NATIONAL CENTER *for* ANALYSIS *of* LONGITUDINAL DATA *in* EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of North Carolina at Chapel Hill, University of Texas at Dallas, and University of Washington

AIR®

Teachers and Students' Postsecondary Outcomes: Testing the Predictive Power of Test and Nontest Teacher Quality Measures

Ben Backes
James Cowan
Dan Goldhaber
Roddy Theobald

WORKING PAPER No. 270-1022 • October 2022

Teachers and Students' Postsecondary Outcomes: Testing the Predictive Power of Test and Nontest Teacher Quality Measures

Ben Backes
American Institutes for Research/CALDER

James Cowan
American Institutes for Research/CALDER

Dan Goldhaber
American Institutes for Research/CALDER

Roddy Theobald
American Institutes for Research/CALDER

Contents

Acknowledgments

This research was funded by IES Research Grant R305S210012 to the Massachusetts Department of Elementary and Secondary Education and American Institutes for Research, and supported by the National Center for the Analysis of Longitudinal Data in Education Research (CALDER), which is funded by a consortium of foundations. For more information about CALDER funders, see www.caldercenter.org/about-calder. The data were provided by the state of Massachusetts and the state had the right to review the paper prior to public release in order to ensure that the paper does not disclose any personally identifiable information provided by the state. The authors thank Claire Abbott, Ron Ferguson, Andrew Ho, Kirabo Jackson, Pierre Lucien, Elana McDermott, and Aubree Webb for comments that improved the paper.

CALDER working papers have not undergone final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication. Any opinions, findings, and conclusions expressed in these papers are those of the authors and do not necessarily reflect the views of our funders. All errors are our own.

CALDER • American Institutes for Research
1400 Crystal Drive 10th Floor, Arlington, VA 22202
202-403-5796 • www.caldercenter.org

Teachers and Students' Postsecondary Outcomes: Testing the Predictive Power of Test and Nontest Teacher Quality Measures

Ben Backes, James Cowan, Dan Goldhaber, and Roddy Theobald

CALDER Working Paper No. 270-1022
October 2022

Abstract

We examine how different measures of teacher quality are related to students' long-run trajectories. Comparing teachers' *test-based* value-added to *nontest* value-added – based on contributions to student absences and grades – we find that test and nontest value-added have similar effects on the average quality of colleges that students attend. However, test-based teacher quality measures have more explanatory power for outcomes relevant for students at the top of the achievement distribution such as attending a more selective college, while nontest measures have more explanatory power for whether students graduate from high school and enroll in college at all.

1. Introduction

Understanding the different ways in which teachers influence student learning is a pressing policy and research concern. Test-based value-added measures—which capture the impact of teachers on student test scores—are a primary way of understanding the effects of teachers on student outcomes in policy and research. Thirty-four states now require an objective measure of student growth in their teacher evaluation systems, with more than half of these states using data from standardized tests (National Council on Teacher Quality [NCTQ], 2019). Researchers have used value-added measures to examine how equitably teachers are distributed across students (Goldhaber et al., 2017; Isenberg et al., 2016; Williams et al., 2016); changes in the teacher workforce over time (Nagler et al., 2019); how well teacher evaluation systems identify effective teachers (Goldhaber, 2007; Jacob et al., 2018; Kane et al., 2011); and how well licensure testing serves as a screening instrument for prospective teachers (Cowan et al., 2020; Goldhaber et al., 2017). Policy and research interest in test-based value-added measures is warranted because these measures appear to matter for students' later life outcomes. For example, Chetty and colleagues (2014b) find that being assigned to a teacher with high test value-added is associated with increases in future college quality and adult earnings.

Test-based value-added measures also have clear limitations. For example, although these measures can give us a clear sense of teacher contributions to student test scores, they have little to say about how teachers affect other important outcomes (e.g., a student's persistence in school). Recent research has begun to address this shortcoming and broaden the field's understanding of teacher quality and value-added by applying value-added techniques to nontest outcomes (e.g., high school graduation, attendance, and course grades). This research tends to find that nontest quality measures are not highly correlated with test-based quality measures (e.g., Backes & Hansen, 2018; Gershenson, 2016; Jackson, 2018; Kraft, 2019; Liu & Loeb, 2021). This literature has also found that nontest teacher quality measures are more predictive of high school outcomes, such as graduation, than test-based measures (Jackson, 2018; Liu & Loeb, 2021). These results raise a fundamental conundrum: why is test value-added predictive of postsecondary outcomes, such as those related to college quality and earnings as identified by Chetty and colleagues, but not predictive of the secondary outcomes that we expect would mediate these relationships?

In this paper, we address this conundrum by connecting both test and nontest value-added to students' secondary and postsecondary outcomes to examine how different teacher quality measures predict their students' later outcomes. To date, the nontest value-added literature has primarily examined later secondary outcomes, including SAT scores and high school graduation. While important, these outcomes are limited for two reasons. First, because the vast majority of students graduate from high school (90% in the sample used in this paper, for example), this binary measure may not detect teacher effects on students in the upper portion of the achievement distribution. Second, most of the existing research uses proxies for postsecondary outcomes, but there is limited empirical evidence on the extent to which these proxies (e.g., improvements in SAT taking or plans to attend college) translate into gains in college enrollment or college quality. In contrast, the measure we use in our main results —college quality as proxied by the median future earnings of graduates of the college—is a continuous long-run outcome measure that has been found to be associated with future labor market outcomes (e.g., Chetty et al., 2017).

Using a sample of students from Massachusetts, we begin by replicating several prior findings on teacher quality. We document teacher effects on both tests and a composite of nontest outcomes used in prior research: course grades, suspensions, attendance, and grade promotion (Jackson, 2018). As in prior work, we find that these nontest effects predict a range of secondary outcomes, such as dropout and graduation, that are not well predicted by teacher value-added to student test scores (Jackson, 2018; Liu & Loeb, 2021). We then examine teacher effects on several postsecondary outcomes, such as college enrollment and college quality. We find that teacher value-added to test and nontest outcomes both play an important role in predicting these long-run outcomes. However, some outcomes, like enrollment in a 4-year college, are better predicted by teacher effects on nontest outcomes, while others, like enrolling in a selective college, are better predicted by test-based value-added. These results are robust when it comes to the various ways of accounting for student-teacher sorting, such as the addition of school-track fixed effects or using the quasi-experimental approach developed by Chetty and colleagues (2014a) that identifies effects based on teachers who switch schools or grades.

To help explain the divergent results between teacher test and nontest quality measures, we next consider the distributional effects of test and nontest value-added using a continuous outcome measure – college quality – capable of summarizing teacher effects on students at different points in the postsecondary outcomes distribution. Using a quantile regression approach, we find that college quality is more sensitive to teachers' test value-added at the top of the outcome distribution. In contrast, nontest measures of teaching quality have larger effects for students at the bottom of the outcome distribution. These findings suggest that discrepancies in prior results comparing test and nontest teacher measures may be driven, in part, by the types of students affected by different teaching skills.

With this paper, we make two primary contributions. First, to our knowledge, this is the first paper to connect nontest value-added measures to student college enrollment outcomes. We find that nontest value-added affects college enrollment, the likelihood of enrolling in a 4-year college, and overall college quality, even though there is no relationship between nontest value-added and some secondary academic outcomes, such as SAT test scores or passing Advanced Placement (AP) tests. Second, we help to explain the conundrum described earlier: test-based value-added measures often do not predict some outcomes in high school that should, in theory, contribute to postsecondary success even though these same measures predict postsecondary outcomes. How is it that Chetty and colleagues (2014b) find long-run effects of test value-added on outcomes such as college quality and adult earnings, but recent papers find near zero effects of test value-added on more immediate outcomes such as SAT test taking (Petek & Pope, 2021; Gilraine & Pope, 2021) and high school graduation (Jackson, 2018; Liu & Loeb, 2021; Gilraine & Pope, 2021)? We replicate these findings and demonstrate that this pattern is driven by (a) different long-run outcomes being more sensitive to students' placement in the achievement and nontest distributions along with (b) the varying strength of test and nontest impacts across the distribution of outcomes.

This paper adds to a growing body of evidence that highlights the fact that solely focusing on test value-added misses important ways in which teachers affect student outcomes. Even when accounting for teachers' test-based value-added and students' school and academic tracks, we find that a composite measure of teachers' nontest value-added contains independent predictive power for whether a student enrolls in a 4-year college and the quality of a college. Because the nontest composite is created with measures already collected in routine data collection, such measures are readily available to inform real-world policy decisions. For example, education agencies could begin using such measures to better understand the quality and contributions of their teacher workforce and examine how different teacher workforce policies might affect short- and longer run outcomes for students.[1]

2. Background and Prior Literature

The statistical properties of traditional test value-added measures of teacher quality have been rigorously evaluated in both experimental (Bacher-Hicks et al., 2019; Kane et al., 2013; Kane & Staiger, 2008) and nonexperimental (Bacher-Hicks et al., 2014; Chetty et al., 2014a) settings. These studies found that, conditioning on prior student achievement and other data about the student and classroom context, test value-added measures provide a causal estimate of teacher contributions to students' short-run achievement with low bias from classroom context or other confounders. Using these measures, researchers have found substantial variation in teaching effectiveness (Chetty et al., 2014a, 2014b; Rivkin et al., 2005). As already noted, researchers have found that test value-added measures impact later outcomes such as college attendance and even earnings (Chetty et al., 2014b), although they appear to capture a small portion of teachers' overall contributions toward these outcomes (Chamberlain, 2013).

Several studies have examined the long-run effects of teachers using a variety of proxies for student skill acquisition and long-run outcomes (see Table 1). These studies examine the relationship between value-added in the short term and student outcomes in the long run.[2] Chetty and colleagues (2014b) provided some of the first evidence on the long-run effects of teachers operating through test scores. They find that elementary and middle school students receiving instruction from teachers with a one standard deviation higher value-added are estimated to be

[1] Because the reporting of these nontest measures is directly manipulable by teachers and schools in a way that does not necessarily apply to test scores, it is not clear how well they would serve for high-stakes accountability purposes.

[2] An alternative approach would be to directly estimate teacher value-added to long-run outcomes. However, these measures would be impractical for policy applications due to the need to wait for years between exposure to a teacher and, for example, observed college enrollment.

0.7 percentage points more likely to enroll in college and 0.8 percentage points more likely to complete at least 4 years of college by age 22 and to raise income at age 28 by about $300–$350 per year.

Jackson (2018) demonstrated that teachers also have long-run effects operating through effects that are not fully captured by standardized test scores. He constructs two measures of teacher value-added: one using test scores and another using an index created from four nontest outcomes (grades, absences, suspensions, and grade promotion) and investigates the degree to which these affect high school graduation. He finds teacher effects of a similar magnitude across test and nontest outcomes, but the correlation between the two types of teacher effects is quite small. The effects of teaching skills also differ significantly across the measures. Assignment to a teacher one standard deviation higher on the nontest value-added measure improves on-time high school graduation by about 1.5 percentage points and reduces the dropout rate by about 0.4 percentage points; a one standard deviation increase in test value-added increases on-time high school completion by only 0.1 percentage points and has no detectible effect on the dropout rate. Teachers with higher nontest value-added also have larger effects on cumulative grade point averages (GPAs), whether students take the SAT and whether they intend to enroll in college. On the other hand, test value-added is more strongly linked to SAT scores than to nontest teacher quality measures.

Several studies (including this one) have also used Jackson’s nontest index of teacher quality to assess teachers’ causal contributions to future student outcomes. Petek and Pope (2021) and Gilraine and Pope (2021) construct a nontest factor and find the nontest value-added of elementary teachers is more predictive of not repeating a grade (Petek & Pope, 2021) and also of being more likely to graduate from high school (Gilraine & Pope, 2021). Liu and Loeb (2021) study the effects of teachers on absences in middle and high school. They use data from Grades 7–11 that links absences to the particular class that students missed, which allows them to construct a course-specific measure of unexcused absences. As with other research on nontest teacher effects, test and nontest measures of teachers are not strong predictors of the other outcome; the effects of test and attendance value-added on the other outcome are both an order of magnitude smaller than the effects of value-added to the same outcome. As with the prior studies mentioned, Liu and Loeb (2021) find that attendance value-added is a better predictor of high school completion. Specifically, they find that a one standard deviation increase in attendance value-added improves the on-time graduation rate by 0.7 percentage points and reduces the dropout rate by about 0.3 percentage points. They also find that test value-added has no detectible effect on either high school completion or dropout rates. However, the results differ when they look at the effects on participation in AP courses: the effects of test value-added are about 50% larger for both the number of AP courses taken and the total number of AP credits earned compared with the effects of attendance value-added.

Mulhern and Opper (2022) also study the long-run effects of elementary and middle school teachers. They find little evidence that teacher value-added constructed from individual short run test or nontest measures affects high school completion outcomes. The main exception is in middle school, where they find that a one standard deviation increase in test value-added increases the likelihood of earning a Regents diploma by about 0.2 percentage points. They also find some evidence that teacher value-added to attendance may reduce high school completion rates, although teachers who improve attendance in the next school year do appear to increase completion rates. Nonetheless, they do find that combining teacher skills on tests, nontest, and future academic outcomes into a single teacher skill index does better predict student long run outcomes. For example, a one standard deviation increase on the combined metric improves high school graduation by about 2–3 percentage points, and the effects are about two times as large in elementary school compared with middle school.[3]

Although the results differ by study and context, there appear to be some common trends in this emerging literature. First, teachers do appear to have long-run effects on students that operate through improvements in both short-run academic achievement and learning behaviors. Second, the effects of teachers on nontest outcomes are generally not highly correlated with teacher effects on test scores. The lack of correlation is consistent with the notion that test and nontest student measures may capture different teacher skills; such distinctions also resonate with evidence that suggests measures of teacher practice and students' perceptions of their teachers pick up distinct contributions that teachers make to student learning (Danielson & Ferguson, 2014). The idea that these different measures can capture different dimensions of teacher quality is buttressed by new experimental evidence that finds negative correlations between measures of teacher contributions to student math test scores and their contributions to student-reported measures of classroom engagement (Blazer & Pollard, 2022).

Third, the literature suggests that teacher effects on test scores have larger effects for outcomes that are proximate to college success (e.g., AP credits, SAT scores) and little to no effect on high school completion outcomes. Last, prior research suggests that teacher effects that operate through improvements to nontest outcomes appear to have larger effects on outcomes that are more proximate to high school completion or to the college enrollment/non-enrollment

[3] Nontest value-added has also been shown to predict nonacademic outcomes. For example, Rose and colleagues (2022) estimate teachers' impacts on contact with the criminal justice system. Using teachers of students in Grades 4–8, the authors found that value-added to absences and suspensions substantially reduced future arrests, in contrast to value-added to test scores, which was unrelated to future arrests. Like the papers discussed above, Rose and colleagues (2022) found that nontest value-added better predicted high school graduation rates compared with test value-added.

margin. As we describe in the sections that follow, we explore these issues further by focusing on different margins of students' postsecondary educational achievement.

3. Data and Measures

We use a sample of students from Massachusetts matched to teachers, end-of-year standardized tests and nontest outcomes, and long-run outcomes. These data, obtained through a data sharing agreement with the Massachusetts' Department of Elementary and Secondary Education (DESE), include student-teacher matches between 2012 and 2019 and students' postsecondary outcomes through 2021. To ensure sufficient cohorts of students who can be connected both to their K–12 teachers and to their postsecondary outcomes, we focus on students in Grades 7, 8, and 10 (i.e., grades in which both test and nontest outcomes are available and that are sufficiently proximate to postsecondary outcomes to permit linkages during the available data panel). The final sample includes teachers in math and English language arts (ELA) in Grades 7 (2012–2015), 8 (2012–2016), and 10 (2012–2018). The matched sample included about 85–90% of students in each school year and grade. Summary statistics for the matched and unmatched samples are included in Table 2.[4]

3.1 Data and Measures

Math and ELA Achievement: We use standardized test data for math and ELA in Grades 7, 8, and 10. We standardize each test to be mean zero, standard deviation one within each grade, subject, and year given that Massachusetts implemented multiple standardized tests during this period.[5]

[4] The racial composition of the analysis samples is broadly similar to official reported numbers: https://profiles.doe.mass.edu/statereport/enrollmentbyracegender.aspx

[5] Prior work has found that test-based value-added is relatively stable when states change from one assessment to another (Backes et al., 2018). We apply a normal curve equivalent transformation to the test scale scores given that in some years Massachusetts applies a nonlinear transformation to the individual scores to obtain scaled scores (Jacob & Rothstein, 2017).

Nontest Index: Following Jackson (2018), we use four nontest outcomes commonly found in state administrative data systems (absences, discipline, grades, and grade progression) to construct a behavioral index measure using exploratory factor analysis.[6] The student enrollment data report the total days a student was enrolled and in attendance for at least half the school day. We calculate the number of days absent and use the log of total absences (plus 1) as an outcome. The administrative data collection also includes a report of all disciplinary actions that result in suspension.[7] Following prior studies, we use the log of total days suspended (plus 1). The student transcript data includes courses and grades reported on a numeric (0–100%) or grade point (0.0–4.0) scale. We convert numeric grades to a GPA (i.e., 3.7 for a score from 90 to 93 on the numeric scale) and calculate a student's GPA in the current school year. Finally, we identify grade promotion using enrollment data. We define grade progression as a student enrolling in the next grade during the following school year.

We match the student data to several secondary and postsecondary outcomes. For each of the secondary and postsecondary outcomes, we use data from all years following the teacher assignment through the academic year after scheduled graduation.

Secondary Outcomes:[8] We use student enrollment and transcript data to identify several key secondary outcomes. We use student enrollment records to measure credits earned through

[6] We use all students enrolled in Grades 7–12 to estimate the factor model using the Bartlett scoring method. The factor weights days absent (-0.57), days suspended (-0.36), GPA (0.76), and grade retention (-0.24). For comparison, Jackson (2018) computed -0.21 for absences, -0.15 for suspensions, 0.38 for GPA, and -0.31 for grade retention.

[7] Before 2013, the discipline data only includes infractions related to drug, violent, or criminal offenses (and the resulting disciplinary action). Starting in 2013 and thereafter, the data include all disciplinary actions that resulted in suspensions. Drug, violent, and criminal offenses comprised 34% of all suspensions in 2013 and later. In addition, the state implemented a law in the 2014–15 school year intended to reduce the number of out-of-school suspensions. The average number of days suspended increased from 0.12 days in 2010–11 and 2011–12 to 0.25 days between 2012–13 and 2013–14 and falls to 0.16 days between 2014–15 and 2018–19.

[8] Most of the secondary student outcomes are only measured for students who enrolled in public high schools in Massachusetts. Among students in our sample in Grades 7 and 8, we observed 93% with public high school enrollments. We limited the sample in these grades to students enrolling in public high schools.

AP courses (i.e., from passing AP courses [as distinct from AP tests] in high school). In addition, we use the linked AP data to measure the total number of actual AP tests taken and passed in subsequent years. We use the student enrollment data to measure dropout and graduation events. The enrollment records track confirmed dropouts, but this may understate the true dropout rate among students with unknown enrollment status (Sorensen, 2019).

A number of prior studies have used various measures as proxies for college plans (e.g., self-reported plans to attend a 4-year college after graduation as in Rose et al. [2022], or whether a student takes the SAT). Because we have information on actual college enrollment (described below), we do not need to use such a proxy for postsecondary enrollment. However, to reconcile our findings with prior work, we also use an indicator for whether the student takes the SAT.[9]

Postsecondary Outcomes: The student data are linked to postsecondary enrollment using data from the National Student Clearinghouse (NSC). The NSC covers about 92% of all college enrollments in the United States and about 95% of all college enrollments in Massachusetts (Dynarski et al., 2015). We use the NSC data to measure enrollment in college the year after high school graduation. We identify the level (2-year or 4-year) of the college a student initially attends.

We then match enrollment data to the College Mobility Report Card constructed by Chetty and colleagues (2017). Following Chetty and colleagues (2017), we use an index of college quality based on the median earnings of students at ages 33–35 who attended the college

[9] Prior to taking the SAT, students fill out a Student Data Questionnaire that inquires about students' college degree goals (among other topics). The vast majority of students who take the SAT intend to obtain an associate's degree or higher (86%), with the bulk of the remainder being undecided (13%). Source: 2021 College Board Annual Report, https://reports.collegeboard.org/media/2022-04/2021-total-group-sat-suite-of-assessments-annual-report%20%281%29.pdf

(or did not attend college at all) from the 1980–1982 birth cohorts.[10] This index is available for students who do not enroll in college and thus measured for the entire sample.[11] We supplement the Chetty and colleagues (2017) measure with additional data on high school non-completers from the American Community Survey (ACS). To match the procedures used by Chetty and colleagues (2017) as closely as possible, we consider the median earnings in 2011–2015 for people born in Massachusetts who were ages 33–35 during the previous year.[12] The non-completer group includes those who obtain a GED or other alternative credential, those reporting 12 years of education but no high school degree, and those reporting fewer than 12 years of education. We impute these earnings for all students in our sample who fail to complete high school and are not observed to enroll in college.

Second, we create a binary measure denoting whether a student enrolled in a highly selective college. We identify highly selective schools using the tier categories in the Report Card data, which includes the "Ivy Plus" group (the eight Ivy League schools plus MIT, Stanford, Chicago, and Duke); "other elite" (examples include Georgetown, CMU, and the University of Virginia); and "highly selective" group (examples include the University of Michigan and Boston University). About 12% of the sample attends a highly selective school.

3.2 Teacher Value-Added Measures

[10] An alternative approach would be to use the average SAT or ACT scores of entrants to the college (e.g., Hoxby, 2009). However, this would be unable to capture important margins such as college enrollment and between selective and non-selective (i.e., do not require SAT or ACT scores) colleges.

[11] Some colleges are not separately identifiable in the tax data used by Chetty and colleagues (2017) and are aggregated into a single unit. In our data, this is most common among public 4-year universities in Massachusetts. Students attending University of Massachusetts – Amherst, University of Massachusetts – Boston, University of Massachusetts – Dartmouth, and University of Massachusetts – Lowell are combined in our earnings and mobility measures.

[12] Chetty et al. (2017) use earnings data from 2014, which is closest to what the prior 12 month earning measure reported in the 2015 ACS. Because there are only 145 people with less than a high school education in the 1980-82 birth cohorts in that sample, we pool data from the 2011-2015 ACS. The resulting sample has 761 people with less than a high school education. The median earnings were $11372 in 2015.

Estimation of Teacher Value-Added

We estimate teacher value-added using all available data on students on Grades 7, 8, 10 (2012–2019) in a first step. We estimated a standard one-step value-added that includes student, classroom, and school covariates:

$$Y_i = X_i\beta + \mu_j + \epsilon_i, \tag{1}$$

where Y_i represents either test scores or the nontest factor. The control vector X_i includes student race/ethnicity, gender, free and reduced-price lunch status, participation in special education or English learner programs, cubic polynomials of prior math and ELA standardized test scores, lagged suspensions, grades, absences, and grade progression, the second lag of each of these outcome variables, and the classroom and school-grade-year means of each of these covariates.[13] We estimate Equation 1 separately by subject and grade level. Several random assignment experiments and quasi-experimental validations have found that value-added models similar to Equation 1 provide nearly unbiased forecasts of teacher effectiveness in subsequent school years (Bacher-Hicks et al., 2019; Chetty et al., 2014a; Kane & Staiger, 2008). But other studies have identified potential sources of bias from student tracking (Backes & Hansen, 2018; Jackson, 2014; Opper, 2019; Rothstein, 2010). Because our focus is on the effects of assignment to teachers with particular skills (i.e., test or nontest value-added), and not on the individual teacher estimates, we defer a comprehensive discussion of identifying assumptions to the next section.

We then formed leave-out empirical Bayes predictions of teacher quality to use as regressors following prior work (Chetty et al., 2014a).[14] In particular, we first constructed

[13] Students are not tested in math or ELA in ninth grade in Massachusetts. For students in 10th grade, the lagged achievement data are from seventh and eighth grade. The second lag of student GPA is not available for students in seventh grade, so we include only one lag of the grade variable.

[14] We use the Stata program by Stepner (2013) to estimate the teacher value-added.

student residuals based on Equation 1 and then estimated a teacher-year effect by averaging the student residuals:

$$\hat{\mu}_{jt} = \sum_{i:j(i,t)=j}\left(Y_i - X_i\hat{\beta}\right)/n_{jt}.$$

We then constructed a leave-out estimate of teacher quality in year *t* by taking a weighted average of the teacher effects in other years (both before and after year *t*)

$$\hat{\theta}_{j,-t} = \Omega_{jt}\hat{\mu}_{j-t},$$

where $\hat{\mu}_{j-t}$ is the vector of teacher-year means in years other than *t* and Ω_{jt} is a vector of weights (Chetty et al., 2014a; Stepner, 2013). The resulting prediction $\hat{\theta}_{j,-t}$ is a leave-out estimate of teacher quality in year *t* based on data from other school years, which we obtain for both test and nontest value-added. We use these empirical Bayes estimates of teacher quality as regressors because the shrinkage factor approximates the attenuation bias resulting from the use of noisy estimates of teacher quality in place of its true value. The empirical Bayes estimator shrinks the estimated teacher effects by an estimate of the attenuation bias, an approach similar to the heteroskedastic measurement error model considered by Sullivan (2001) for bivariate regression (Jacob & Lefgren, 2008).

4. Empirical Methods

4.1 Statistical Model

We use a subset of the contemporaneous outcomes (test scores and the nontest factor) to estimate teacher value-added as described in Section 3.2 above and then use these estimates of test and nontest value-added to assess the long-run effects of teachers and how they operate through different contemporaneous student outcomes (test and nontest). In other words, our objective is to understand how assignment to specific teachers with skill measures $(\hat{\theta}_{j,-t}^{test}, \hat{\theta}_{j,-t}^{nontest})$ affects student outcomes. Following prior studies of teacher effects on longer run academic outcomes, we rely primarily on a selection on observables design (Jackson, 2018; Liu & Loeb, 2021). The sample includes several short- and long-run outcomes for students in Grades 7, 8, and 10. The statistical model is

$$Y_{ijst} = X_{it}\beta + \hat{\theta}_{j,-t}\,\delta + \alpha_{st} + \epsilon_{ijt} \tag{2}$$

where X_{it} contains cubic functions of prior scores in math and ELA, prior grade retention, prior absences, prior days suspended, prior GPA, the second lag of each of these variables,[15] demographic information consisting of limited English proficiency, race, ethnicity, gender, free and reduced-price lunch status, and special education, and the mean of each of these values at the grade and school levels. In addition, $\hat{\theta}_{j,-t}$ is a vector containing predicted teacher effects on each of test and nontest outcomes, and α_{st} represents school-subject-grade-year (or track-year) fixed effects. We cluster standard errors at the school level in all models. The key identifying assumption is that student unobservables are not correlated with measured teacher quality, $\hat{\theta}_{j,-t}$, conditional on school effects and the control vector.

We include school-subject-grade-year fixed effects to mitigate concerns about the independent effects that schools and teachers can have on student outcomes. Several studies have documented the fact that schools do affect student outcomes and that these effects are correlated across different kinds of outcomes (Jackson et al., 2020). Although some of the variation in school quality appears to be driven by differences in teacher quality, the variation in teacher effectiveness across schools does not appear sufficient to explain the full effect of schools (Mansfield, 2015). The primary concern in this case is that failure to account for school effects might bias estimates of the effects of teacher quality by conflating school and teacher effects.

[15] We do not include a second lag of GPA, which is not available for students in seventh grade.

The direction of the bias is unclear a priori, particularly if teachers and schools differ in the correlation in effects across outcomes.

In addition to concerns about school effects, prior studies have raised potential concerns about within-school sorting of students and teachers. Rothstein (2010) finds that current teaching assignments predict student's *past* achievement, which is evidence of sorting of students to teachers based on unobservable determinants of achievement. In the current application, the Rothstein critique would suggest that findings could be driven by persistent assignment biases; that is, the error term in the value-added regression Equation 1 and the long-run regression in Equation 2 are correlated. However, in another paper, Rothstein (2009) finds that controlling for multiple lags of student outcomes can remove much of the implied bias in value-added models under several plausible student tracking policies. In our application, we include two lags of each of the outcome variables to mitigate potential bias from student sorting. Moreover, Koedel and Betts (2011) find that teacher assignments are not persistent over time and that the Rothstein (2009) critique is less applicable when teacher value-added is estimated across multiple years. These findings suggest that correlation in error terms across classrooms may not be a serious concern.

At the high school level, Jackson (2014) has found that models such as Equation 2 overstate the importance of teacher quality because they fail to account for educational inputs that are bundled with student "track" assignments. Backes and Hansen (2018) find similar results for value-added to nontest outcomes, which exhibits greater bias at higher grade levels where tracking is more common. Similarly, Opper (2019) has found teachers influence students outside their classrooms through peer-to-peer spillover effects. We address these concerns by estimating models that replace the school-subject-grade-year effects in Equation 2 with school-subject-track-year effects. We follow Jackson (2014) and construct track identifiers using the 10 most common courses in each grade level. Because the Massachusetts transcript data uses a standardized course coding system, we can construct the track identifiers in each school. We assign students to a track based on their participation in each of the 10 most common courses, their school, and their grade level. The courses and their enrollment rates are listed in Appendix Table A1. The track assignment is relatively straightforward in high school where course names reliably differentiate the content area of the class. However, middle schools may offer multiple sections of courses aligned to the state core curriculum that are nonetheless tracked by student achievement. We therefore supplement the track indicators for indicators for whether a student took any of the following courses: an advanced math class, an art elective, a foreign language, an English as a Second Language (ESL) class, or a supplemental or tutorial class.[16] The inclusion of track effects weakens the identifying assumptions described above. We can relax this assumption to conditional exogeneity of teacher skill measures based on the set of courses in which a student enrolls rather than just the school and grade.

A final concern is that the teaching effectiveness measures $\hat{\theta}_{j,-t}$ are estimated and not known. The use of the leave-out empirical Bayes estimates in place of the estimated teacher fixed effects avoids a mechanical endogeneity: students who have exceptionally high achievement conditional on covariates are also likely to have unexpectedly strong postsecondary outcomes. Therefore, using the same set of students to estimate teacher value-added and the effects on long-run outcomes would likely result in an upward bias in the estimated coefficient on teacher value-added in the long-run outcome. We follow the convention in the value-added

[16] We differentiate core art classes (e.g., "Grade 8 Art") from art electives using the SCED codes assigned to the class. Art electives are typically courses like "chorus" or "drama." We similarly define ESL and supplemental/tutorial classes by the SCED code. In Appendix A, we show enrollment and student characteristics for courses that enroll at least 5% of students in each grade.

literature and use leave-out estimates of teacher effects that omit data from the current school year when assessing long-run teacher effects (Chetty et al., 2014b). Because the leave-out estimates do not use the same students in the estimation of value-added and the estimation of the effects of teaching effectiveness on long-run outcomes, we avoid this issue.

We also estimate quantile effects of teaching quality for continuous outcomes using the unconditional quantile regression of Firpo and colleagues (2009).[17] We use specifications similar to Equation 2 for each decile of the outcome distribution. The estimated coefficients δ_q provide the partial effect of teacher skills on the *q*th quantile of the outcome distribution.[18]

4.2 Assessing the Plausibility of the Research Design

Our key identifying assumption is that students are not sorted to teachers with varying test or nontest value-added within tracks based on unobserved determinants of future educational outcomes. Prior research on the effects of teachers has found that such selection on observables designs yield results similar to experimental or quasi-experimental assessments of teacher effects (Chetty et al., 2014a). Nonetheless, this remains a strong assumption. To assess the plausibility of our research design, we first examine the sorting of teachers to students based on baseline outcomes within schools and tracks. We regress each of the baseline student skill measures Y_{ijst-1} on teacher value-added and either school-subject-grade-year or track-grade-year fixed effects without other covariates:

$$Y_{ijst-1} = \hat{\theta}_{j,-t}\,\delta + \alpha_{st} + \epsilon_{ijt} \tag{4}$$

[17] We estimated quantile regressions using the Stata package written by Rios-Avila (2020). We used the RIF-OLS method discussed by Firpo and colleagues (2009), which relies on a linear probability model to estimate the relationship between the independent variable and quantiles of the outcome variable. We used a clustered bootstrap with 200 iterations to estimate standard errors clustered by student and teacher (Cameron et al., 2011).

[18] Standard quantile regression, which is conceptually similar, provides the partial effect of teaching skills on the *q*th quantile of the conditional distribution of outcomes on the included covariates. In the context of value-added models like those described in Equation 2, the conditional quantile regression estimates an effect that is interpretable as an effect on the distribution of student growth. We focus in this paper on effects on distribution of outcomes.

The coefficients δ provide a sense of the direction of unconditional sorting within schools.

In Table 3, we uncover consistent evidence that teachers are non-randomly assigned to students even within the same academic track. Of course, each of the baseline skill measures is included in our regressions, so this fact alone does not necessarily imply bias in our results. In even-numbered columns, therefore, we include student skills in year *t-2* (middle school) or *t-3* (high school) as controls in Equation 4, along with student demographic information and class-level averages. Conditional on prior student outcomes, demographics, and classroom controls, we find little evidence of sorting of students to teachers based on additional lags of the outcome variables, whether when using school fixed-effects models (Columns 2 and 6) or track fixed-effects models (Columns 4 and 8).

We also regress the other-subject teacher value-added on the same variables. We do find positive relationships between same-type value-added (e.g., teacher test value-added and the test value-added of the other-subject teacher of that student). We discuss this further in Appendix C. In addition, in Appendix C, we investigate the robustness of the main results to alternative specifications, including a quasi-experimental design developed by Chetty and colleagues (2014). We generally find that results are robust to a variety of alternative specifications.

5. Statistical Properties of Test-Based and Nontest Value-Added

In Table 4, we consider the relationship between the measures of teaching effectiveness and contemporaneous student outcomes. As in Table 3, we present results for two sets of fixed effects: one at the school level and one at the track level. In Table 4, results are similar across both specifications. A one standard deviation increase in test-based value-added is associated with an increase in test scores of 0.13-0.14 standard deviations; this is consistent in magnitude with prior work (e.g., Chetty et al., 2014a; Hanushek and Rivken, 2010). However, test value-added is associated with very little change in nontest outcomes (Panel B, Columns 1 and 2).

Likewise, nontest value-added primarily affects students' nontest outcomes (Panels A and B, Columns 3 and 4) and not test scores.[19] A one standard deviation in nontest value-added is associated with about a 0.12 standard deviation increase in the student nontest factor. While this estimate is larger than Jackson's (2018) estimate of 0.06, Jackson also finds a smaller relationship between test value-added and student achievement (0.07) than is commonly found in the literature, perhaps because the variance of teacher effects tends to be smaller in high school. When examining the components of the nontest factor individually, increases in the nontest factor are most related to GPA and grade progression, with minimal estimated effects on absences and suspensions.[20]

The correlation between test and nontest value-added for a given teacher in a given year is 0.15. This happens to be precisely the correlation reported in both Petek and Pope's (2021) study using elementary school teachers and Jackson's (2018) study of ninth grade teachers.[21] Thus, our findings add to a growing body of evidence that test-based and nontest value-added are positively correlated but also largely capture different facets of teacher skill. We display a scatterplot of teacher-level test and nontest value-added in Figure 1.

[19] In Appendix Table B1, we display results with performance ratings from teacher evaluations as well as licensure test scores in place of student outcomes. Test and nontest value-added are each individually predictive of overall performance ratings; however, the relationship is stronger for test value-added than nontest value-added. The relationship between test value-added and performance ratings are especially strong for curriculum planning and teaching all students (i.e., creating a respectful environment for students from diverse backgrounds). In addition, test value-added is much more strongly related to subject matter knowledge than nontest value-added.

[20] The relationship between the nontest factor and student absences in a given year is similar when replacing absences with unexcused absences on the lefthand side. We use absences in the primary specifications because of unexplained fluctuations over time in the average number of unexcused absences. Using total absences (rather than unexcused absences only) is common in the nontest value-added literature because unexcused absences often is not present in available administrative data (e.g., Jackson, 2018 using North Carolina data).

[21] To facilitate comparison with Jackson (2018), we report the raw correlations (i.e., not corrected for measurement error). The correlation corrected for measurement error is 0.16.

6. Teacher Value-Added and Long-Run Student Outcomes

6.1. Intermediate Secondary Outcomes

Before examining the relationship between test and nontest value-added in postsecondary outcomes, we first consider the effects of test and nontest value-added on a range of student outcomes at the secondary (high school) level in order to benchmark our results against prior studies. Results are shown in Table 5. As in prior studies, results are mixed in terms of whether test or nontest value-added carries a stronger relationship to later student outcomes. Consistent with prior work (e.g., Jackson, 2018; Liu & Loeb, 2021), we find that teacher nontest value-added is predictive of high school graduation whereas test-based value-added is not. In particular, we find that a one standard deviation change in nontest value-added is associated with a 0.54 percentage point increase in high school graduation. This is in the range of prior studies, with Rose and colleagues (2022) finding an impact of 0.20 percentage points, Liu and Loeb (2021) finding an impact of 0.70 percentage points, Gilraine and Pope (2021) reporting 0.83 percentage points, and Jackson (2018) reporting 1.5 percentage points. We also find that students in classrooms with higher test-based value-added tend to take more AP credits (0.13 credits per teacher standard deviation) and to pass more AP tests (0.04 tests passed), with no relationship between nontest value-added and the AP outcomes. This is consistent with Liu and Loeb (2021), though we find larger differences between test and nontest value-added. Finally, like Petek and Pope (2021) and Gilraine and Pope (2021), we find that nontest value-added is more predictive of whether a student takes the SAT; and like Jackson (2018) and Petek and Pope (2021), we find that test value-added is more predictive of SAT scores.

Overall, the results in Table 5 paint a remarkably clear picture. For the binary outcomes (dropping out, high school graduation, and SAT test-taking) more relevant for students in the middle or bottom of the test achievement distribution, we find larger effects for nontest value-added. For continuous results more sensitive to the top of the achievement distribution, we find larger effects for test-based value-added (AP credits, AP tests taken and passed, and SAT scores). We further explore these distributional patterns in Section 7 below.

6.2. Long-run Postsecondary outcomes

Table 6 displays results for long-run postsecondary outcomes. Three notable findings stand out here. First, there are large differences between the regressions with school fixed effects and the regressions with school-track fixed effects.[22] For example, the point estimate for the relationship between a one standard deviation increase in nontest value-added and college enrollment is 1.10 percentage points in the school fixed-effects model (Column 3), but only 0.67 percentage points in the track fixed effects-model (Column 4). This suggests that, as demonstrated in Jackson (2014), models with school fixed effects, but not track fixed effects, conflated differences in teacher quality with the sorting of students and teachers into different tracks. Second, test value-added (Column 2) and nontest value-added (Column 4) each independently predict later college quality (Panel D). And finally, when combined into the same regression, test and nontest value-added continue to each have predictive power of similar magnitude for college quality (Column 6).

While test and nontest value-added each predict college quality in Panel D, the remaining panels suggest that they do so through different mechanisms. For example, nontest value-added is more strongly related to college enrollment and whether a student enrolls in a 4-year college (Panels A and B). In contrast, test value-added is more strongly related to whether a student enrolls in a selective college (Panel C). Importantly, unlike binary measures like high school

[22] This is in contrast to the findings for the short-run outcomes in Table 4, which are largely insensitive to the inclusion of school fixed effects.

completion or college enrollment, the college quality measure used in Panel D is sensitive to changes along the outcome distribution by capturing both the college-going margin (by imputing average earnings of non-college-attenders) and the college quality margin.

Our test-based teacher value-added findings are generally similar to those reported by Chetty and colleagues (2014b), with the exception of our results not finding any relationship between test-based value-added and the college-going margin. Specifically, we find that having a teacher with one standard deviation higher test-based value-added is associated with a 0.63 percentage point increase in the likelihood of attending a selective college.[23] To put the magnitude of the findings into perspective, Chetty and colleagues (2014b) find that the difference in impact on earnings operating through test-based value-added between a fifth percentile teacher and an average teacher amounts to about $250,000 in lifetime earnings. This large amount is driven by two factors that amplify the impact of teachers: each teacher reaches many students at once, and the impact on students lasts through the entirety of their eventual adult working lives.

7. Heterogeneous Effects of Test and Nontest Value-Added

The results in Tables 5 and 6 suggest that, at least for some outcomes, students are differentially affected by different dimensions of teacher quality as measured by teacher test-based and nontest value-added. In this section, we investigate two factors that play a role in the extent to which value-added affects long-run outcomes. The first factor is the strength of the relationship between short-run student outcomes and a given long-run outcome. In particular, teachers could have an effect on a given short-run outcome, but if that short-run outcome has little relationship with a given long-run student outcome of interest, the estimates of the effect of

[23] Chetty and colleagues (2014b), by comparison, find an effect size of 0.72.

teachers on that long-run outcome will be small. Second, we describe the strength of the relationship between a given measure of teacher quality (test or nontest value-added) and short-run student outcomes for students in different quantiles of the student outcome distribution; this explores whether differential results could be explained by different teaching skills having larger or smaller impacts for different subgroups of students.

7.1. Student Short-Run Versus Long-Run Outcomes

We explore the relationship between short-term student test scores and nontest factors and longer term outcomes in Figure 2. In each portion of Figure 2, a given explanatory factor (test scores or the nontest factor) is divided into 100 equal-sized bins, and an average outcome for each bin is calculated. Beginning with high school graduation in the upper left, once a certain test score threshold is reached, further increases do very little to boost graduation because the rate is already so high (i.e., the slope is very flat for a good portion of the distribution). For the nontest factor, however, we see a very different pattern. Namely, the slope of the line is very steep through the middle of the distribution, suggesting that modest improvements to nontest outcomes for these students translate to differences in later high school graduation. This basic pattern may explain why prior work has tended to find larger relationships between test value-added and later high school graduation than test value-added and high school graduation. A similar pattern emerges for college attendance in the upper right and attending a four-year college in the middle right. For selective college-going and college quality, however, increases in test scores at the top of the distribution do translate to substantial increases in the outcomes of interest. This provides an intuitive explanation for the value-added findings above: certain long-run outcomes appear to be more sensitive to changes in short-term test versus nontest measures in different parts of the distribution.[24]

We elaborate on the above findings in Table 7, where we report the results from regressing long-run student outcomes on students' average test scores, the behavioral factor, and each of the separate components that constitute the behavioral factor. Here we set aside teachers entirely and focus solely on the relationship between student short-run and long-run outcomes and explore mechanisms for the relationships described in Section 6.

Beginning with attending college in the upper left of Table 7, where we control for both test scores and the nontest factor in Column 1, the coefficient is larger on the nontest factor. Again, this is consistent with the results in Table 5. In addition, comparing the separate regressions with test scores only (Column 2) to the nontest factor only (Column 3), the R-squared is substantially higher for the nontest factor than for test scores and is close to the R-squared for the combined regression in column 1. In other words, these results show that the nontest factor has far greater predictive power than test scores when it comes to predicting the college-going margin.

This pattern does not hold when looking at college quality (rather than college-going on its own). The R-squared in Columns 2 and 3 are similar; both are much smaller than the combined Column 1 R-squared. This suggests that student test scores have more signal for the *type* of college attended rather than the college-going margin. Meanwhile, nontest outcomes predicted both the college-going margin and the likelihood of attending a 2-year college. The

[24] The nonlinearity displayed in Figure 2 is between *student* short-term outcomes and *student* long-run outcomes. If there were a similar pattern between teacher value-added and student long-run outcomes, it may not be appropriate to rely on linear probability models for regressions of binary outcomes such as high school graduation on teacher value added, as done here and in prior literature. However, in results available from authors, replacing student outcomes in the x-axis of Figure 2 with teacher value-added yields slopes that suggest a linear relationship.

remaining panels for college-going exhibit a similar pattern to the postsecondary group earnings panel: test and nontest outcomes both predict these later outcomes. Combining the two measures produces the best predictions. Finally, for high school graduation and dropping out, the R-squared is generally much lower, but the nontest factor offers more explanatory power than test scores.

7.2. Differential Test-Based and Nontest Teacher Value-Added Effects

To explore whether the results can be explained in part by the effects of different teachers' skills for students in different parts of the outcomes distribution, we estimate quantile effects of teaching quality discussed in Section 4.3 using each decile of the outcomes distribution to estimate the partial effect of teacher skills on quantiles of the outcomes distribution. Results for the short-term test and the nontest factors are shown in Figure 3. As shown in Panel A, at every point in the test score distribution, test value-added is estimated to have a larger impact on test scores; the same is true for nontest value-added for the behavioral factor in Panel B. In addition, as discussed above and found in prior work, the cross-skill relationship (i.e., nontest value-added on test scores) is extremely weak, suggesting that test and nontest value-added capture different facets of teacher skill.

In addition, we see differences between test and nontest value-added regarding which types of students are most affected by differences in teacher quality. In particular, the estimated impact of value-added on test scores is largest for students at the top of the test score distribution, while for nontest value-added, its impact on the nontest factor is largest for students at the bottom. The impact of the nontest factor being largest for students at the bottom of the distribution is consistent with recent evidence from Jackson et al. (2022), who find that high schools' impacts on students not captured by test scores tend to be largest for less advantaged students. One possible explanation for this pattern is the greater variation across students in absences, suspensions, and grade repetition at the bottom of the distribution due to all of these being close to zero for many students near the top of the distribution.[25]

Finally, we examine the relationship between college quality and both types of value-added in Figure 4. For test value-added, we again see the largest impacts at the top of the distribution. For nontest value-added, effects are concentrated in the 30th to 60th percentiles. This is a possible explanation for the differences that prior work has found, with outcomes like high school graduation being more affected by nontest value-added and others like AP test taking by test value-added.

8. Discussion

Teachers' test-based and nontest value-added both play important and explanatory roles for long-run student outcomes, including whether and where a student enrolls in college. Although the effect size for college quality is similar for test-based and nontest value-added, this finding masks important differences in mechanisms and distributional effects. Nontest value-added primarily works through the margins of college attendance and the 2-year versus 4-year decision, while test-based value-added is more relevant at the top of the distribution for outcomes, such as whether a student attends a selective college. An important takeaway is that the different types of value-added are more relevant for different subsets of students, and the teacher skills these measures capture are only weakly correlated for a given teacher (the correlation between test and nontest value-added for a given teacher in a given year is 0.15).

[25] Jackson et al. (2022) conduct a formal test for what they term "mechanical heterogeneity" between marginal effects and baseline probabilities in which they examine the relationship between a quantile group effect and the distance of the group mean from 0.5 (see Appendix B in Jackson et al., 2022). They find strong evidence of mechanical heterogeneity for student outcomes underlying the behavioral factor such as being chronically absent and ever being suspended; i.e., effects are largest for quantiles whose means are further away from 0.5. In results available from authors, we find similar evidence for chronic absences (10 or more absences in a given year) and whether a student was suspended in a given year.

These results also suggest that focusing on test or nontest value-added in isolation likely misses key contributions that teachers make to student learning. Moreover, the unique contributions that teachers make to different student outcomes may be relevant both for thinking about the equity of teachers across students (Goldhaber et al., 2017) and for thinking about teacher assignments. For example, teachers with high test value-added being assigned to high-performing students does not necessarily create greater inequality if teachers with high nontest value-added are simultaneously disproportionately assigned to lower-performing students because this latter group of students appears to benefit more from high nontest value-added teachers. Thus, the literature that measures teacher quality gaps only along one dimension (test score value-added) may miss other important considerations.

Finally, these results provide some evidence about how interventions to improve teacher quality in the short term might influence student outcomes in the long term. For example, an intervention that improves teacher value-added to test outcomes should ultimately improve student AP attainment, SAT scores, and college quality, while an intervention that improves nontest value-added should improve student SAT participation, high school graduation, and college attendance and quality. Understanding these relationships also helps quantify the potential "scope for change" of various teacher policies (e.g., licensure and assignment policies) for downstream outcomes for different subsets of students.

References

Bacher-Hicks, A., Billings, S., & Deming, D. (2019). *The school to prison pipeline: Long-run impacts of school suspensions on adult crime* (Working paper no. 26257). National Bureau of Economic Research. https://doi.org/10.3386/w26257

Bacher-Hicks, A., Kane, T., & Staiger, D. (2014). *Validating teacher effect estimates using changes in teacher assignments in Los Angeles* (Working paper no. w20657). National Bureau of Economic Research. https://doi.org/10.3386/w20657

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review*, *73*, 101919. https://doi.org/10.1016/j.econedurev.2019.101919

Backes, B., Cowan, J., Goldhaber, D., Koedel, C., Miller, L. C., & Xu, Z. (2018). The common core conundrum: To what extent should we worry that changes to assessments will affect test-based measures of teacher performance? *Economics of Education Review*, 62, 48–65.

Backes, B., & Hansen, M. (2018). The impact of Teach for America on non-test academic outcomes. *Education Finance and Policy*, *13*(2), 168–193. https://doi.org/10.1162/edfp_a_00231

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis*, *39*(1), 146–170.

Blazar, D., & Pollard, C. (2022). *Challenges and tradeoffs of "good" teaching: The pursuit of multiple educational outcomes* (EdWorkingPaper 22-591). Annenberg Institute at Brown University.

Chamberlain, G. E. (2013). Predictive effects of teachers and schools on test scores, college attendance, and earnings. Proceedings of the National Academy of Sciences, 110(43), 17176–17182.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632. https://doi.org/10.1257/aer.104.9.2593

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–2679. https://doi.org/10.1257/aer.104.9.2633

Chetty, R., Friedman, J. N., Saez, E., Turner, N., & Yagan, D. (2017). *Mobility report cards: The role of colleges in intergenerational mobility* (Working paper no. 23618). National Bureau of Economic Research.

Cowan, J., Goldhaber, D., Jin, Z., & Theobald, R. (2020). Teacher Licensure Tests: Barrier or Predictive Tool? Working Paper No. 245-1020. National Center for Analysis of Longitudinal Data in Education Research (CALDER).

Danielson, C. & Ferguson, R. (2014). How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching. In T. J. Kane, K. A. Kerr, & R. C.

Pianta (Eds.), *Designing teacher evaluation systems: New guidance form the Measures of Effective Teaching project* (pp. 98–143). Jossey-Bass.

Firpo, S., Fortin, N. M., & Lemieux, T. (2009). Unconditional quantile regressions. Econometrica, 77(3), 953-973.

Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. Education Finance and Policy, 11(2), 125–149.

Gilraine, M., & Pope, N. G. (2021). *Making teaching last: Long-run value-added* (Working paper no. 29555). National Bureau of Economic Research. https://doi.org/10.3386/w29555

Goldhaber, D. (2007). Everyone's doing it, but what does teacher testing tell us about teacher effectiveness? The Journal of Human Resources, 42(4), 765–794.

Goldhaber, D., Quince, V., & Theobald, R. (2017). Has it always been this way? Tracing the evolution of teacher quality gaps in U.S. public schools. American Educational Research Journal. https://doi.org/10.3102/0002831217733445

Hoxby, C. M. (2009). The changing selectivity of American colleges. *Journal of Economic perspectives*, *23*(4), 95-118.

Isenberg, E., Max, J., Gleason, P., Johnson, M., Deutsch, J., & Hansen, M. (2016). Do low-income students have equal access to effective teachers? Evidence from 26 districts (No. NCEE 2017-4007). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Jacob, B., & Rothstein, J. (2017). The Measurement of Student Ability in Modern Assessment Systems. Journal of Economic Perspectives, 30(3), 85-108.

Jackson, C. K. (2014). Teacher quality at the high school level: The importance of accounting for tracks. *Journal of Labor Economics*, *32*(4), 645–684. https://doi.org/10.1086/676017

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, *126*(5), 2072–2107.

Jackson, C. K., Porter, S. C., Easton, J. Q., & Kiguel, S. (2020). Who benefits from attending effective schools? Examining heterogeneity in high school impacts (No. w28194). National Bureau of Economic Research.

Jackson, C. K., Kiguel, S., Porter, S. C., & Easton, J. Q. (2022). Who Benefits From Attending Effective High Schools?

Jackson, C. K., Porter, S. C., Easton, J. Q., Blanchard, A., & Kiguel, S. (2020). School effects on socioemotional development, school-based arrests, and educational attainment. *American Economic Review: Insights*, *2*(4), 491–508. https://doi.org/10.1257/aeri.20200029

Jacob, B. A., Rockoff, J. E., Taylor, E. S., Lindy, B., & Rosen, R. (2018). Teacher applicant hiring and teacher performance: Evidence from DC public schools. Journal of Public Economics, 166, 81–97.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment [MET Project Research Paper]. Bill & Melinda Gates Foundation.

Kane, T., & Staiger, D. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (Working paper no. 14607). National Bureau of Economic Research. https://doi.org/10.3386/w14607

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. The Journal of Human Resources, 46(3), 587–613.

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. Education Finance and policy, 6(1), 18-42.

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. Journal of Human Resources, 54(1), 1–36.

Liu, J., & Loeb, S. (2021). Engaging teachers: Measuring the impact of teachers on student attendance in secondary school. *Journal of Human Resources*, *56*(2), 343–379. https://doi.org/10.3368/jhr.56.2.1216-8430R3

Mansfield, R. K. (2015). Teacher quality and student inequality. *Journal of Labor Economics*, *33*(3), 751–788. https://doi.org/10.1086/679683

Mulhern, C., & Opper, I. (2022). Measuring and Summarizing the Multiple Dimensions of Teacher Effectiveness.

Nagler, M., Piopiunik, M., & West, M. R. (2019). Weak markets, strong teachers: Recession at career start and teacher effectiveness. Journal of Labor Economics. Advance online publication.

Opper, I. M. (2019). Does helping John help Sue? Evidence of spillovers in education. *American Economic Review*, *109*(3), 1080–1115. https://doi.org/10.1257/aer.20161226

Petek, N., & Pope, N.G. (2021). *The multidimensional impacts of teachers on students*. Unpublished manuscript.

Reardon, S. F., Ho, A. D., Shear, B. R., Fahle, E. M., Kalogrides, D., Jang, H., & Chavez, B. (2021). *Stanford Education Data Archive* (Version 4.1). http://purl.stanford.edu/db586ns4974.

Rios-Avila, F. (2020). Recentered influence functions (RIFs) in Stata: RIF regression and RIF decomposition. *The Stata Journal*, *20*(1), 51–94. https://doi.org/10.1177/1536867X20909690

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. Econometrica, 73(2), 417-458.

Rose, E. K., Schellenberg, J. T., & Shem-Tov, Y. (2022). The Effects of Teacher Quality on Adult Criminal Justice Contact (No. w30274). National Bureau of Economic Research.

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

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, *125*(1), 175–214.

Rothstein, J. (2017). Measuring the impacts of teachers: Comment. *American Economic Review*, *107*(6), 1656–1684. https://doi.org/10.1257/aer.20141440

Ruggles, S., Flood, S., Foster, S., Goeken, R., Pacas, J., Schouweiler, M., & Sobek, M. (2021). *IPUMS USA: Version 11.0* [dataset]. IPUMS. https://doi.org/10.18128/D010.V11.0

Sorensen, L. C. (2019). "Big data" in educational administration: An application for predicting school dropout risk. *Educational Administration Quarterly*, 55(3), 404–446.

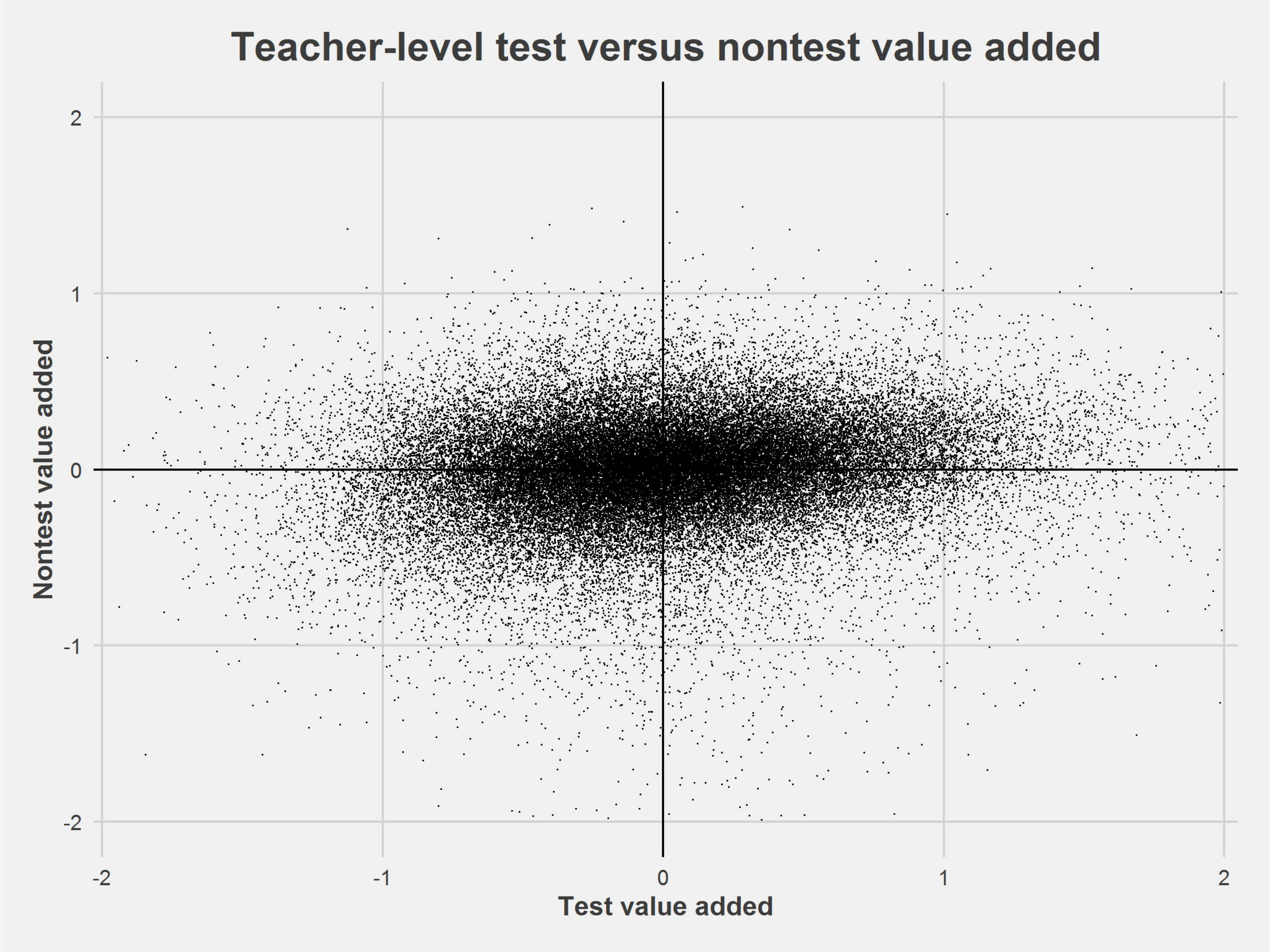
Stepner, M. (2013). VAM: Stata module to compute teacher value-added measures.

Sullivan, D. G. (2001). *A note on the estimation of linear regression models with heteroskedastic measurement errors* (No. 2001–23). Federal Reserve Bank of Chicago. https://www.chicagofed.org/publications/working-papers/2001/2001-23

Williams, W., Adrien, R., Murthy, C., & Pietryka, D. (2016). Equitable access to excellent educators: An analysis of states' educator equity plans. U.S. Department of Education, Office of Elementary and Secondary Education.

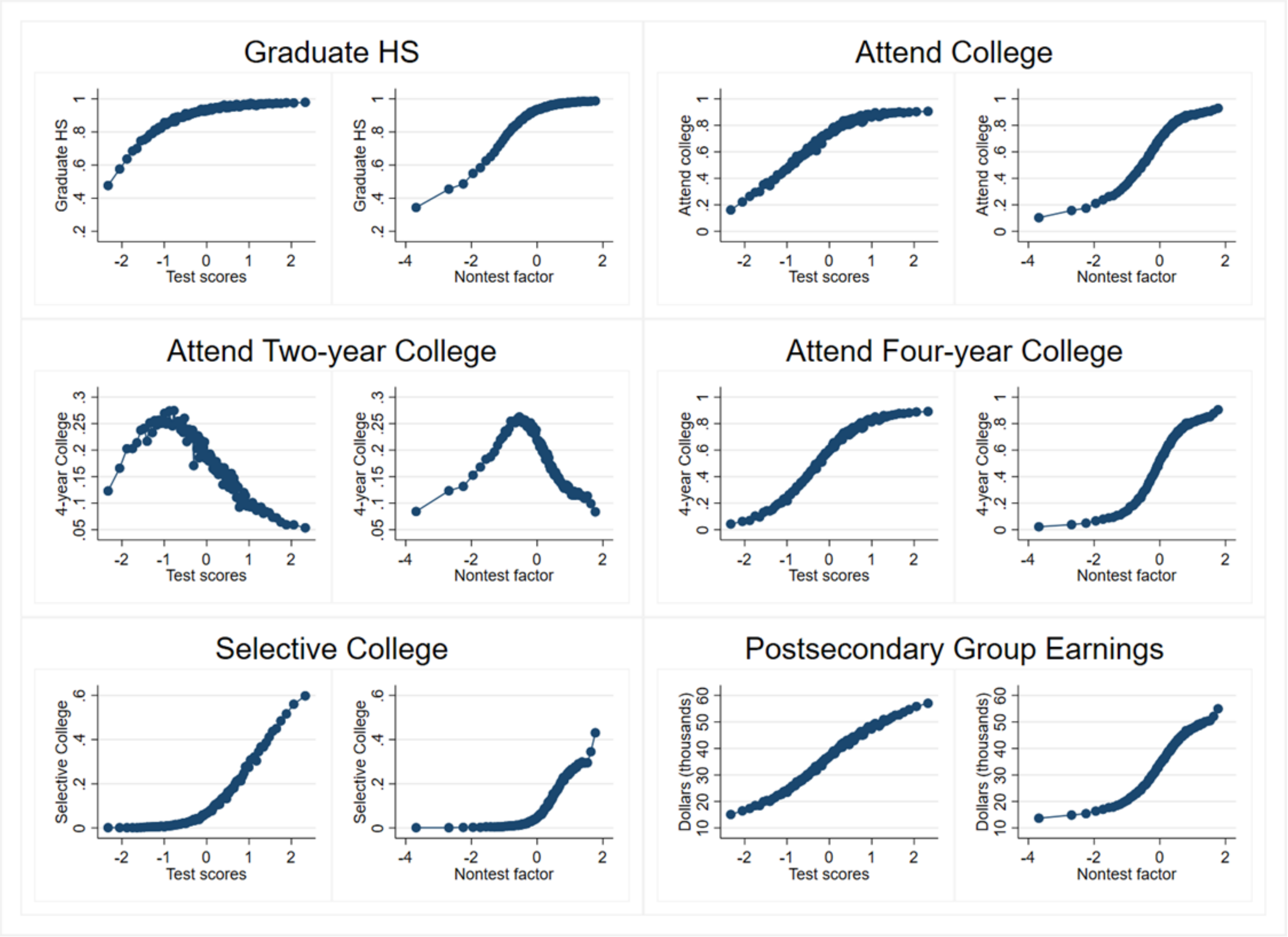
Tables and Figures

Figure 1.



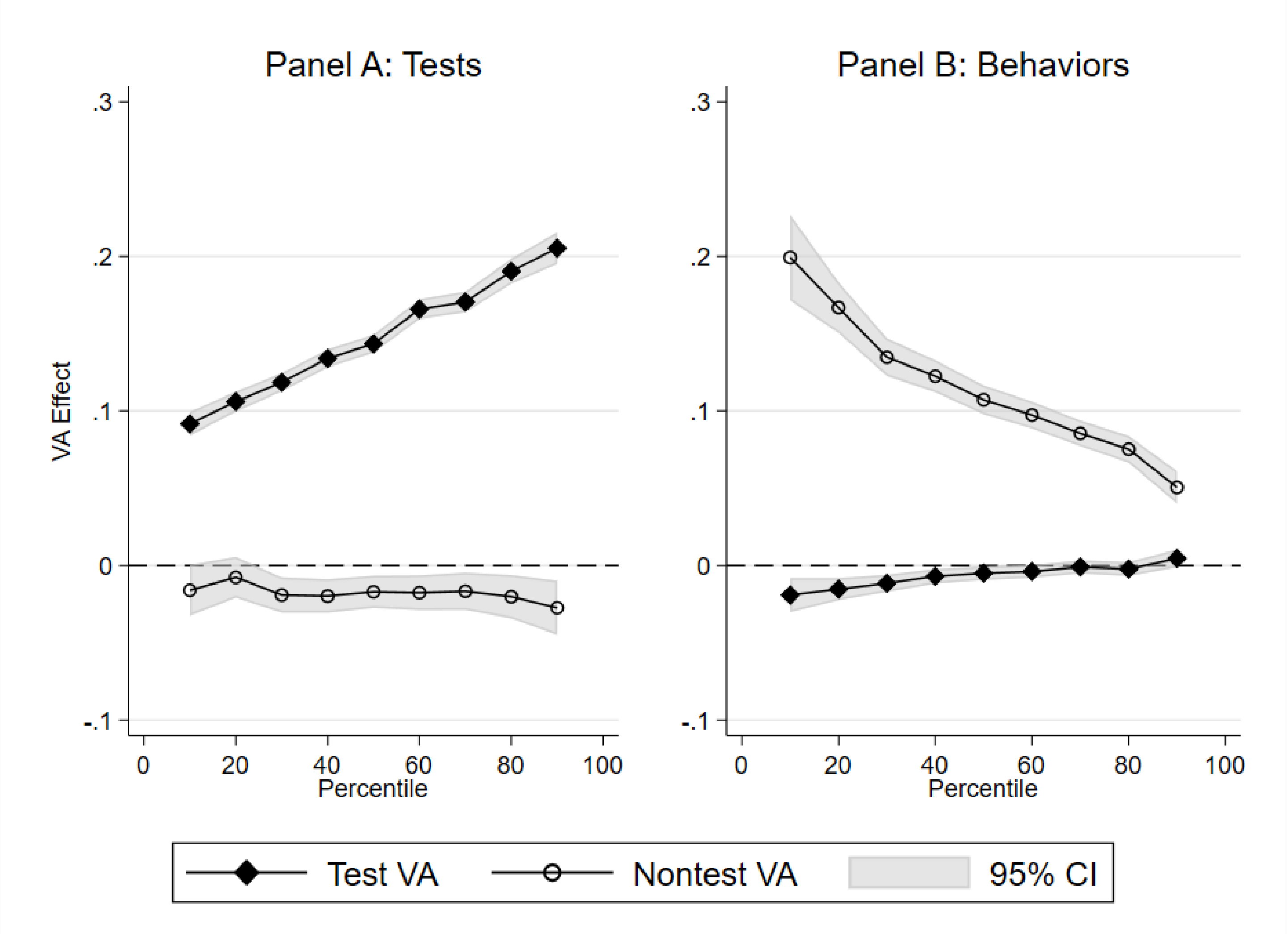
Notes: Each point represents test and nontest value-added for a given teacher in a given year. Points outside of -2 and 2 omitted for clarity.

Figure 2.



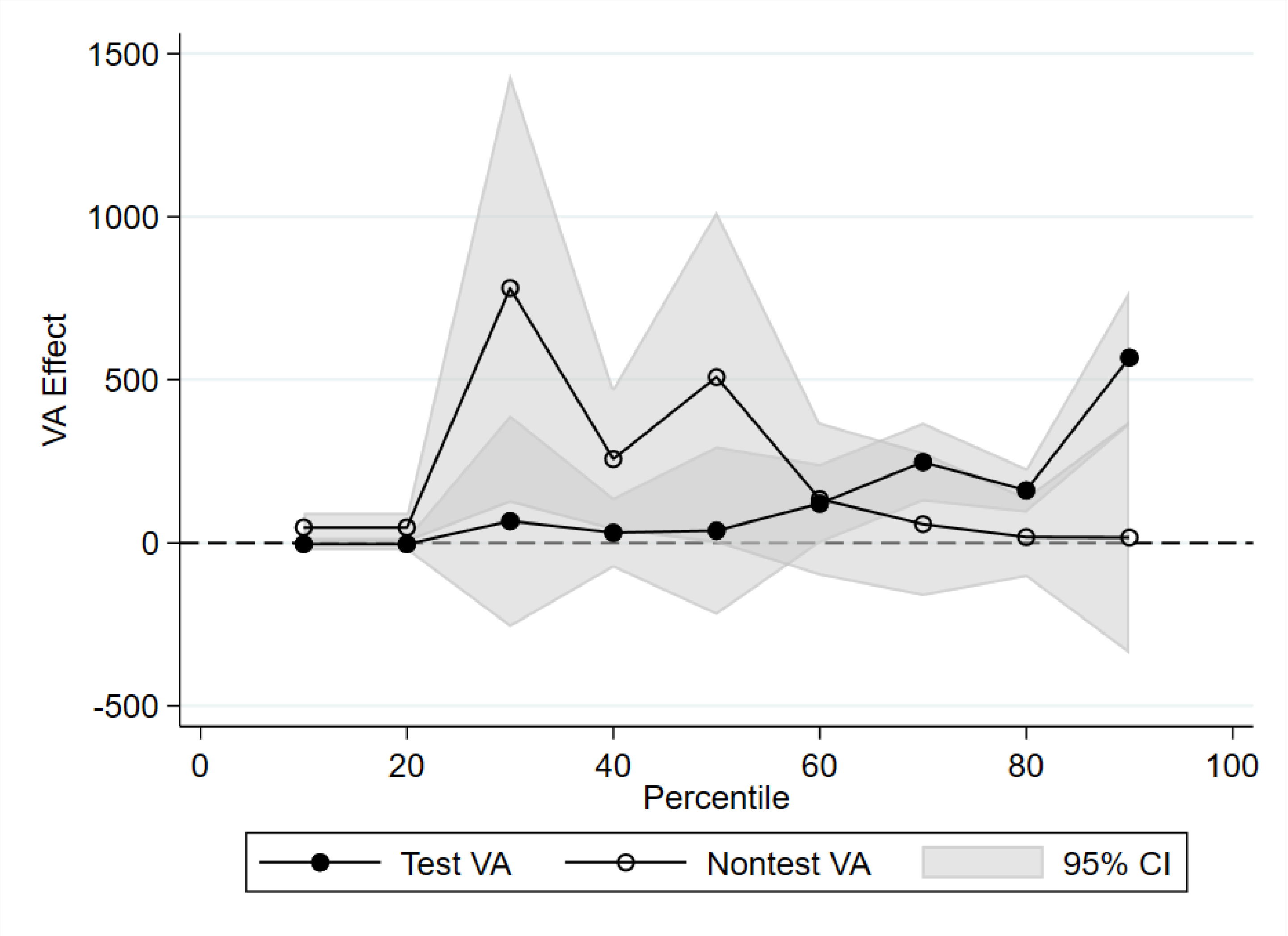
Notes: Students are divided into 100 bins based on test scores (left panels) or nontest factors (right panels). For each bin, vertical axis represents the average of a given outcome within that bin.

Figure 3. Quantile Effects on Contemporaneous Outcomes



Notes: Each point shows the estimated effect of a one standard deviation change in value-added on a given quantile of the test score (Panel A) or nontest factor (Panel B) distribution.

Figure 4. Quantile Effects on College Quality



Notes: Each point shows the estimated effect of a one standard deviation change in value-added on a given quantile of the college quality distribution.

Table 1. Prior Evidence on Value-Added and Long-Run Outcomes

			Impact of increase of one standard deviation of teacher VA	
Study	Setting	Nontest VA	Test VA	Nontest VA
Chetty et al. (2014b)	Gr. 4-8, 1989-2009, NYC	--	College enroll 0.7 pp College quality $156-266 High Q college 0.7 pp	--
Jackson (2018)	Gr. 9, 2005-2012, NC	Factor[1]	HS grad: 0.1 pp Take SAT: 1.2 pp SAT: 0.60 Intend 4yr: 0.1 pp	HS grad: 1.5 pp Take SAT: 0.1 pp SAT: -0.23 Intend 4yr: 1.3 pp
Liu and Loeb (2021)	Gr. 7-11, 1 district in CA, 2004-2014	Unexcused abs	HS grad: 0.1 pp AP courses: 0.02 AP credits: 0.11	HS grad: 0.7 pp AP courses: 0.01 AP credits: 0.08
Mulhern and Opper (2022)	Gr. 5-7, NYC, 2005-2014	Attendance, grades[3]	HS grad: 0.2 pp	HS grad: -.6 to -.8 pp
Petek and Pope (2021)	Gr. 3-5, Los Angeles USD, 2003-2015	Factor[1]	Dropout: -0.2 pp Held back: 0.1 pp Take SAT: -0.2 pp SAT: 6.3 points	Dropout: -0.3 pp Held back: -0.6 pp Take SAT: 1.0 pp SAT: 2.0 points
Gilraine and Pope (2021)	Gr. 3-5, 1 large district, 2003-2017	Factor[1]	HS grad: 0.12 pp Take SAT: 0.05 pp SAT score: 2.9 points	HS grad: 0.83 pp Take SAT: 0.33 pp SAT score: 6.59 points
Rose et al., (2022)	Gr. 4-8, NC, 1996-2013	Factor[1,2]	HS grad: 0.11 pp Arrested: -0.08 pp	HS grad: 0.20 pp Arrested: -0.36 pp

(1) Factor consists of absences, suspensions, GPA, grade progression originally developed in Jackson (2018).
(2) Rose et al. (2022) do not include GPA in factor.
(3) The measures presented in Mulhern and Opper (2022) are conditional on other test + nontest measures and are thus not directly comparable to the other studies.

Table 2. Summary Statistics

	Contemporaneous Outcomes			Long Run Outcomes		
	Mean	SD	N	Mean	SD	N
ELA test	-0.018	0.904	3031202	-0.028	0.895	1886153
Math test	-0.008	0.913	3035364	-0.017	0.901	1889123
Nontest index	0.059	0.946	2900521	0.073	0.929	1834574
Retained	0.006	0.080	3113806	0.008	0.087	1942344
Absences	8.682	10.265	3112801	8.625	10.325	1941593
Days suspended	0.247	1.897	3113806	0.253	1.974	1942344
GPA	2.961	0.901	2901193	2.904	0.898	1835050
Next year GPA	2.884	0.930	2827368	2.826	0.914	1829994
AP credits				4.273	8.426	1942344
AP tests taken				1.327	2.079	1942344
AP tests passed				0.887	1.783	1942344
Takes SAT				0.689	0.463	1578281
SAT scores (standard deviations)				0.067	1.001	1087702
Graduate				0.898	0.302	1942344
Dropout				0.034	0.182	1942344
Attends college				0.684	0.465	1942344
Attends 2-year college				0.181	0.385	1942344
Attends 4-year college				0.546	0.498	1942344
College quality				35997.441	20344.001	1942344
Lag math test	-0.010	0.918	2917855	-0.017	0.911	1814993
Lag ELA test	-0.021	0.910	2912808	-0.029	0.901	1810437
Lag retention	0.007	0.080	3038144	0.008	0.088	1900953
Lag absences	7.571	8.485	3027259	7.467	8.469	1893282
Lag days suspended	0.172	1.493	3038145	0.177	1.597	1900953
Lag GPA	3.015	0.879	2655421	2.958	0.875	1708477
Limited English proficient	0.050	0.218	3113806	0.043	0.203	1942344
Male	0.502	0.500	3113806	0.500	0.500	1942344
Free- or reduced-price lunch	0.347	0.476	3113806	0.352	0.478	1942344
Full inclusion special education	0.111	0.314	3113806	0.106	0.308	1942344
Partial inclusion special education	0.021	0.145	3113806	0.023	0.149	1942344
Substantially separate special education	0.006	0.076	3113806	0.006	0.080	1942344
Black student	0.118	0.323	3113806	0.115	0.319	1942344
Asian student	0.084	0.277	3113806	0.079	0.270	1942344
American Indian student	0.029	0.169	3113806	0.028	0.164	1942344
Pacific Islander student	0.010	0.099	3113806	0.010	0.100	1942344
Hispanic student	0.186	0.389	3113806	0.172	0.377	1942344
Takes advanced math	0.265	0.441	3113806	0.271	0.444	1942344
Takes art elective	0.280	0.449	3113806	0.219	0.414	1942344
Takes advanced language	0.105	0.306	3113806	0.143	0.350	1942344
Takes supplemental course	0.096	0.295	3113806	0.090	0.286	1942344
Takes ESL course	0.028	0.165	3113806	0.024	0.154	1942344

Table 3. Student-Teacher Sorting Within Schools and Tracks

	Test VA				Behavior VA			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Prior math score	0.130***	0.006**	0.031***	0.003	-0.174***	-0.005	-0.084***	-0.006
	(0.018)	(0.003)	(0.008)	(0.002)	(0.034)	(0.005)	(0.015)	(0.004)
Prior ELA score	0.112***	0.002	0.021***	-0.003	-0.154***	-0.003	-0.072***	-0.004
	(0.016)	(0.003)	(0.007)	(0.002)	(0.030)	(0.005)	(0.015)	(0.005)
Prior GPA	0.108***	0.001	0.028***	0.003	-0.129***	-0.004	-0.056***	-0.006
	(0.014)	(0.004)	(0.006)	(0.003)	(0.027)	(0.009)	(0.014)	(0.008)
Repeating grade	-0.001*	0.000	-0.000	-0.000	0.004***	0.003**	0.002**	0.001*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Prior absences	-0.060***	-0.007***	-0.018***	-0.004	0.036**	-0.010*	0.006	-0.008
	(0.008)	(0.003)	(0.004)	(0.003)	(0.014)	(0.006)	(0.009)	(0.005)
Prior suspensions	-0.006***	0.001	-0.002*	-0.000	0.007*	-0.002	0.006**	0.002
	(0.002)	(0.001)	(0.001)	(0.001)	(0.004)	(0.002)	(0.003)	(0.002)
Other teacher test VA	0.089***	0.085***	0.085***	0.085***	0.016	0.019	0.019	0.020
	(0.017)	(0.017)	(0.019)	(0.019)	(0.026)	(0.026)	(0.029)	(0.029)
Other teacher nontest VA	0.003	0.004	0.004	0.004	0.214***	0.213***	0.226***	0.225***
	(0.006)	(0.006)	(0.007)	(0.007)	(0.024)	(0.024)	(0.027)	(0.027)
Has long-run outcomes	-0.001	0.000	0.000	0.000	0.004**	0.003*	0.002	0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
N	1291882	1291882	1260277	1260277	1291882	1291882	1260277	1260277
Controls		Y		Y		Y		Y
School-grade-year FE	Y	Y			Y	Y		
School-track-year FE			Y	Y			Y	Y

Notes: Coefficients on test scores and nontest factor from regressions with a given outcome as the dependent variable. The sample includes all students in the matched long-run sample described in the text. All regressions include grade-by-year fixed effects.

Table 4. Teacher Effects on Students' Short-Run Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Test Scores						
Test VA	0.137***	0.130***			0.138***	0.131***
	(0.003)	(0.003)			(0.003)	(0.003)
Nontest VA			-0.005	-0.004	-0.019***	-0.016***
			(0.007)	(0.007)	(0.004)	(0.005)
Panel B. Behavioral Factor						
Test VA	-0.005*	-0.004			-0.009***	-0.007***
	(0.003)	(0.003)			(0.002)	(0.002)
Nontest VA			0.124***	0.119***	0.125***	0.120***
			(0.008)	(0.008)	(0.008)	(0.008)
Panel C. GPA						
Test VA	-0.011***	-0.009**			-0.016***	-0.013***
	(0.004)	(0.004)			(0.003)	(0.003)
Nontest VA			0.182***	0.176***	0.184***	0.177***
			(0.011)	(0.012)	(0.011)	(0.012)
Panel D. Absences						
Test VA	-0.004*	-0.003			-0.003	-0.003
	(0.002)	(0.002)			(0.002)	(0.002)
Nontest VA			-0.005	-0.003	-0.005	-0.003
			(0.004)	(0.004)	(0.004)	(0.004)
Panel E. Days Suspended						
Test VA	-0.000	-0.000			-0.000	-0.000
	(0.001)	(0.001)			(0.001)	(0.001)
Nontest VA			-0.003	-0.002	-0.003	-0.002
			(0.002)	(0.002)	(0.002)	(0.002)
Panel F. Retained						
Test VA	0.045*	0.044*			0.056**	0.052**
	(0.026)	(0.023)			(0.026)	(0.023)
Nontest VA			-0.395***	-0.312***	-0.403***	-0.316***
			(0.058)	(0.058)	(0.059)	(0.058)
N	2222044	2174082	2217606	2169847	2215947	2168275
School-Grade-Year FE	Y		Y		Y	
School-Track-Year FE		Y		Y		Y

Notes: Coefficients on test scores and nontest factor from regressions with contemporaneous student outcomes as the dependent variable. The sample includes all students in the matched long-run sample described in the text. All regressions include grade-by-year fixed effects.

Table 5. Teacher Effects on Students' Secondary Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. AP Credits						
Test VA	0.250***	0.131***			0.253***	0.134***
	(0.083)	(0.040)			(0.084)	(0.041)
Nontest VA			0.048	-0.006	0.009	-0.024
			(0.117)	(0.069)	(0.119)	(0.070)
Panel B. AP Tests Taken						
Test VA	0.080***	0.050***			0.081***	0.051***
	(0.015)	(0.009)			(0.015)	(0.009)
Nontest VA			-0.028	-0.001	-0.038	-0.007
			(0.023)	(0.016)	(0.023)	(0.016)
Panel C. AP Tests Passed						
Test VA	0.071***	0.043***			0.074***	0.044***
	(0.013)	(0.009)			(0.013)	(0.009)
Nontest VA			-0.055***	-0.015	-0.065***	-0.020
			(0.020)	(0.012)	(0.020)	(0.013)
Panel D. Took SAT						
Test VA	-0.234	-0.012			-0.270	-0.028
	(0.205)	(0.167)			(0.208)	(0.168)
Nontest VA			1.311***	0.715*	1.347***	0.731*
			(0.401)	(0.377)	(0.404)	(0.377)
Panel E. SAT Scores (standard deviations)						
Test VA	0.028***	0.017***			0.029***	0.017***
	(0.003)	(0.003)			(0.003)	(0.003)
Nontest VA			-0.015**	-0.004	-0.019***	-0.006
			(0.006)	(0.006)	(0.006)	(0.006)
Panel F. Graduate HS						
Test VA	-0.109	-0.060			-0.115	-0.065
	(0.091)	(0.098)			(0.092)	(0.099)
Nontest VA			0.521**	0.564**	0.540**	0.573**
			(0.228)	(0.251)	(0.230)	(0.253)
Panel G. Dropout						
Test VA	0.116**	0.050			0.122**	0.052
	(0.054)	(0.058)			(0.055)	(0.058)
Nontest VA			-0.262**	-0.213	-0.278**	-0.226
			(0.128)	(0.141)	(0.129)	(0.141)
N	1436422	1398764	1434198	1396658	1432748	1395292
School-Grade-Year FE	Y		Y		Y	
School-Track-Year FE		Y		Y		Y

Notes: Coefficients on test scores and nontest factor from regressions with student secondary outcomes as the dependent variable. The sample includes all students in the matched long-run sample described in the text. All regressions include grade-by-year fixed effects.

Table 6. Teacher Effects on Postsecondary Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Enroll in College						
Test VA	0.017	0.053			-0.008	0.040
	(0.153)	(0.153)			(0.153)	(0.153)
Nontest VA			1.099***	0.666**	1.103***	0.664**
			(0.334)	(0.328)	(0.333)	(0.328)
Panel B. Enroll in Four-Year College						
Test VA	-0.050	0.031			-0.091	0.016
	(0.176)	(0.149)			(0.176)	(0.150)
Nontest VA			1.532***	0.856**	1.551***	0.869**
			(0.349)	(0.336)	(0.350)	(0.338)
Panel C. Enroll in Selective College						
Test VA	0.986***	0.632***			1.019***	0.642***
	(0.159)	(0.129)			(0.161)	(0.130)
Nontest VA			-0.534**	0.135	-0.660***	0.065
			(0.245)	(0.196)	(0.251)	(0.200)
Panel D. College Quality Index						
Test VA	188.419***	181.702***			180.709***	179.698***
	(61.551)	(60.217)			(61.617)	(60.265)
Nontest VA			382.999***	231.209**	364.754***	216.027*
			(116.405)	(115.217)	(115.676)	(114.356)
N	1436422	1398764	1434198	1396658	1432748	1395292
School-Grade-Year FE	Y		Y		Y	
School-Track-Year FE		Y		Y		Y

Notes: Coefficients on test scores and nontest factor from regressions with student postsecondary outcomes as the dependent variable. The sample includes all students in the matched long-run sample described in the text. All regressions include grade-by-year fixed effects.

Table 7. Predicting Postsecondary Outcomes with Tests and Nontest Factor

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Attend College				Attend Four-Year College			
Tests	0.10	0.19			0.16	0.26		
Nontest Factor	0.17		0.22		0.18		0.26	
Retained				-0.09				0.08
Log Absences				-0.05				-0.05
Log Days Suspended				-0.06				-0.04
GPA				0.21				0.27
R Sq.	0.22	0.14	0.20	0.22	0.29	0.22	0.24	0.28
	College Quality Index							
Tests	7120.73	11054.23						
Nontest Factor	7249.65		10837.64					
Retained				2644.05				
Log Absences				-2327.07				
Log Days Suspended				-1212.51				
GPA				11179.63				
R Sq.	0.32	0.24	0.25	0.29				
	Graduate				Dropout			
Tests	0.02	0.07			-0.00	-0.03		
Nontest Factor	0.11		0.12		-0.05		-0.06	
Retained				-0.21				0.07
Log Absences				-0.04				0.02
Log Days Suspended				-0.08				0.05
GPA				0.09				-0.04
R Sq.	0.13	0.05	0.14	0.15	0.08	0.03	0.09	0.09
Observations	1785667	1888012	1834574	1834574	1785667	1888012	1834574	1834574

Notes: Coefficients on test scores and nontest factor from regressions with long-run student outcomes as the dependent variable. The sample includes all students in the matched long-run sample described in the text. All regressions include grade-by-year fixed effects.

Appendix A. Construction of Tracks

We follow Jackson (2014, 2018) and construct academic tracks using the 10 most enrolled classes in each grade level. In each case, students in each track have the same enrollment status in each of the 10 classes and the academic level of the math and ELA classes (basic, general, advanced, or postsecondary). In **Table A.1**, we show the distribution of the number of teachers in each track for both the full sample of matched students in 7th, 8th, and 10th grades as well as the restricted sample with long-run outcomes. Tracks do tend to be smaller than the school-grade cells as a whole, although most tracks do have multiple teachers. The modal number of teachers within school-grades is 5 (4 for the long-run sample); it is 4 (3 for the long-run sample) within tracks.

We show summary statistics for each of these courses in **Tables A.2** through **A.4**. The italicized courses are the 10 most popular in each grade. As shown in the tables, popular courses better differentiate students' academic ability in high school than in middle school. This is primarily because there are fewer courses in each subject in the middle school course categorization. To better use information on tracking embedded in class assignments, we construct five additional covariates used in both the value-added models and the regression analyses. The courses used to construct these indicators (among the courses with at least 5% of students enrolled) are indicated in bold in Tables A.2 through A.4. We construct an indicator for advanced math courses if students take Pre-Algebra or Algebra in 7th grade; Algebra in 8th grade; or Algebra II in 10th grade. We define advanced foreign languages if students take any foreign language in 7th or 8th grade; or if students take a third year foreign language class in 10th grade. We construct an arts elective for students in 7th or 8th grade who take an art course other than the grade-specific Art or Music course. As we show in Tables A.2 and A.3, these are mostly band, chorus, and drama courses. We additionally construct an indicator for supplemental courses for students who take either a Tutorial class or a Supplemental course. The Supplemental courses are usually offered in math. Finally, we construct an indicator for students who take an English as a Second Language (ESL) class. Not all students classified as English language learners take an ESL class, so this indicator is distinct from the limited English proficient indicator. As can be seen in Tables A.2 and A.3, some of the arts, foreign language, supplemental, and ESL classes – although not among the top 10 most enrolled – are strongly predictive of student outcomes.

Table A.1. Distribution of the Number of Teachers per Track

Number of Teachers	VA: School-Grade	VA: School-Track	LR: School-Grade	LR: School-Track
1	41583	132474	47571	124124
2	99638	260521	127583	228759
3	179880	342377	179663	288004
4	214988	363952	191459	244545
5	275275	334323	178376	219786
6	248808	298897	160429	174789
7	246521	236340	116749	141314
8	175893	201033	101256	110912
9	184247	194960	81627	90961
10	175851	154941	67217	78649
11	128839	100427	50696	55536
12	129783	92983	57657	42403
13	96656	76803	51485	38264
14	89577	66884	52181	41517
15+	826267	256891	563700	148086

Notes: Counts of teachers per school-grade or school-track cells for the value-added (2012-2019, [VA]) and long-run [LR] samples.

Table A.2. Summary Statistics by Course Enrollment (Grade 7)

Course	N	LEP	Prior ELA Score	Prior Math Score	Prior Retention	Prior Absences	Prior Days Suspended	Prior GPA	Special Education
French	50529	0.01	0.43	0.40	0.00	6.26	0.04	3.57	0.07
General Band	56214	0.03	0.30	0.35	0.00	5.69	0.05	3.49	0.09
Spanish	159304	0.01	0.27	0.26	0.00	6.48	0.06	3.47	0.08
Drama (grade 7)	48237	0.04	0.26	0.26	0.00	6.45	0.06	3.46	0.15
Foreign Language (grade 7)	91557	0.02	0.25	0.25	0.00	6.49	0.06	3.36	0.09
Family and Consumer Science—Comprehensive	33008	0.02	0.26	0.24	0.00	6.27	0.04	3.48	0.14
Pre-Algebra	145368	0.06	0.19	0.23	0.00	6.90	0.09	3.29	0.13
World Geography	60920	0.01	0.21	0.20	0.00	6.33	0.04	3.47	0.16
Pre-Engineering Technology	49371	0.03	0.17	0.17	0.00	6.80	0.06	3.36	0.15
Chorus	85350	0.03	0.22	0.13	0.00	6.75	0.05	3.43	0.13
Engineering and Technology—Other	35484	0.02	0.10	0.12	0.00	6.85	0.07	3.42	0.15
Engineering Technology	60196	0.02	0.11	0.11	0.00	6.94	0.08	3.31	0.17
Computer Applications	42167	0.02	0.02	0.01	0.00	7.04	0.07	3.30	0.16
Health Education	351108	0.06	0.01	0.01	0.00	7.18	0.12	3.26	0.16
Introduction to Computers	52026	0.03	0.04	0.01	0.00	7.04	0.08	3.23	0.16
Health and Fitness	39924	0.04	-0.02	0.00	0.00	7.78	0.13	3.34	0.15
Art (grade 7)	492507	0.06	0.00	0.00	0.00	7.28	0.12	3.26	0.16
Music (grade 7)	296091	0.06	-0.03	-0.02	0.00	7.26	0.12	3.24	0.15
Physical Education (grade 7)	652874	0.06	-0.03	-0.03	0.00	7.32	0.14	3.21	0.17
Writing (grade 7)	38072	0.03	-0.06	-0.03	0.00	7.18	0.14	3.22	0.16
Computer and Information Technology	77996	0.05	-0.02	-0.04	0.00	7.48	0.14	3.28	0.15
Language Arts (grade 7)	624103	0.05	-0.05	-0.05	0.00	7.53	0.15	3.17	0.17
Social Studies (grade 7)	501381	0.06	-0.07	-0.06	0.01	7.63	0.16	3.13	0.17
Science (grade 7)	591268	0.06	-0.07	-0.07	0.00	7.57	0.15	3.16	0.17
Technological Literacy	63938	0.06	-0.08	-0.09	0.00	7.44	0.15	3.19	0.16
Computer Literacy	67280	0.08	-0.11	-0.09	0.00	7.41	0.14	3.10	0.16
Study Skills	86406	0.04	-0.12	-0.13	0.00	7.63	0.12	3.08	0.30
World History—Overview	47541	0.12	-0.12	-0.14	0.01	7.18	0.17	3.26	0.18
Mathematics (grade 7)	507623	0.07	-0.16	-0.17	0.00	7.74	0.17	3.12	0.18
Exploratory	34426	0.07	-0.15	-0.19	0.00	7.92	0.22	3.06	0.16
Reading (grade 7)	72460	0.06	-0.29	-0.29	0.00	7.67	0.14	3.08	0.25

Grade 7	35996	0.06	-0.40	-0.42	0.00	8.16	0.21	3.02	0.47
Tutorial	73386	0.10	-0.49	-0.49	0.01	8.30	0.31	2.92	0.29
Mathematics—Supplemental	33873	0.11	-0.50	-0.54	0.01	9.13	0.33	2.78	0.22
English as a Second Language	33139	0.94	-1.40	-1.20	0.01	7.96	0.25	2.50	0.14

Notes: Summary statistics for students enrolled in courses in grade 7 (2012-2019) with enrollments of at least 5% of the total enrollment. Courses in bold are included in the course type indicators used as covariates in the regression analyses. Courses indicated in italics are used to construct academic tracks.

Table A.3. Summary Statistics by Course Enrollment (Grade 8)

Course	N	LEP	Prior ELA Score	Prior Math Score	Prior Retention	Prior Absences	Prior Days Suspended	Prior GPA	Special Education
French	45887	0.01	0.48	0.46	0.00	6.53	0.05	3.48	0.05
General Band	48437	0.03	0.32	0.40	0.00	5.79	0.06	3.44	0.08
Algebra I	193580	0.05	0.31	0.39	0.00	6.98	0.10	3.36	0.10
U.S. History—Comprehensive	33902	0.02	0.30	0.31	0.00	7.03	0.07	3.37	0.15
Foreign Language (grade 8)	87693	0.01	0.26	0.29	0.00	6.78	0.08	3.29	0.08
Spanish	178715	0.01	0.27	0.27	0.00	6.75	0.09	3.38	0.07
Drama (grade 8)	37190	0.04	0.21	0.23	0.00	7.22	0.12	3.37	0.16
Family and Consumer Science—Comprehensive	34621	0.02	0.16	0.18	0.00	6.93	0.08	3.34	0.13
Pre-Engineering Technology	52838	0.03	0.10	0.14	0.00	7.35	0.12	3.25	0.15
Chorus	77240	0.03	0.23	0.14	0.00	7.32	0.09	3.35	0.13
Introduction to Computers	40820	0.03	0.06	0.13	0.00	7.02	0.11	3.20	0.14
Engineering and Technology—Other	36815	0.02	0.09	0.12	0.00	7.45	0.12	3.30	0.15
Engineering Technology	60303	0.02	0.10	0.12	0.00	7.55	0.13	3.20	0.16
Writing (grade 8)	38593	0.03	0.04	0.07	0.00	7.77	0.21	3.20	0.15
Health Education	336912	0.05	0.04	0.05	0.00	7.58	0.16	3.18	0.16
Health and Fitness	37198	0.04	0.01	0.05	0.00	8.18	0.16	3.18	0.14
Computer and Information Technology	71086	0.05	-0.03	0.01	0.00	7.74	0.17	3.19	0.15
Art (grade 8)	471194	0.05	-0.02	0.00	0.00	7.87	0.18	3.15	0.16
Physical Education (grade 8)	648621	0.06	-0.03	-0.01	0.00	7.80	0.20	3.12	0.16
World History—Overview	84768	0.06	-0.03	-0.02	0.00	7.41	0.19	3.21	0.16
Language Arts (grade 8)	616852	0.05	-0.05	-0.05	0.00	8.11	0.21	3.08	0.16
Social Studies (grade 8)	476767	0.05	-0.06	-0.05	0.00	8.18	0.21	3.06	0.16
Music (grade 8)	257261	0.06	-0.07	-0.05	0.00	7.92	0.18	3.12	0.15

Science (grade 8)	591897	0.06	-0.08	-0.07	0.00	8.18	0.22	3.07	0.17
Technological Literacy	68029	0.07	-0.07	-0.08	0.00	7.87	0.21	3.12	0.16
Computer Literacy	65203	0.08	-0.14	-0.13	0.00	8.06	0.20	3.00	0.16
Exploratory	35381	0.06	-0.11	-0.17	0.00	8.66	0.28	3.03	0.15
Study Skills	78544	0.05	-0.26	-0.25	0.00	8.55	0.24	2.98	0.33
Mathematics (grade 8)	364597	0.08	-0.27	-0.29	0.00	8.70	0.29	2.94	0.20
Pre-Algebra	75596	0.04	-0.25	-0.32	0.00	8.71	0.18	2.94	0.20
Tutorial	68992	0.10	-0.48	-0.48	0.00	8.92	0.41	2.84	0.29
Mathematics—Supplemental	37397	0.09	-0.43	-0.48	0.01	9.56	0.39	2.78	0.21
Reading (grade 8)	46992	0.07	-0.49	-0.49	0.01	8.86	0.29	2.81	0.32
English as a Second Language	32411	0.93	-1.43	-1.21	0.01	8.64	0.33	2.46	0.13

Notes: Summary statistics for students enrolled in courses in grade 8 (2012-2019) with enrollments of at least 5% of the total enrollment. Courses in bold are included in the course type indicators used as covariates in the regression analyses. Courses indicated in italics are used to construct academic tracks.

Table A.4. Summary Statistics by Course Enrollment (Grade 10)

Course	N	LEP	Prior ELA Score	Prior Math Score	Prior Retention	Prior Absences	Prior Days Suspended	Prior GPA	Special Education
French III	40224	0.00	0.66	0.64	0.00	5.20	0.02	3.32	0.02
Algebra II	170280	0.01	0.49	0.58	0.01	5.96	0.10	3.27	0.04
Spanish III	140810	0.00	0.46	0.47	0.00	5.52	0.05	3.22	0.03
Chemistry	262342	0.02	0.22	0.25	0.01	6.76	0.13	3.01	0.08
Modern World History	54697	0.03	0.21	0.24	0.01	7.25	0.21	2.91	0.16
Health and Fitness	73106	0.03	0.09	0.10	0.01	7.51	0.19	2.87	0.16
Integrated Math—multi-year equivalent	42822	0.07	0.03	0.05	0.02	8.30	0.21	2.80	0.20
Physical Education/Health/Drivers' Education	75195	0.03	0.01	0.01	0.01	6.93	0.18	2.89	0.16
Health Education	139773	0.04	-0.02	0.00	0.01	8.10	0.28	2.82	0.14
English/Language Arts II (10th grade)	567927	0.03	-0.05	-0.04	0.01	7.96	0.26	2.77	0.15
Spanish II	129778	0.02	-0.09	-0.07	0.01	7.74	0.21	2.74	0.09
Visual Arts—Comprehensive	43093	0.06	-0.04	-0.09	0.02	8.93	0.34	2.74	0.15
Early U.S. History	216734	0.04	-0.09	-0.09	0.01	8.12	0.29	2.71	0.15
Physical Education	314700	0.06	-0.13	-0.10	0.02	8.52	0.34	2.69	0.15
Modern U.S. History	116438	0.04	-0.15	-0.14	0.02	8.15	0.31	2.68	0.15
U.S. History—Comprehensive	141176	0.08	-0.19	-0.17	0.02	9.09	0.30	2.67	0.15

Biology	259602	0.07	-0.25	-0.25	0.02	9.15	0.37	2.57	0.18
Geometry	393656	0.06	-0.26	-0.27	0.01	8.70	0.32	2.60	0.15
Spanish I	50259	0.04	-0.50	-0.51	0.02	11.17	0.57	2.24	0.25
Tutorial	36870	0.07	-0.66	-0.71	0.03	12.02	0.68	2.23	0.61
Study Skills	59947	0.05	-0.69	-0.72	0.03	12.24	0.54	2.24	0.64
Algebra I	32889	0.24	-0.70	-0.81	0.06	15.25	0.93	1.90	0.25
English as a Second Language	49243	0.96	-1.64	-1.35	0.07	9.73	0.35	2.23	0.07

Notes: Summary statistics for students enrolled in courses in grade 10 (2012-2019) with enrollments of at least 5% of the total enrollment. Courses in bold are included in the course type indicators used as covariates in the regression analyses. Courses indicated in italics are used to construct academic tracks.

Appendix B. Value-added and Teacher Performance Ratings

Table B1. Teacher Value-added and Teacher Performance Ratings

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Performance Rating						
Test VA	0.339***	0.336***			0.337***	0.334***
	(0.024)	(0.025)			(0.024)	(0.025)
Nontest VA			0.115***	0.146***	0.085**	0.119***
			(0.044)	(0.046)	(0.043)	(0.045)
Panel B. Performance Rating: Curriculum Planning and Assessment						
Test VA	0.162***	0.160***			0.162***	0.160***
	(0.013)	(0.014)			(0.013)	(0.014)
Nontest VA			0.053**	0.065***	0.039*	0.052**
			(0.022)	(0.023)	(0.022)	(0.023)
Panel C. Performance Rating: Teaching All Students						
Test VA	0.157***	0.156***			0.156***	0.155***
	(0.013)	(0.014)			(0.013)	(0.014)
Nontest VA			0.053**	0.061**	0.040*	0.048**
			(0.023)	(0.024)	(0.023)	(0.024)
Panel D. Performance Rating: Family and Community Engagement						
Test VA	0.062***	0.060***			0.061***	0.059***
	(0.011)	(0.011)			(0.011)	(0.011)
Nontest VA			0.027	0.031*	0.022	0.026
			(0.017)	(0.018)	(0.017)	(0.018)
Panel E. Performance Rating: Professional Culture						
Test VA	0.122***	0.122***			0.120***	0.121***
	(0.013)	(0.013)			(0.013)	(0.013)
Nontest VA			0.041*	0.051**	0.031	0.042*
			(0.023)	(0.024)	(0.023)	(0.024)
Panel F. Communications and Literacy Skills						
Test VA	0.095***	0.096***			0.096***	0.098***
	(0.034)	(0.037)			(0.035)	(0.038)
Nontest VA			-0.043	-0.038	-0.053	-0.046
			(0.054)	(0.059)	(0.056)	(0.060)
Panel G. Subject Matter Knowledge						
Test VA	0.210***	0.213***			0.213***	0.216***
	(0.039)	(0.043)			(0.040)	(0.044)
Nontest VA			-0.046	-0.041	-0.069	-0.061
			(0.061)	(0.065)	(0.063)	(0.067)
School-Grade-Year FE	Y		Y		Y	
School-Track-Year FE		Y		Y		Y

Notes: Coefficients on test scores and nontest factor from regressions with teacher performance ratings as the dependent variable. The sample includes all students in the matched long-run sample described in the text. All regressions include grade-by-year fixed effects.

Appendix C. Robustness Checks and Alternate Specifications

In this section, we investigate the sensitivity of the results in Tables 6 and 7 to alternative estimation strategies. Results are shown in Table C1 below, with Columns 1–4 estimating the relationship between test-based value-added and future student outcomes. To estimate Column 1, we first regress outcomes on the controls used in Eq. (1), along with grade-subject-year fixed effects, and a teacher fixed effect, to obtain residualized outcomes as in Chetty et al. (2014). We then regress these outcomes on test value-added. Column 2 follows this same procedure but with additional controls for district-level socioeconomic controls in the residualization procedure.[26] In general, Columns 1 and 2 are consistent with the results in Tables 5 and 6, but are larger in magnitude. For example, the coefficient on AP tests passed rises from about 0.04 to 0.14 when using this residualization procedure.

When assessing the plausibility of the research design by looking for evidence of sorting, we found that conditional on twice-lagged test scores, test value-added for a student's teacher in a given subject was positively correlated with test value-added for that student's teacher in the other subject. In Column 3, we add controls for other-teacher test and nontest value-added to our base specifications. Results are quite similar to the main specification. For example, the coefficient on college quality in Table 6 ($181.7) is very similar to Column 3 in Table C1 ($184.8).

In Column 4, we employ a teacher switching design as in Chetty et al. (2014). In particular, we collapse the outcome and test value-added to subject-grade-year-school cells and regress aggregate outcomes on aggregate value-added. Again, results are generally consistent with Tables 5 and 6, though the aggregation process results in substantially less imprecise estimates. Columns 5–8 repeat the sequence of robustness checks conducted in Columns 1–4 but with nontest value-added in place of test value-added. Results are again similar to the main results in Tables 5 and 6.

Table C1. Robustness to Alternative Specifications

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)

[26] These controls are obtained from the Stanford Education Data Archive and include percentage of students in urban, rural, and town locale schools, the percentage of students in families with a BA or higher, unemployment rate, SNAP receipt rate, poverty rate, single mother household rate, and an SES composite.

Panel A. AP Tests Passed								
Test VA	0.192***	0.136***	0.038***	0.056**				
	(0.031)	(0.031)	(0.008)	(0.028)				
Nontest VA					0.021	-0.009	-0.003	-0.061
					(0.064)	(0.059)	(0.011)	(0.049)
Panel B. SAT Scores (standard deviations)								
Test VA	0.166***	0.092***	0.015***	0.011				
	(0.011)	(0.013)	(0.003)	(0.016)				
Nontest VA					0.077**	0.005	-0.000	-0.022
					(0.031)	(0.029)	(0.005)	(0.030)
Panel C. Graduate High School								
Test VA	0.301*	0.307*	-0.028	-0.503				
	(0.161)	(0.157)	(0.093)	(0.318)				
Nontest VA					1.211***	0.662*	0.404**	-0.074
					(0.359)	(0.373)	(0.201)	(0.479)
Panel D. College Quality Index								
Test VA	2662.764***	1106.650***	184.798***	188.311				
	(191.189)	(177.198)	(60.891)	(239.059)				
Nontest VA					1639.451***	464.056	192.069*	345.887
					(437.802)	(324.551)	(98.304)	(434.992)
Panel E. Enroll in College								
Test VA	3.597***	1.379***	0.109	-0.075				
	(0.406)	(0.367)	(0.146)	(0.492)				
Nontest VA					2.118***	0.770	0.489*	0.403
					(0.748)	(0.629)	(0.279)	(0.990)
Panel F. Enroll in Four-Year College								
Test VA	6.153***	2.523***	-0.015	0.476				
	(0.489)	(0.462)	(0.148)	(0.581)				
Nontest VA					3.966***	1.181	0.672**	0.136
					(1.096)	(0.826)	(0.297)	(1.169)
Panel G. Enroll in Selective College								
Test VA	5.616***	2.412***	0.637***	0.762*				
	(0.484)	(0.358)	(0.129)	(0.455)				
Nontest VA					3.218***	0.888	0.204	0.360
					(0.944)	(0.630)	(0.174)	(0.751)
N	1321755	1321755	1218104	7776	1319582	1319582	1216797	7776
Standard Controls	Y	Y	Y		Y	Y	Y	
Additional Controls		Y				Y		
Other-teacher VA Controls			Y				Y	
Switching Design				Y				Y

Notes: Columns 1–2 and 5-6 use residualized outcomes. Columns 2 and 6 include additional controls for district-level SES; see above. Columns 3 and 7 include controls for the test and nontest value-added of student's teacher in other subject (e.g., when estimating the impact of VA of a math teacher, controls for test and nontest VA of the student's ELA teacher). Columns 4 and 8 use grade-school-year aggregates of value-added and outcomes as in Chetty et al. (2014).

Appendix D. Comparison with Prior Research

Several prior studies have examined the long-run effects of assignment to teachers with differing value-added. In this appendix, we investigate the sensitivity of the results to the methodological choices in the literature.

Chetty, Freidman, and Rockoff ([CFR], 2014b) estimate the effects of teachers on long-run outcomes in New York City. We follow their approach for constructing value-added predictions using residuals from the regression

$$Y_{ij} = X_i\beta + \theta_j + \epsilon_{ij} \tag{A.1}$$

To estimate the effects of teachers on long-run outcomes, they work with outcome residuals constructed in a similar manner. They first regress outcomes on student and classroom characteristics and teacher fixed effects and then deduct the fitted values (using student and classroom characteristics only). They then regress these residuals on the predicted teacher value-added:

$$Y_i^* = \alpha + \hat{\theta}_{jt}\delta + \eta_i \tag{A.2}$$

In practice, CFR estimate Eq. (A.2) on data aggregated to the classroom level. We use the student-level data so that results are more directly comparable to the baseline models. The research design described by Eqs. (A.1) and (A.2) is a pure selection on observables design that relies on the control vector to adjust for the influences of schools, families, and other student unobservables. CFR (2014a) test the plausibility of the research design by including omitted covariates in their construction of Y_i^* and assessing the stability of the coefficient on teacher VA, δ. Their data includes prior lags of student test scores and parental income data from tax returns. To test the plausibility of the design in our setting, we pursue a similar approach using data on district demographics from the Stanford Education Data Archive (Reardon et al., 2021).[27]

In addition to the selection on observables design, CFR propose an estimator that leverages changes in teaching assignments. In particular, they construct versions of their EB teacher value-added predictions that omit student data from year *t* and either year *t+1* or year *t-1*. Let .[28] Methods similar to those proposed by CFR are also used in several other studies (Gilraine & Pope, 2021; Petek & Pope, 2021).

Our primary research design more closely follows that proposed by Jackson (2018), which assumes that teacher assignments are exogenous conditional on school and track assignments in addition to student covariates. However, the methods used in Jackson (2018) differ somewhat from our baseline specification. Jackson (2018) estimates empirical Bayes predictions of teacher value-added that directly incorporate school and track effects into the first stage estimation. In particular, he estimates a value-added model

$$Y_{ij} = X_i\beta + \theta_j + \epsilon_{ij} \tag{A.3}$$

The VAM in Eq. (A.3) omits teacher effects and includes school-track and school-year fixed effects. An advantage of this approach is that it is an unbiased estimate of teacher quality within tracks under In addition, Jackson (2018) leverages variation in teacher effectiveness within tracks across cohorts. That is, his second stage regression is

[27] The variables we include in this analysis are district urbanicity, log median income, proportion of adults with a bachelors degree or higher, the unemployment rate, participation in the Supplemental Nutrition Assistance Program, the poverty rate, the proportion of single mothers, and a general socioeconomic status measure.

[28] Our teacher switching design differs somewhat from that proposed by CFR. We use hold-out data from the sample of students lacking long-run outcomes and estimate empirical Bayes predictions for teachers in the long-run sample. We then aggregate these predictions to the school-grade-subject-year level and estimate models with school-grade-subject and grade-subject-year fixed effects using the aggregated TVA as instruments for the assigned teacher. This approach is similar to one taken by Jackson (2018).

$$Y_{ij} = X_i\beta + \theta_j + \epsilon_{ij} \tag{A.4}$$

The research design in Eq. (A.4) incorporates both variation in teacher value-added resulting from the turnover of teachers as well as variation resulting from the leave-out nature of the EB estimate. A similar approach is taken by Liu and Loeb (2021). Like our application, they estimate a first stage VAM without school effects; as in Jackson (2018), they then estimate a second stage regression with school fixed effects.