes that were statistically significant remain at least marginally significant. We also recomputed *p*-values for the index outcomes, allowing for multiple-hypothesis testing across all three outcomes. The statistical significance is unaffected, except that the *p*-values for test score value-added on GPA are only marginally significant for some adjustments (table A.33).

Broadly, these results show that we detect long-term, significant effects of having a teacher with a higher test score value-added and a higher behavior value-added and that our results are not driven by multiple-hypothesis testing. However, the significance of some individual outcomes is not robust to these adjustments. Notably, the evidence that behavior value-added has long-term effects on test scores, conditional on test score value-added, is weaker, although behavior value-added still significantly affects other outcomes.

VI. Applications of Non–Test Score Measures

In this section, we demonstrate that non–test score measures of achievement are useful for answering additional questions related to the effects of teachers on students. We first examine teacher effects over the educational life cycle and ask to what extent there could be gains from moving high value-added teachers between grades and what are the cumulative benefits from increasing teacher quality. We then take the approach used to construct non–test score value-added measures to compute GPA value-added for specific subjects in order to test the long-term value of having a better teacher in different subjects.

[15] We use code from Jones, Molitor, and Reif (2019) and Clarke, Romano, and Wolf (2020).

A. *Effects of Teacher Quality over the Educational Life Cycle*

The approach in this analysis is to compute the test score and behavior value-added for teachers in grades 3–12 and ask how having a standard deviation better teacher in each grade affects outcomes as measured in eleventh or twelfth grade.[16] We do not compute learning skills value-added because we do not have learning skills data for middle and high school students. Previous work on teacher effects by grade has estimated the effects of test score value-added for grades 4–8 (Chetty, Friedman, and Rockoff 2014b) and test score and non–test score value-added for grade 9 (Jackson 2018).

Figure 7 reports the results of this analysis for outcome variables measured as late as possible in a student's career. In each graph, we also report the sum of the coefficients across all grades. With some assumptions, particularly no tracking of students and no diminishing returns to having consecutive high-quality teachers (Kinsler 2016), this sum reflects an upper bound on the cumulative effect of having a teacher with a standard deviation higher test score or behavior value-added in each grade from the third through the twelfth grade. If tracking students plays a large role, or if diminishing returns exist, this sum overestimates the cumulative effect and is an upper bound. However, Chetty, Friedman, and Rockoff (2014b) find evidence for only a small amount of tracking, conditional on controls for lagged achievement in grades 4–8, and we include a robustness check that directly controls for tracking.

We find that having a teacher with a standard deviation higher test score value-added in grades 3–12 has a beneficial effect on taking the SAT, SAT scores, and math and English test scores. The cumulative effect for each of these outcomes is quite large. Using the cross-sectional relationship between test scores and earnings and the cumulative effects on math and English tests scores, having a teacher with a standard deviation higher test score value-added in each grade increases a student's adult earnings by 2.7%–5.2%.

We also find that having a teacher with a standard deviation higher behavior value-added in grades 3–12 has a large beneficial effect on dropping out of high school, graduation, taking the SAT, the three GPA measures, absences, suspensions, and grade retention. For example, the cumulative effects suggest that having a teacher with a standard deviation higher behavior value-added in each grade decreases the likelihood of dropping out of high school by 9.0 percentage points. Once adjusted for dropping out of high school being overestimated because of students leaving the LAUSD, this effect is still a 5.9 percentage point decrease.

[16] We cannot compute teacher value-added in twelfth grade, so value-added measures for teachers in twelfth grade use estimates of teacher quality in earlier grades.

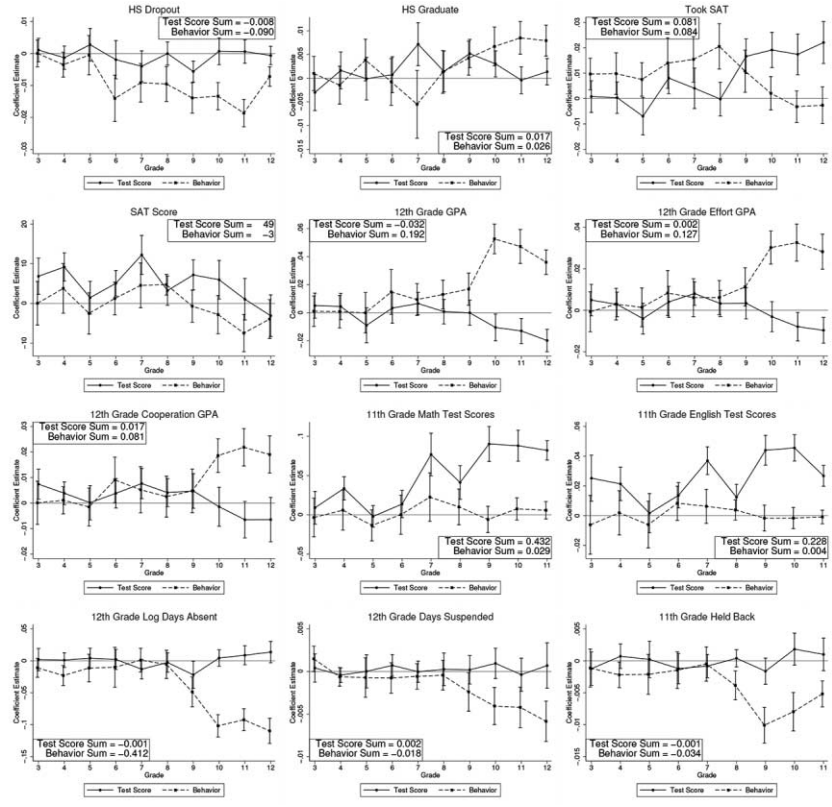

FIG. 7.—Effects of test score and behavior teacher value-added by grade. The sample includes students in grades 3–12 who attended high school in the LAUSD. The unit of observation is a student–academic year. The figure shows plots of the effect of test score and behavior teacher value-added on high school (HS) outcomes by the grade level of the student. The plotted coefficients and standard errors (clustered at the modal–high school level) are from a regression of the high school outcome variable on teacher test score and behavior value-added, and the vector of controls for high school students specified in section IV.A, estimated separately for each grade.

We also show that the magnitude of the cumulative effects of better teachers in figure 7 is not driven by tracking. We reestimate the value-added scores for middle and high school teachers after adding controls for contemporaneous tracking. Then we reestimate the grade-by-grade effects in elementary school, controlling for future tracking. For middle and high school grades, we use the value-added measures estimated with tracking controls, and when estimating the effect of teacher value-added on longer-term outcomes, we include controls for contemporaneous tracking (fig. A.7). We control for future tracking using the same approach as in table A.28 and control for contemporaneous tracking with

fixed effects for the number of English electives, English helper classes, English AP classes, classes on each of the three math pathways, math helper or below-grade-level classes, and AP math classes in each year. The effects of a better teacher on each outcome are generally smaller than those in figure 7, but only slightly, and our conclusions are unaffected. For example, cumulative effects on SAT taking fall from 8.1 and 8.4 percentage points for test score and behavior value-added to 7.8 and 7.7 percentage points, respectively.

We reestimate the cumulative effects using an alternative approach where we compute the average test score and behavior value-added for each student's teachers across grades 4–11 and then regress the outcomes reported in figure 7 on the average test score and behavior value-added and controls measured in fourth grade.[17] The support of the distribution includes ±1 standard deviation in teacher value-added (fig. A.8). The effects of behavior value-added on long-term outcomes are systematically smaller than those in figure 7, but they still imply substantial effects of providing students with teachers who have better behavior value-added for a number of outcomes (table A.34). The effect of test score value-added on outcomes is approximately as large and for some outcomes is notably larger than the estimates reported in figure 7. The outcomes for which the effect of test score value-added increases the most tend to be ones where behavior value-added decreases substantially, which suggests that this empirical approach puts more weight on better test score value-added teachers than on better behavior value-added teachers.

We also report the cumulative effect of having a standard deviation better teacher in the full grade 3–12 period, estimated with the correction for dynamic selection described in section V.F.4. We find large effects of test score value-added on test taking, GPA, and test scores that are generally consistent with the results estimated with OLS (compare table A.29 and fig. 7), although there is more evidence of an effect on improved GPA and less evidence of an effect on dropping out. For behavior value-added, we see large reductions in dropping out, suspensions, absences, and being held back and small increases in GPA. While behavior value-added may slightly lower students' test scores, the substantial reduction in dropping out changes the composition of students in a way that will tend to reduce performance on all other outcomes.

Since we have data on all grades 3–12, we can also examine whether having a high value-added elementary school teacher or high school teacher has a larger effect on students' high school outcomes under the strong assumption that a 1–standard deviation change in teacher value-added

[17] We start in fourth grade rather than third grade because it substantially increases our sample size. We include all students with at least seven observations of teacher value-added in the 8 years from grade 4 to grade 11.

induces the same amount of learning in each grade. Models of human capital formation in which past human capital production is complementary to current human capital production suggest that having a high value-added elementary school teacher is more important (Kinsler 2016). Alternatively, the substantial fade-out we observe in the effect of teacher value-added suggests that high school teachers will have a larger effect on student outcomes.

The results generally show that having a high value-added English or math teacher in middle school or high school has a bigger impact on high school outcomes than having a high value-added elementary school teacher. This pattern of results is especially clear for dropping out of high school, test scores, the GPA measures, absences, suspensions, and grade retention. For example, having a teacher with a standard deviation higher behavior value-added has little effect on whether a student drops out of high school in grades 3–5, but it reduces the likelihood of dropping out by about 1 percentage point per year in grades 6–12. Exceptions exist, notably for taking the SAT, but the pattern of results is fairly clear. The strength of the middle school and high school effects is somewhat surprising, because we calculate value-added only for English and math teachers, with whom a student spends less than half of her school day, whereas, in elementary school, the value-added measures are calculated for a classroom teacher with whom students spend much more time. However, these results should be interpreted with an abundance of caution, since common cardinality of the value-added measures across grades is difficult to fully defend, given that achievement measures are standardized within each grade-year. It is possible that 1 standard deviation in the value-added distribution at later grades represents a larger-magnitude shift in an underlying distribution of skill relative to a 1–standard deviation shift in earlier grades.

B. Measuring Teacher Quality in Untested Subjects

A significant shortcoming of the test score value-added framework is that a test must be administered in a subject to measure a teacher's quality in that subject. Consequently, we cannot evaluate teachers with value-added measures in subjects or grades that are untested or compare the importance of high-quality teachers across untested subjects. We extend the approach for calculating non–test score value-added described in section IV.A to compute value-added measures for elementary school teachers by subject using students' grades in each subject. We measure a teacher's quality using the grade each student receives in a subject in year $t + 1$, controlling for the baseline controls from year $t - 1$.

Table A.35 shows long-term student outcomes regressed on students' grades and the standard set of controls. Better grades in virtually all

subjects improve students' long-term outcomes. Two exceptions are speaking and PE (physical education), which indicate that students who perform better in speaking or PE perform worse in high school, even conditional on their grades in other subjects and prior achievement. However, these estimates may not detect the true effect of ability in a particular subject but may instead detect unobserved characteristics associated with both elementary school grades and high school outcomes.

Table 8 reports the results after redoing this analysis with teacher value-added for each subject. Teachers who excel at teaching reading and health have students who perform better in high school across several measures. The effects for math GPA value-added are subject specific. The effect is positive and marginally significant only for math CAHSEE scores. Unexpectedly, students with better elementary school math teachers are more likely to be held back in high school.

The effects of reading GPA value-added are more widespread, with significant or marginally significant effects on dropping out, taking the SAT, SAT score, and both math and English CAHSEE scores. Writing GPA value-added significantly affects the probability of being held back. Health GPA value-added significantly affects SAT taking and cooperation GPA and is marginally significant for both math and English CAHSEE scores. Social studies GPA value-added has positive and marginally significant effects on two of the grade measures.

Speaking GPA value-added has a positive and significant effect on dropout rate and negative and significant or marginally significant effects on SAT scores, both CAHSEE score measures, and all three GPA measures. Perhaps talking in class is not well rewarded in high school. The coefficients on PE, arts, and science also generally suggest negative effects on students but in most cases are not statistically significant.

The confidence intervals on the statistically insignificant estimates are such that they often cannot reject effects of the same magnitude found for reading or health GPA, although, for example, the effect of reading GPA on taking the SAT and test scores is outside the confidence interval for many of the other subjects. Similarly, several of the negative effects of speaking GPA value-added (e.g., dropping out and SAT score) are often rejected by the confidence intervals of the other value-added measures.

These results broadly support the traditional view that reading is a building-block skill that has long-term benefits. The health results are unexpected, suggesting that health knowledge at a young age could have long-term benefits, though this explanation should be interpreted with caution. Besides the negative effect of speaking, we find relatively little evidence of negative effects in other subjects. These results suggest that elementary schools could potentially create long-term benefits for students by hiring and retaining strong reading teachers.

TABLE 8

EFFECT OF ELEMENTARY SCHOOL TEACHER SUBJECT VALUE-ADDED (VA) ON HIGH SCHOOL OUTCOMES

Pooled Grades 3–5 VA	LAUSD Dropout	Took SAT	SAT Score	GPA	Effort GPA	Cooperation GPA	Math CAHSEE	English CAHSEE	Days Suspended	Log Absences	Held Back
Math GPA	.004	−.006	5.997	−.005	−.002	.001	.021*	.009	−.000	.001	.015***
	(.005)	(.007)	(4.139)	(.010)	(.006)	(.005)	(.011)	(.009)	(.004)	(.011)	(.005)
Reading GPA	−.011**	.029***	9.830**	.007	.006	.005	.030**	.028***	.002	.006	−.005
	(.005)	(.008)	(4.345)	(.009)	(.005)	(.006)	(.012)	(.010)	(.004)	(.011)	(.005)
Writing GPA	−.001	−.012	2.202	.016	.008	.007	−.003	.002	.002	−.005	−.012**
	(.005)	(.008)	(4.964)	(.010)	(.006)	(.006)	(.014)	(.011)	(.004)	(.012)	(.006)
Listening GPA	−.001	−.010	4.579	.003	.002	.004	.017	.011	−.003	−.013	.001
	(.006)	(.008)	(4.258)	(.011)	(.007)	(.007)	(.015)	(.011)	(.005)	(.010)	(.006)
Speaking GPA	.014**	−.003	−11.323**	−.025**	−.013*	−.014*	−.028*	−.025**	.001	.009	.001
	(.005)	(.007)	(4.700)	(.010)	(.007)	(.007)	(.016)	(.012)	(.004)	(.010)	(.005)
History/social science GPA	−.000	−.003	1.483	.009	.008*	.008*	−.002	.001	−.001	.006	−.006
	(.006)	(.006)	(3.824)	(.008)	(.005)	(.005)	(.011)	(.007)	(.003)	(.012)	(.006)
Science GPA	−.001	.000	−6.470	−.000	−.005	−.009	−.024*	−.013	−.003	−.010	−.003
	(.007)	(.008)	(4.824)	(.010)	(.006)	(.006)	(.013)	(.010)	(.004)	(.009)	(.006)
Health education GPA	.003	.015**	5.572	.007	.008	.012**	.023*	.019*	−.002	−.000	.002
	(.006)	(.007)	(4.198)	(.010)	(.006)	(.006)	(.013)	(.010)	(.003)	(.010)	(.005)
PE GPA	−.007	.011	−.223	.003	−.005	−.011*	−.008	.000	−.003	−.012	−.002
	(.006)	(.007)	(4.795)	(.009)	(.005)	(.006)	(.011)	(.009)	(.004)	(.009)	(.004)
Arts GPA	−.002	.000	−4.331	−.004	−.005	−.007	−.009	−.015*	.003	.007	.003
	(.006)	(.007)	(5.169)	(.010)	(.006)	(.006)	(.012)	(.008)	(.003)	(.011)	(.005)
Observations	136,125	102,822	60,875	293,569	233,529	233,529	152,693	152,162	316,740	277,858	222,174
R^2	.293	.146	.617	.244	.234	.240	.500	.512	.039	.266	.108

NOTE.—This table reports the effect of a standard-deviation increase in the 10 subject teacher VA measures on students' high school outcomes. Specifically, each column of the table reports the coefficients on each of the 10 subject VA measures from an OLS regression of the students' high school outcome on the 10 subject VA measures for the students' teachers in grades 3, 4, or 5, along with the baseline controls described in sec. IV.A. The baseline controls include lags of a cubic polynomial of the student's math test score, English test score, GPA, and effort GPA; a cubic polynomial of lagged class- and grade-level means of each of those variables; current English-learner status, lagged log days absent, lagged suspensions, lagged being held back, and the class- and grade-level means of these variables. Each of these variables, except English-learner status, is fully interacted with grade fixed effects, and a control for class size is included. Standard errors clustered at the modal–high school level are reported in parentheses.

* $p < .10$.

** $p < .05$.

*** $p < .01$.

VII. Conclusion

The results demonstrate that teacher quality is multidimensional. We show that teachers' test score value-added has significant effects on long-term outcomes and that adding controls for behavior and learning skills value-added does not influence the estimated effects. This finding indicates that the long-term effects of having a teacher with high test score value-added may not be biased upward by omitting measures of behavior or learning skills value-added.

We also find that a teacher value-added measure that combines the teacher value-added for GPA, absences, suspensions, and grade retention affects many high school outcomes. These effects are similar in magnitude to those of test score value-added. We find little evidence that learning skills value-added individually affects high school outcomes. The second dimension of teacher quality is only weakly correlated with test score value-added and allows for a substantial enhancement in the measurement of teacher quality. For example, a policy that uses both dimensions and 3 years of data to identify the bottom 5% of teachers and replaces them with average teachers increases the efficacy of the policy by over 50% versus a policy that uses only test score value-added for dropout rates, the likelihood of taking the SAT, GPA, effort GPA, absences, and being held back. Despite substantial gains in many areas, high school test scores experience only minimal declines.

We then demonstrate that this value-added framework can be extended to analyze effects by grade and all elementary school subjects. We find that the cumulative effect of better elementary, middle, and high school teachers is large, even after controlling for tracking and dynamic selection. We also show that teachers who are relatively better at teaching reading and health improve their elementary school students' high school outcomes, whereas teachers who are better at teaching speaking worsen them. Teaching reading may have long-term benefits for students, which suggests that schools should focus on increasing teaching quality in that area.

Overall, this paper shows the multifaceted role that teachers play in influencing the outcomes of their students. In addition to benefiting students through increasing performance on tests, teachers meaningfully affect students' long-term outcomes through additional channels.

Data Availability

Code replicating the tables and figures in this article can be found in Petek and Pope (2022) in the Harvard Dataverse, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BAE6IH.

References

Ad Council. 2015. "California Attendance Parent Survey Results." https://oag.ca.gov/sites/all/files/agweb/pdfs/tr/toolkit/QuantitativeResearchReport.pdf.

Araujo, M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady. 2016. "Teacher Quality and Learning Outcomes in Kindergarten." *Q. J.E.* 131 (3): 1415–53.

Bacher-Hicks, Andrew, Thomas J. Kane, and Douglas O. Staiger. 2014. "Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles." Working Paper no. 20657 (November), NBER, Cambridge, MA.

Blazar, David, and Matthew A. Kraft. 2017. "Teacher and Teaching Effects on Students' Attitudes and Behaviors." *Educ. Evaluation and Policy Analysis* 39 (1): 146–70.

Carrell, Scott E., and James E. West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." *J.P.E.* 118 (3): 409–32.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2011. "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." Working Paper no. 17699 (December), NBER, Cambridge, MA.

———. 2014a. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *A.E.R.* 104 (9): 2593–632.

———. 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *A.E.R.* 104 (9): 2633–79.

———. 2017. "Measuring the Impacts of Teachers: Reply." *A.E.R.* 107 (6): 1685–717.

Clarke, Damian, Joseph P. Romano, and Michael Wolf. 2020. "The Romano-Wolf Multiple Hypothesis Correction in Stata." *Stata J.* 20 (4): 812–43.

Cook, Philip J., Kenneth A. Dodge, Elizabeth J. Gifford, and Amy B. Schulting. 2017. "A New Program to Prevent Primary School Absenteeism: Results of a Pilot Study in Five Schools." *Children and Youth Services Rev.* 82:262–70.

Cunha, Flávio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3): 883–931.

Daniel, Rhian M., Bianca L. De Stavola, and Simon N. Cousens. 2011. "gformula: Estimating Causal Effects in the Presence of Time-Varying Confounding or Mediation Using the g-Computation Formula." *Stata J.* 11 (4): 479–517.

Figlio, David N., Morton O. Schapiro, and Kevin B. Soter. 2015. "Are Tenure Track Professors Better Teachers?" *Rev. Econ. and Statis.* 97 (4): 715–24.

Flèche, Sarah. 2017. "Teacher Quality, Test Scores, and Non-Cognitive Skills: Evidence from Primary School Teachers in the UK." Working paper.

Fryer, Roland G. 2013. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." *J. Labor Econ.* 31 (2): 373–407.

Gershenson, Seth. 2016. "Linking Teacher Quality, Student Attendance, and Student Achievement." *Educ. Finance and Policy* 11 (2): 123–49.

Goldhaber, Dan, and Michael Hansen. 2010. "Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions." Working Paper no. 31 (February), Nat. Center Analysis Longitudinal Data Educ. Res., Arlington, VA.

Goodman, Sarena, and Lesley Turner. 2013. "The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program." *J. Labor Econ.* 31 (2): 409–20.

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. "Identifying Effective Teachers Using Performance on the Job." Discussion Paper 2006-01, Brookings Inst., Washington, DC.

Hanushek, Eric A. 1971. "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data." *A.E.R.* 61 (2): 280–88.
———. 2011. "Valuing Teachers: How Much Is a Good Teacher Worth?" *Educ. Next* 11 (3): 41–45.
Hanushek, Eric A., and Steven G. Rivkin. 2010. "Generalizations about Using Value-Added Measures of Teacher Quality." *A.E.R.* 100 (2): 267–71.
Hanushek, Eric A., and Ludger Woessmann. 2008. "The Role of Cognitive Skills in Economic Development." *J. Econ. Literature* 46 (3): 607–68.
Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes." *A.E.R.* 103 (6): 2052–86.
Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *J. Labor Econ.* 24 (3): 411–82.
Hernán, Miguel A., and James M. Robins. 2020. *Causal Inference: What If*. Boca Raton, FL: Chapman & Hall/CRC.
Horn, Sandra P., and William L. Sanders. 1994. "The Tennessee Value-Added Assessment System (TVAAS): Mixed-Model Methodology in Education Assessment." *J. Personnel Evaluation Educ.* 8 (3): 299–311.
Jackson, C. Kirabo. 2018. "What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes." *J.P.E.* 126 (5): 2072–107.
Jackson, C. Kirabo, Jonah E. Rockoff, and Douglas O. Staiger. 2014. "Teacher Effects and Teacher-Related Policies." *Ann. Rev. Econ.* 6:801–25.
Jacob, Brian A., Lars Lefgren, and David P. Sims. 2010. "The Persistence of Teacher-Induced Learning." *J. Human Resources* 45 (4): 915–43.
Jennings, Jennifer L., and Thomas A. DiPrete. 2010. "Teacher Effects on Social and Behavioral Skills in Early Elementary School." *Sociology Educ.* 83 (2): 135–59.
Jones, Damon, David Molitor, and Julian Reif. 2019. "What Do Workplace Wellness Programs Do? Evidence from the Illinois Workplace Wellness Study." *Q.J.E.* 134 (4): 1747–91.
Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment." MET Project res. paper, Bill & Melinda Gates Found., Seattle.
Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Working Paper No. 14607 (December), NBER, Cambridge, MA.
Kinsler, Josh. 2016. "Teacher Complementarities in Test Score Production: Evidence from Primary School." *J. Labor Econ.* 34 (1): 29–61.
Kraft, Matthew A. 2019. "Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies." *J. Human Resources* 54 (1): 1–36.
Kraft, Matthew A., and Todd Rogers. 2015. "The Underutilized Potential of Teacher-to-Parent Communication: Evidence from a Field Experiment." *Econ. Educ. Rev.* 47:49–63.
Lazear, Edward P. 2003. "Teacher Incentives." *Swedish Econ. Policy Rev.* 10 (3): 179214.
McQuiggan, Meghan, and Mahi Megra. 2017. "Parent and Family Involvement in Education: Results from the National Household Education Surveys Program of 2016." NCES 2017-102, Nat. Center Educ. Statis., US Dept. Educ., Washington, DC.
Mihaly, Kata, Daniel F. McCaffrey, Douglas O. Staiger, and J. R. Lockwood. 2013. "A Composite Estimator of Effective Teaching." MET Project res. brief, Bill & Melinda Gates Found., Seattle.

Mulligan, Casey B. 1999. "Galton versus the Human Capital Approach to Inheritance." *J.P.E.* 107 (S6): S184–S224.

Murnane, Richard J., John B. Willett, Yves Duhaldeborde, and John H. Tyler. 2000. "How Important Are the Cognitive Skills of Teenagers in Predicting Subsequent Earnings?" *J. Policy Analysis and Management* 19 (4): 547–68.

National Council on Teacher Quality. 2015. *2015 State Teacher Policy Yearbook: National Summary*. Washington, DC: Nat. Council Teacher Quality.

Neal, Derek. 2011. "The Design of Performance Pay in Education." In *Handbook of the Economics of Education*, vol. 4, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 495–550. Amsterdam: North-Holland.

Petek, Nathan, and Nolan G. Pope. 2022. "Replication Code for 'The Multidimensional Impact of Teachers on Students.'" Harvard Dataverse. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BAE6IH.

Robins, James. 1986. "A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect." *Math. Modelling* 7 (9–12): 1393–512.

Robinson, Carly D., Monica G. Lee, Eric Dearing, and Todd Rogers. 2018. "Reducing Student Absenteeism in the Early Grades by Targeting Parental Beliefs." *American Educ. Res. J.* 55 (6): 1163–92.

Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *A.E.R.* 94 (2): 247–52.

Rockoff, Jonah, E., and Cecilia Speroni. 2010. "Subjective and Objective Evaluations of Teacher Effectiveness." *A.E.R.* 100 (2): 261–66.

Rogers, Todd, and Avi Feller. 2018. "Reducing Student Absences at Scale by Targeting Parents' Misbeliefs." *Nature Human Behaviour* 2 (5): 335–42.

Romano, Joseph P., and Michael Wolf. 2016. "Efficient Computation of Adjusted *p*-Values for Resampling-Based Stepdown Multiple Testing." *Statis. and Probability Letters* 113: 38–40.

Rothstein, Jesse. 2007. "Does Competition among Public Schools Benefit Students and Taxpayers? Comment." *A.E.R.* 97 (5): 2026–37.

———. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Q.J.E.* 125 (1): 175–214.

———. 2017. "Measuring the Impacts of Teachers: Comment." *A.E.R.* 107 (6): 1656–84.

Ruzek, Erik A., Thurston Domina, AnneMarie M. Conley, Greg J. Duncan, and Stuart A. Karabenick. 2015. "Using Value-Added Models to Measure Teacher Effects on Students' Motivation and Achievement." *J. Early Adolescence* 35 (5–6): 852–82.

Westfall, Peter H., and S. Stanley Young. 1993. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York: Wiley.