How to Measure a Teacher: The Influence of Test and Nontest Value-Added on Long-Run Student Outcomes

Ben Backes
James Cowan
Dan Goldhaber
Roddy Theobald

April 2023

WORKING PAPER No. 270-0423-2

CALDER
National Center for Analysis of
Longitudinal Data in Education Research

AIR®

How to Measure a Teacher: The Influence of Test and Nontest Value-Added on Long-Run Student Outcomes

Ben Backes
American Institutes for Research / CALDER

James Cowan
American Institutes for Research / CALDER

Dan Goldhaber
American Institutes for Research / CALDER
University of Washington

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

How to Measure a Teacher: The Influence of Test and Nontest Value-Added on Long-Run Student Outcomes

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

CALDER Working Paper No. 270-0423-2
April 2023

Abstract

This paper examines how different measures of teacher quality are related to students' long-run educational trajectories. We estimate teachers' *test-based* and *nontest* value-added (the latter based on contributions to student absences, suspensions, grade progression, and grades) and assess how these predict various student postsecondary outcomes. We find that both types of value-added have positive effects on student outcomes. 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 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 extent to which a teacher contributes to students' test score growth beyond what would be expected given their starting points—have long been a primary way of understanding the effects of teachers on student outcomes. For example, shortly before the COVID-19 pandemic, thirty-four states required 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, 2019).[1] Policymaker interest in using test-based value added is bolstered by evidence that these measures are causally linked to students' later life outcomes (Bacher-Hicks and Koedel, 2022). Chetty et al. (2014b), for instance, find that being assigned to teachers with higher test value-added improves a variety of students later life outcomes, including college quality and adult earnings.

But test-based teacher quality measures also have clear limitations. Because schools do not test in all grades and subjects, value-added measures are only available for a small fraction (typically 20%) of the teacher workforce. Value-added measures are also controversial and unpopular among teachers, and their use in high-stakes settings has also prompted lawsuits in several settings (American Federation of Teachers, 2017; Paige, 2020). And, as a newer body of research shows, test-based measures of teachers fail to capture important ways in which teachers contribute to student success in schools.

[1] Researchers have used value-added measures to examine how equitably teachers are distributed across students (Goldhaber et al., 2017; Isenberg et al., 2022; Williams et al., 2016); changes in the quality of the teacher workforce over time (Nagler et al., 2019); and how well teacher preparation, licensure and evaluation systems identify effective teachers (Boyd et al., 2009; Clotfelter et al., 2007, 2010; Cowan et al., 2020; Goldhaber, 2007; Jacob et al., 2018; Kane et al., 2011).

In light of these limitations, recent research has developed new methods for measuring teacher contributions to *nontest* outcomes, such as attendance and grades (described in more detail in Sections 2 and 3.1). This research finds that nontest teacher quality measures are only weakly correlated with test-based quality measures, and that nontest teacher quality measures predict critical outcomes on the road to college enrollment, such as SAT taking and high school graduation (e.g., Backes & Hansen, 2018; Gershenson, 2016; Jackson, 2018; Kraft, 2019; Liu & Loeb, 2021). These findings pose what appears to be, at a high-level, an empirical conundrum. One the one hand, recent papers comparing test and nontest measures of teacher quality have found near zero effects of test value-added on outcomes that we would expect to be related to future college outcomes, such as SAT test taking (Petek & Pope, 2021; Gilraine & Pope, 2021) and high school graduation (Jackson, 2018; Liu & Loeb, 2021; Gilraine & Pope, 2021). On the other, Chetty and colleagues (2014b) find long-run effects of test value-added on outcomes such as college quality and adult earnings. Taken together, research on test and nontest measures of teacher quality raises a host of puzzling questions about how we should understand and measure the different ways in which teachers impact student success in K-12 and beyond.

In this paper, we address some of these puzzles by examining both how test and nontest measures of teacher quality (also referred to as teacher value-added) predict students' secondary and postsecondary outcomes. Specifically, using a sample of students from Massachusetts, we replicate prior findings that nontest teacher quality measures 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 test and nontest measures both play important roles in predicting long-run postsecondary

outcomes, but they appear to operate on different margins depending on the student outcomes in question. Outcomes like enrollment in college and enrollment in 4-year college are better predicted by nontest teacher quality measures. But other outcomes, like enrolling in a selective college, are better predicted by test-based quality measures. These results are robust to the various ways of accounting for student-teacher sorting, such as the addition of school-track fixed effects.

To help explain the divergent results between teacher test and nontest quality measures, we then 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 long-run outcomes distribution. We find that college quality is sensitive to test-based measures of teacher quality throughout the outcome distribution, but nontest measures of teacher quality have significantly larger effects for students at the bottom and middle of the college quality 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 teacher quality measures to student college enrollment outcomes.[2] We find that nontest teacher quality affects college enrollment, the likelihood of enrolling in a 4-

[2] As we describe in Section 2, 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).

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 teacher 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. We replicate prior findings and further demonstrate that this conundrum can be explained by: (a) differential impacts of the teacher quality measures on different short-term student outcomes; and b) differential relationships between short-run and long-run outcomes for students in different parts of the distribution.

2. Background and Prior Literature

The statistical properties of traditional test-based 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 student and classroom covariates, 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.[3] Using these measures, researchers have found substantial variation in teaching effectiveness (Aaronson et al., 2007; Chetty et al., 2014a, 2014b; Rivkin et al., 2005). And, as noted above, researchers have found that test value-added measures impact later outcomes such

[3] For example, using tax data, Chetty et al. (2014a) estimate an upper limit of the degree of bias induced by not including factors such as parental income and 401(k) contributions of 0.25 percent. The authors attribute this minimal degree of bias to the fact that the typical set of controls in value-added models, especially prior test scores, capture much of the unobserved variation in parental advantage. Wee Bacher-Hicks and Koedel (2022) for a more thorough discussion.

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).

Chetty and colleagues (2014b) provide 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 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) demonstrates that teachers also have long-run effects operating through effects that are not fully captured by test-based measures of teacher quality. 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 similar variation in teacher effects across test and nontest measures, but the correlation between the two types of teacher effects is quite small.

The estimated effects of the two teacher quality measures differ significantly. Jackson (2018) finds that assignment to a teacher one standard deviation higher on the nontest value-added measure is estimated to improve on-time high school graduation by about 1.5 percentage points and reduces the dropout rate by about 0.4 percentage points, whereas a one standard deviation increase in test-based value-added is only estimated to increase on-time high school completion by 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-based value-added is more strongly linked to SAT scores than nontest teacher quality measures are.

Several studies have also used Jackson's nontest index of teacher quality to assess teachers' contributions to future student outcomes. Petek and Pope (2021) and Gilraine and Pope (2021) construct a nontest factor and find that nontest teacher quality measures for elementary teachers is more predictive of not repeating a grade (Petek & Pope, 2021) and 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 the prior studies mentioned, Liu and Loeb (2021) find that attendance value-added is a better predictor of high school completion. In addition, they find that 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.[4] 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

[4] The exception is for middle school teachers: a one standard deviation increase in test-based value-added of teachers in middle school is estimated to increase the likelihood of earning a Regents diploma by about 0.2 percentage points.

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.[5]

Although the results differ by study and context, there appear to be some common trends in this literature, which we summarize in Table 1. 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 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).[6] Third, the literature suggests that test-based measures 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 whereas nontest measures of teacher quality appear to have larger effects on outcomes that are more proximate to the high school completion and college enrollment/non-enrollment margins. As a crude measure of this, taking a simple average of the estimated impacts from the four studies which examine high school graduation reveals that a one standard deviation increase in test-based measures of teacher quality raises high school graduation by 0.11 percentage points, compared to 0.81 percentage points for nontest measures

[5] 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.

[6] The idea that these different measures 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).

of teacher quality (Jackson, 2018; Liu & Loeb, 2021; Gilraine & Pope, 2021; Rose et al., 2022). As we describe in the sections that follow, we explore these issues further by focusing on different margins of students' postsecondary educational enrollment.

The outcomes used in existing research have two features that may tend to overstate the importance of nontest teacher quality measures for student outcomes in future years. First, as we show in Section 6.2 below, much of the prior work has focused on outcomes, such as high school completion or college plans, that exhibit relatively large returns to soft skills near the bottom of the student skills distribution. Therefore, conclusions about the long-run effects of teaching skills may depend on how teacher effects vary across the short-run outcomes distribution. In particular, if certain teaching skills are disproportionately important for lower achieving students, we should expect the extant literature to find these skills to be the most important for certain long-run outcomes.

This is potentially a concern because nontest outcomes such as attendance and discipline exhibit substantial skew. Jackson et al. (2022) refer to this possibility as *mechanical heterogeneity* because the skill measures available in administrative datasets are more sensitive to educational interventions for students near the thresholds of the margins for suspensions, grade promotion, or attendance. Therefore, the conclusions of prior studies – that teacher skills operating through nontest outcomes have larger long-run effects than those operating through test scores – may be sensitive to the outcomes available in prior research. One of the main contributions of this paper is to use a non-binary postsecondary outcome that has a stronger relationship with student short-run measures throughout the student test and nontest distributions. Using this outcome, we find closer returns to teacher test and nontest skills than those indicated in prior research. Taken together, the results in our study suggest that the choice of outcome

measures matters but that the greater long-run returns to teacher nontest skills found in the recent literature on nontest measures of teacher quality cannot be explained solely by the choice of long run outcome measures.

3. Data and Measures

We use a sample of students from Massachusetts matched to teachers, end-of-year standardized tests and various nontest short- and long-run outcomes (described in more detail below). 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.[7]

3.1 Data and Measures

Short-run outcomes: There are two types of short-run measures used on the left-hand side of Eq. (1) below when constructing the two measures of teacher quality. First, for test-based measures of teacher quality, we use standardized test data for math and ELA in Grades 7, 8, and

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

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.[8]

Second, following Jackson (2018), to construct nontest measures of teacher quality, 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.[9] 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.[10] 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.

[8] 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).

[9] We use all students enrolled in Grades 7–12 to estimate the factor model using the Bartlett scoring method. The factor weights are days absent (-0.57), days suspended (-0.36), GPA (0.76), and grade retention (-0.24). For grade 6 (used as a prior-year control for grade 7), we estimate a factor model that excludes GPA.

[10] 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.

Intermediate Secondary Outcomes:[11] 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. We use student enrollment records to measure credits earned through 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).

Prior studies have used various proxy measures for college plans (e.g., self-reported plans to attend a 4-year college after graduation, 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.[12]

Long-Run 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.

[11] 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 limit the sample in these grades to students enrolling in public high schools.

[12] Prior to taking the SAT, students fill out a Student Data Questionnaire that asks about students' college degree goals (among other topics). The vast majority of students who take the SAT intend to obtain an associate 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

We then match enrollment data to the College Mobility Report Card constructed by Chetty et al. (2017). Following Chetty et al. (2017), we use an index of college quality based on the median earnings of students at ages 33–35 who attended the college (or did not attend college at all) from the 1980–1982 birth cohorts.[13] This index is available for students who do not enroll in college and thus measured for the entire sample.[14] We supplement this college quality 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.[15] 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

[13] 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.

[14] 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.

[15] 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.

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 Student Short-Run Outcomes Versus Student Long-Run Outcomes

We explore the relationship between short-run student test scores and nontest factors and long-run outcomes in Table 3, 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. The relative magnitude of test scores and the behavioral factor varies quite a lot according to the outcome in question. Focusing on Panel B, where we focus on the likelihood of graduating high school, the coefficient on the nontest factor is two times as large as the coefficient on test scores (where we control for both test scores and the nontest factor in Column 5). And comparing the separate regressions with test scores only (Column 6) to the nontest factor only (Column 7), 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 5. In other words, these results show that the nontest factor has far greater predictive power than test scores when it comes to predicting the high school graduation margin. A similar pattern emerges for the college-going margin (Panel E). Test scores, however, are more predictive than the nontest outcomes for other outcomes, like passing AP tests (Panel C). And when looking at overall college quality (Panel H), 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 relatively more signal for the *type* of college attended than the college-going margin.

3.3 Teacher Value-Added Measures

We follow the approach to estimating out-of-sample value-added taken by Chetty et al. (2014a) and Chetty et al. (2014b). Briefly, we first construct student residuals from the following

equation that includes student, classroom, and school-grade covariates as well as school fixed effects using a sample of students in Grades 7, 8, and 10:

$$Y_{it} = X_{it}\beta + \mu_s + \epsilon_{it}, \quad (1)$$

where Y_{it} represents either test scores or the nontest factor.[16] The control vector X_{it} includes student race/ethnicity, gender, free and reduced-price lunch status, participation in special education or English learner programs, and cubic polynomials of prior test scores (math and ELA separately) and prior-year nontest factor, along with the school and classroom means of each of these covariates.[17] We estimate Eq. (1) separately by subject (math or ELA), grade level (i.e., 7 and 8 versus 10), and outcome (test or nontest).

We then form leave-out empirical Bayes predictions of teacher quality to use as regressors following prior work (Chetty et al., 2014a).[18] In particular, we first construct student residuals based on Eq. (1) and then obtain a teacher-year-subject-outcome effect by averaging the student residuals:

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

We then construct 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*)

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

[16] As shown in Table C4, results are similar when residualizing on teacher fixed effects (as in Chetty et al., 2014) rather than school fixed effects in the first stage. The main difference is the estimates for nontest value-added, which are attenuated when using teacher fixed effects rather than school fixed effects (Table C4, Panels C-F). This is driven in part by the estimates of the standard deviation of nontest teacher effects used to standardize value added, which are, for example, 0.21 for middle school ELA when residualizing on teacher and 0.14 when residualizing on school. The former likely overstates the variance across teachers by conflating teacher and school effects; we use school fixed effects in Equation (1) as this specification performs substantially better in the quasi-experimental tests (Figure 1 and Appendix Table C1).

[17] Students are not tested in math or ELA in ninth grade in Massachusetts. For students in 10th grade, the lagged achievement data are from eighth grade.

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

where $\hat{\mu}_{j-t}$ is the vector of teacher-year means in years other than *t* and $\Omega_{j,-t}$ is a vector of weights (Chetty et al., 2014a; Stepner, 2013). The resulting prediction $\hat{\theta}_{j,-t}^{U}$ 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 denote this measure with a *U* superscript to highlight that these are not the standardized estimates that will be used for later results; i.e., these measures have not yet been scaled to represent teacher standard deviations.

Several random assignment experiments and quasi-experimental validations have found that value-added models similar to Eq. (1) provide nearly unbiased forecasts of teacher effectiveness in subsequent school years (Bacher-Hicks et al., 2019; Chetty et al., 2014a; Kane & Staiger, 2008). In Figure 1 we re-produce Figures 2 and 4 in Chetty et al. (2014a), using both test and non-test value added. Results are similar to Chetty et al. (2014a) for test scores and we cannot reject unbiasedness. For the nontest factor, while we again cannot reject unbiasedness, results are less precisely estimated.[19] Overall, we conclude that, consistent with prior work, estimates of $\hat{\theta}_{j,-t}^{U}$ appear to capture inherent teacher skills that improve the short-term test and nontest outcomes of their students in a manner that is not driven by the sorting of students to teachers.

Finally, we re-scale the estimates of teaching effectiveness to be expressed in terms of teacher standard deviations by dividing $\hat{\theta}_{j,-t}^{U}$ by an estimate of the standard deviation of teacher effectiveness. As in Chetty et al. (2014b), we use the square root of the covariance of mean score

[19] In Appendix Table C1, we re-produce Table 4 in Chetty et al. (2014a), which regresses changes in student outcomes in school-grade-year cells on changes in value added forecasts in those cells. Again, results are somewhat imprecise because we only have one grade from high schools (grade 10), making the number of school-year cells equal to the number of school-grade-year cells in high school. We thus have about one-quarter of the number of cells as in Chetty et al. (2014a). With the caveat about precision, we are unable to reject unbiasedness.

residuals across classrooms within the same year as our estimate of the true standard deviation of teacher effects.[20] We refer to this standardized version as $\hat{\theta}_{j,-t}$ for the remainder of the paper.

Statistical Properties of Test-Based and Nontest Value-Added

In Table 4, we consider the relationship between the two out-of-sample measures of teacher quality and contemporaneous student outcomes. We present results for two sets of fixed effects: one at the school-grade-year level and one at the school-grade-track-year level. In Table 4, results are similar across both specifications. A one standard deviation increase in out-of-sample test-based value-added is associated with an increase in test scores of 0.11-0.12 standard deviations; this is consistent in magnitude with prior work (e.g., Chetty et al., 2014a; Hanushek and Rivken, 2010). In addition, test value-added is associated with very little change in nontest outcomes (Panel B, Columns 1 and 2).

Likewise, out-of-sample nontest value-added primarily affects students' nontest outcomes (Panels A and B, Columns 3 and 4) and not test scores.[21] A one standard deviation in nontest value-added is associated with a 0.12-0.13 standard deviation increase in the student nontest factor. While this estimate is larger than Jackson's (2018) estimate of 0.06 obtained from ninth graders, Jackson also finds a smaller relationship between test value-added and student

[20] For test value added, we estimate the standard deviation of teacher effectiveness to be 0.122 for middle school math, 0.089 for high school math, 0.130 for middle school ELA, and 0.092 for high school ELA. For nontest value added, we estimate 0.122 for middle school math, 0.089 for high school math, 0.140 for middle school ELA, and 0.092 for high school ELA

[21] 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.

achievement (0.07) than is commonly found in the literature, perhaps in part because the variance of teacher effects tends to be smaller in high school.[22]

The correlation between test and nontest value-added for a given teacher in a given year is 0.10, which is in the range of prior estimates (Table 1). These findings add to a growing body of evidence that test-based and nontest value-added are positively correlated but capture distinct facets of teacher skill.

4. Empirical Methods

4.1 Statistical Model

Our objective is to understand how assignment to specific teachers with the out-of-sample skill measures constructed in Section 3.3 above ($\hat{\theta}_{j,-t}^{test}, \hat{\theta}_{j,-t}^{nontest}$) affects later 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

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

where X_{it} contains the same regressors as in Equation (1), with the exception of school-level averages which would be absorbed by the school-year (or school-track-year) fixed effects. In addition, $\hat{\theta}_{j,-t}$ is a vector containing predicted teacher effects on test and nontest outcomes estimated out-of-sample (i.e., in years other than *t*), 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 estimated teacher

[22] When examining the components of the nontest factor individually, increases in nontest factor predict value added increase each of the four sub-components, although the estimated effect on student absences is cut in half when adding track fixed effects.

quality, $\hat{\theta}_{j,-t}$, conditional on school effects and the control vector.

We worry about two primary sources of bias in Eq. (2). The first is that the teacher effects themselves $\hat{\theta}_{j,-t}$ could be estimated with bias. As discussed above, prior work has found that the inclusion of prior-year controls is sufficient to remove potential biases associated with student-teacher sorting (Chetty et al., 2014a), and our replication of these tests finds similar results (Figure 1; Appendix Table C1). The second potential source of bias is one that can arise even $\hat{\theta}_{j,-t}$ itself is unbiased: correlation between the error term and both $\hat{\theta}_{j,-t}$ and Y_{ijst} conditional on the regressors X_{it}. This could arise if, for example, high or low quality teachers were systematically sorted to schools that had differential school-level impacts on student long-run outcomes.

We thus include school-subject-grade-year fixed effects to mitigate concerns about the independent effects that schools and teachers can have on long-run 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 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. The direction of the bias is unclear a priori, particularly if teachers and schools differ in the correlation in effects across outcomes.

In addition, at the high school level, Jackson (2014) has found that models such as Eq. (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.[23] In addition, because middle schools may offer multiple sections of courses aligned to the state core curriculum that are nonetheless tracked by student achievement, we 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.[24] 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.

Finally, when exploring heterogeneous impacts of value added, we estimate Eq. (2) across 10 different deciles of incoming educational advantage, defined as the average of prior test and nontest outcomes as in Jackson et al. (2022). We then report decile-specific coefficients δ for test and nontest value-added where both teacher quality measures are included in the same regression.

[23] 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.

[24] 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.

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

5.1 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 high school level; this allows us 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, which is in the range of prior studies.[25] We also find that students in classrooms with higher test-based value-added tend to take more AP credits (0.10 credits per teacher standard deviation) and to pass more AP tests (0.03 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 – although our point estimates are not statistically significant – 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 that are more relevant for students in the middle or bottom of the test achievement distribution

[25] Rose and colleagues (2022) find an impact of 0.20 percentage points, Liu and Loeb (2021) find an impact of 0.70 percentage points, Gilraine and Pope (2021) report 0.83 percentage points, and Jackson (2018) reports 1.5 percentage points.

(dropping out, high school graduation, and SAT test-taking), 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 6 below.

5.2 Long-run Postsecondary Outcomes

Table 6 displays results for long-run postsecondary outcomes. Three notable findings stand out. First, some outcomes have large differences between the regressions with school fixed effects and the regressions with school-track fixed effects.[26] For example, the point estimate for the relationship between a one standard deviation increase in test value-added and enrolling in a selective college is 0.59 percentage points in the school fixed-effects model (Column 1), but only 0.40 percentage points in the track fixed effects-model (Column 2). This suggests that, as demonstrated in Jackson (2014), models with school fixed effects but not track fixed effects may conflate differences in teacher quality with the sorting of students and teachers into different tracks, at least for certain long-run outcomes. Second, test value-added (Column 2) and nontest value-added (Column 4) each independently predict later college quality (Panel E). And finally, when combined into the same regression, test and nontest value-added continue to each have predictive power for college quality (Column 6), with nontest value added estimated to have the larger effect.

While test and nontest value-added each predict college quality in Panel E, the remaining panels suggest that they do so through somewhat different mechanisms. For example, nontest value-added is more strongly related to college enrollment and whether a student enrolls in a 4-

[26] This is in contrast to the findings for the secondary outcomes in Table 5, which are largely insensitive to the inclusion of school fixed effects. A possible explanation for this pattern is that sorting into tracks is more strongly related to college than high school outcomes.

year college (Panels A and C). In contrast, test value-added is more strongly related to whether a student enrolls in a selective college (Panel D). Importantly, unlike binary measures like high school completion or college enrollment, the college quality measure used in Panel E 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 somewhat smaller than those reported by Chetty et al. (2014b). We find that having a teacher with one standard deviation higher test-based value-added is associated with a $165 increase in college quality, compared to Chetty et al.'s (2014b) $299. 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 per classroom. 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. Finally, in contrast to Chetty et al. (2014b), we do not find a statistically significant relationship between test-based value-added and the college-going margin.

5.3 Mechanical Relationships Between Nontest Value-added and Long-run Outcomes

In this section, we investigate whether the results in Tables 5 and 6 could be driven by teachers with easier grading standards (Gershenson et al., 2022) rather than a true improvement in student skills due to being assigned to the teacher – i.e., a grading effect. Because students cannot graduate high school without passing their math and English courses, and because admission to four-year institutions – and especially selective four-year institutions – is predicted by high school grades, it is important to separate a grading effect from a true teacher effect. To

investigate this possibility, we re-estimate the outcomes in Tables 5 and 6 that could be sensitive to this concern about mechanical correlation – high school graduation along with the college outcomes – in a sample of students in grades 7 and 8 only. The intuition is that student grades in middle school should not have a mechanical correlation with graduation and postsecondary outcomes because they are not directly used in determining eligibility to graduate and in college applications.

Results are shown in Table C2. Patterns are broadly similar, which is notable because course grades in grades 7 and 8 aren't mechanically related to high school graduation or college admission. To further explore the contribution of course grades to the nontest measure, we estimate the key outcomes shown in Table C2 with an alternate version of nontest value-added that does not use course grades. In addition to exploring the potential for mechanical relationships between value added models and student later outcomes, this is the version of nontest value-added that would likely be necessary for students in earlier grades where grade point average is not calculated. Results are shown in Table C3 and broadly similar. This is consistent with Jackson's (2018) finding that excluding GPA from the nontest measure still generates nontest value-added that is predictive of later high school graduation. Overall, these patterns are consistent with nontest value added representing a true contribution to student skills rather than a grading effect.

5.4 Robustness to Alternative Estimation Strategies

In this section, we investigate the sensitivity of the results in Tables 5 and 6 to alternative estimation strategies. Results are shown in Appendix Table C4 below, with Columns 1–4 estimating the relationship between test-based value-added and future student outcomes. Column 1 shows our base specification for comparison. To estimate Column 2, we add controls for other-

teacher test and nontest value-added to our base specifications. In Column 3, we use within-teacher instead of within-school variation obtaining test score residuals in Eq. (1). In Column 4, we employ a teacher switching quasi-experimental design as in Chetty et al. (2014b) by collapsing the outcome and value-added to subject-grade-year-school cells and regress aggregate outcomes on aggregate value-added.

Results are generally consistent with Tables 5 and 6, though there are some differences. For example, the quasi-experimental results in Column 4 tends to suggest a larger relationship between test value-added and college outcomes, including college enrollment and four-year college enrollment. On the other hand, the estimated relationship is smaller for nontest value added in Column 8, although estimates are very imprecise.

6. Heterogeneous Effects of Test and Nontest Teacher Quality

The results in Tables 5 and 6 show that test and nontest teacher quality each predict some long-run outcomes but not others. In this section, we further explore this pattern by first developing a basic model that illustrates the sensitivity of an outcome to a teacher quality dimension in Section 6.1. We then examine the two factors that the model suggests determine the sensitivity of a long-run outcome to changes in a teacher quality dimension: the sensitivity of the long-run outcome to short-run student outcomes (test and nontest) in Section 6.2, and the sensitivity of the short-run student outcomes (test and nontest) to the two dimensions of teacher quality in Section 6.3.

6.1 Theoretical Model of Teacher Impacts on Long-run Student Outcomes

Consider a model of student skill formation similar to those considered by Cunha and Heckman (2008) and Cunha et al. (2010). Denote the student outcomes used to estimate teacher value added as $\boldsymbol{x_t} = (x_t^{test}, x_t^{nontest})$. In each year, students are assigned to teachers with value

added $\theta_j = (\theta_j^{test}, \theta_j^{nontest})$. For $k = test, nontest$, we assume that the short-run outcome, x_t^k, is a function of baseline outcomes, x_{t-1}, and teacher value added of the corresponding type, θ_j^k:

$$x_t^k = g_k(\boldsymbol{x_{t-1}}, \theta_j^k).$$

We then assume the long-run outcome is a function only of student outcomes at time *t*:

$$y = f(\boldsymbol{x_t}) = f\left(g_C(\boldsymbol{x_{t-1}}, \theta_j^{test}), g_N(\boldsymbol{x_{t-1}}, \theta_j^{nontest})\right).$$

This formulation implicitly assumes that teacher effects on long-run outcomes operate by improving students' short-run skill measures. We can derive the partial effect of teacher skill θ_j^k on long run outcomes via the chain rule:

$$\frac{\partial y}{\partial \theta_j^k} = \frac{\partial y}{\partial x_t^k} \frac{\partial g_k}{\partial \theta_j^k}.$$

That is, the effect of teacher quality on the long-run outcome depends on the product of the derivative of the long-run outcome with respect to the short-run outcome x_t^k and the derivative of the short-run outcome index with respect to teacher quality θ_j^k.

This derivation suggests two important sources of heterogeneity in the partial effects of teacher skills on student long-run outcomes. One, the effects of improving student skills on the long run outcome may differ depending on the combination of short-run skills and long-run outcomes ($\frac{\partial y}{\partial x_t^k}$). For instance, encouraging students to attend class and obtain passing grades may be especially productive for improving high school completion outcomes, for which these intermediate outcomes are often prerequisites.

Second, the effects of teacher skills depends on the product of the teacher effect on the short run skill measure and the effect of increasing the skill measure on the long run outcome. If both of these effects vary over the student skill distribution, then the long run effects of teacher skill may be sensitive to these kinds of heterogeneity. Put differently, we would expect to

observe large long-run effects of a teaching skill if that skill has (a) large effects on the short-run measure for certain students and (b) the effect of improving the short-run measure on the long-run outcome is large for the same set of students.

6.2 Differential Test and Nontest Effects On Long-Run Outcomes

To examine how the relationship between student short-run and long-run outcomes varies over the distribution of student skills, we begin by constructing deciles of the prior test and nontest factor as in Jackson et al. (2022). For simplicity, we show all relationships across deciles of the *average* prior test and nontest indices; however, focusing only on the prior value for the relevant skill measure produces similar patterns. We then estimate 10 versions of Eq. (2), one for each decile, where we include student's test scores and nontest factor instead of the teacher quality estimates. By doing this, we obtain estimates for the impact of a one standard deviation change in the short-run value-added inputs (test and nontest) on a given long-run outcome at different points in the educational advantage distribution.

Results are shown in Figure 2. Beginning with Panel A, high school graduation is relatively insensitive to test score changes throughout the distribution, with the exception of students at the very bottom of the distribution. On the other hand, students at the low end of the incoming educational advantage distribution who experience gains in short-run nontest outcomes have large increases in the likelihood of high school graduation. In contrast, high school graduation is mostly insensitive to gains in short-run nontest outcomes (and test scores) at the top of the distribution because nearly all of these students already graduate.

In Panels B through D, we explore whether students attend college and which type they attend. In Panel B, overall college attendance is again relatively insensitive to test scores gains. However, the middle of the educational advantage distribution experiences large increases in

college-going when the nontest factor increases. As shown in Panel D, this is driven by increases in four-year college-going in the middle of the distribution. For overall college quality in Panel F, the relationship is quite flat for test scores. This reflects moderate increases in the likelihood of attending college at the bottom of the distribution, attending a four-year college in the middle of the distribution, and attending a selective college at the top of the distribution. On the other hand, increases in short-term nontest outcomes most positively increase college quality in the middle of the distribution due to the likelihood of attending college and attending a four-year college.

6.3 Differential Test-Based and Nontest Teacher Value-Added Effects

We next examine how the relationship between teacher value-added and short-run student outcomes (test and nontest) varies by incoming student skill measures. As in Jackson et al. (2022), we obtain the impact of a one standard deviation change in out-of-sample test value-added and in nontest value-added on a given student outcome for each decile on an index of student prior outcomes, controlling for student demographics, prior test scores and behaviors, and school-track fixed effects as in Eq. (2). Results are shown in Panels A and B of Figure 3.

As discussed above, the cross-skill relationship (i.e., the effect of nontest value-added on test scores in Panel A) is extremely weak, suggesting that test and nontest value-added capture distinct facets of teacher skills. In addition, we observe 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 relatively flat throughout the educational advantage distribution (Panel A), while for nontest value-added, its impact on the nontest factor is largest for students at the bottom (Panel B). The latter finding 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.

Returning to the model developed in Section 6.1 above, these patterns present an explanation for the results in Tables 5 and 6 and in prior literature. Specifically, for a long-run student outcome to be sensitive to a particular type of value-added, it needs to be the case that (a) the long-run outcome is sensitive to changes in the short-run measure and (b) the short-run measure is sensitive to changes in value-added throughout a meaningful part of the distribution. For test scores, (b) is straightforward because there is little heterogeneity by test value-added. This would suggest long-run impacts of test value-added on outcomes that are most sensitive to test score increases: selective college attendance and overall college quality. For nontest outcomes, the relationship between nontest outcomes in the short run and nontest value-added is an additional consideration. Because nontest value-added has larger effects on the nontest index for students with lower values of the baseline outcomes, we may tend to find larger effects on the long-run outcomes where improvements for these students are most productive.

We directly examine how the relationship between out-of-sample value-added and long-run outcomes varies by placement in the educational advantage distribution in Panels C through G of Figure 3. For graduating high school, attending college, attending two-year college, and attending four-year college, there is little relationship between test value-added and the long-run outcome (Figures 3C-3F). On the other hand, test value-added impacts attending a selective college at the top of the distribution (Figure 3F) and college quality throughout nearly the entire distribution (Figure 3G).

Unlike with test scores, the relationship between behavior value-added and behaviors is not flat throughout the distribution (Figure 3B). We may thus expect teacher effects on long-run

outcomes to be larger at the bottom of the distribution. While estimates for non-test value-added on long-run outcomes are noisy when broken down into deciles, this is what we observe for some long-run outcomes, with high school graduation, college attendance, and overall college quality having larger effects at lower levels of the baseline skills index (the latter in part because this measure captures the high school graduation and college enrollment margins). However, four-year college enrollment and selective college enrollment do not follow this pattern: the estimated slope at the bottom of the distribution is positive. Notably, although selective college attendance is similarly sensitive to increases in tests and behaviors (Figure 2), students whose nontest outcomes are most sensitive to teachers are not the ones on the margin of attending a selective college.

6.4 Implications for Teacher Effects on Underlying Skills

The results in Figure 3 suggest that the impact of nontest value-added on short-run and some long-run outcomes is larger for students at the bottom of the distribution. As pointed out by Jackson et al., (2022), this pattern could be driven by either (a) teachers truly having larger impacts on the soft skills of these students or (b) the students who are marginal for some of the underlying components of the nontest factor used to estimate nontest value-added (e.g., absences and suspensions) being the same as those who are marginal for binary long-run outcomes. Either of these explanations would tend to amplify the average effect of nontest value-added. However, in the latter case, the patterns observed in Figure 3 may not be reflective of true differences in teacher effects on unmeasured student skills across the skill distribution.

Jackson et al. (2022) propose a test measuring the relationship between the strength of the value-added effect in a given decile and how close students in that decile are to being marginal for an outcome (i.e., the average distance between the decile-specific mean and 0.5). We perform

this exercise in Appendix D and find evidence of mechanical heterogeneity for suspensions and grade retention for nontest value added; i.e., the largest impacts on these outcomes through nontest value-added are for students closest to a 0.5 base rate. In addition, we find similar evidence for college-going and four-year college-going for nontest value-added and for selective college-going for test value added. Thus, many of the relationships observed in Figure 3 are likely mechanical. Put another way, the differential between the test-based and nontest value-added findings for student outcomes is not *necessarily* evidence of heterogeneity of teacher effects on the underlying skills of their students. For example, it is not necessarily the case that teacher value-added predicting selective college-going for students at the top of the distribution means that teacher skills captured by test value-added are most impactful for high-achieving students. This is what one might conclude based on a study where selective college-going were the only long-run outcome available. However, it just happens to be the case that this outcome satisfies the conditions of having a long-run outcome sensitive to changes in the short-run outcome and the short-run outcome being sensitive to changes in teacher quality.

In the lower portion of the distribution, a similar argument holds for the impact of nontest value-added on college attendance. This illustrates a benefit using a continuous long-run outcome measure (college quality): it captures each of these marginal changes at different points in the student distribution to provide a more complete measure of teacher effects. While the estimated value-added effects on this measure are consistent with teachers who raise the nontest skills of their students having the largest impacts for students at the bottom of the distribution, estimates are very noisy.

7. Discussion

Teachers' test-based and nontest value-added both play important and explanatory roles for long-run student outcomes, including where a student enrolls in college. This high-level

finding masks important differences in mechanisms and distributional effects. The explanatory power of nontest value-added is primarily driven by the margins of high school graduation, college attendance, and four-year college attendance; and test-based value-added through attending a selective college. An important takeaway is that the teacher skills that 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 0.10).

These results 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 and for thinking about teacher assignments. For example, there is a literature that measures teacher quality gaps – advantaged students having access to better teachers – along one dimension: test score value-added (e.g., Isenberg et al., 2022; Goldhaber et al., 2017). However, the results of this study suggest that it is *nontest* teacher quality that is especially relevant for disadvantaged students and that gaps in access to effective teachers along the nontest dimension would be even greater cause for concern. Thus, the literature that measures teacher quality gaps only along one dimension 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

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of labor Economics*, 25(1), 95-135.

American Federation of Teachers. (2017). Federal Suit Settlement: End of Value-Added Measures for Teacher Termination in Houston. Retrieved from https://www.aft.org/press-release/federal-suit-settlement-end-value-added-measures-teacher-termination-houston

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.

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., & Koedel, C. (2022). Estimation and interpretation of teacher value-added in research applications. In Hanushek, E. A., Machin, S., & Woessmann, L. (Eds.), Handbook of the Economics of Education (in-press). Elsevier.

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

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational evaluation and policy analysis*, 31(4), 416-440.

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.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, 26(6), 673-682.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2010). Teacher credentials and student achievement in high school a cross-subject analysis with student fixed effects. Journal of Human Resources, 45(3), 655-681.

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).

Cunha, F., & Heckman, J. J. (2008). Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Journal of Human Resources, 43*(4), 738–782. https://doi.org/10.3368/jhr.43.4.738

Cunha, F., Heckman, J. J., & Schennach, S. M. (2010). Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Econometrica, 78*(3), 883–931. https://doi.org/10.3982/ECTA6551

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.

Dynarski, S. M., Hemelt, S. W., & Hyman, J. M. (2015). The missing manual: Using National Student Clearinghouse data to track postsecondary outcomes. *Educational Evaluation and Policy Analysis*, 37, 53S-79S.

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

Gershenson, S., Holt, S., & Tyner, A. (2022). Making the Grade: The Effect of Teacher Grading Standards on Student Outcomes.

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

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, 100(2), 267-271.

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

Isenberg, E., Max, J., Gleason, P., & Deutsch, J. (2022). Do Low-Income Students Have Equal Access to Effective Teachers?. *Educational Evaluation and Policy Analysis*, *44*(2), 234-256.

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.

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., 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

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.

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.

National Council on Teacher Quality. (2019). *State of the States 2019*.

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

Paige, M. (2020). Moving Forward While Looking Back: How Can VAM Lawsuits Guide Teacher Evaluation Policy in the Age of ESSA?. *Education Policy Analysis Archives*, *28*(64).

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

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.

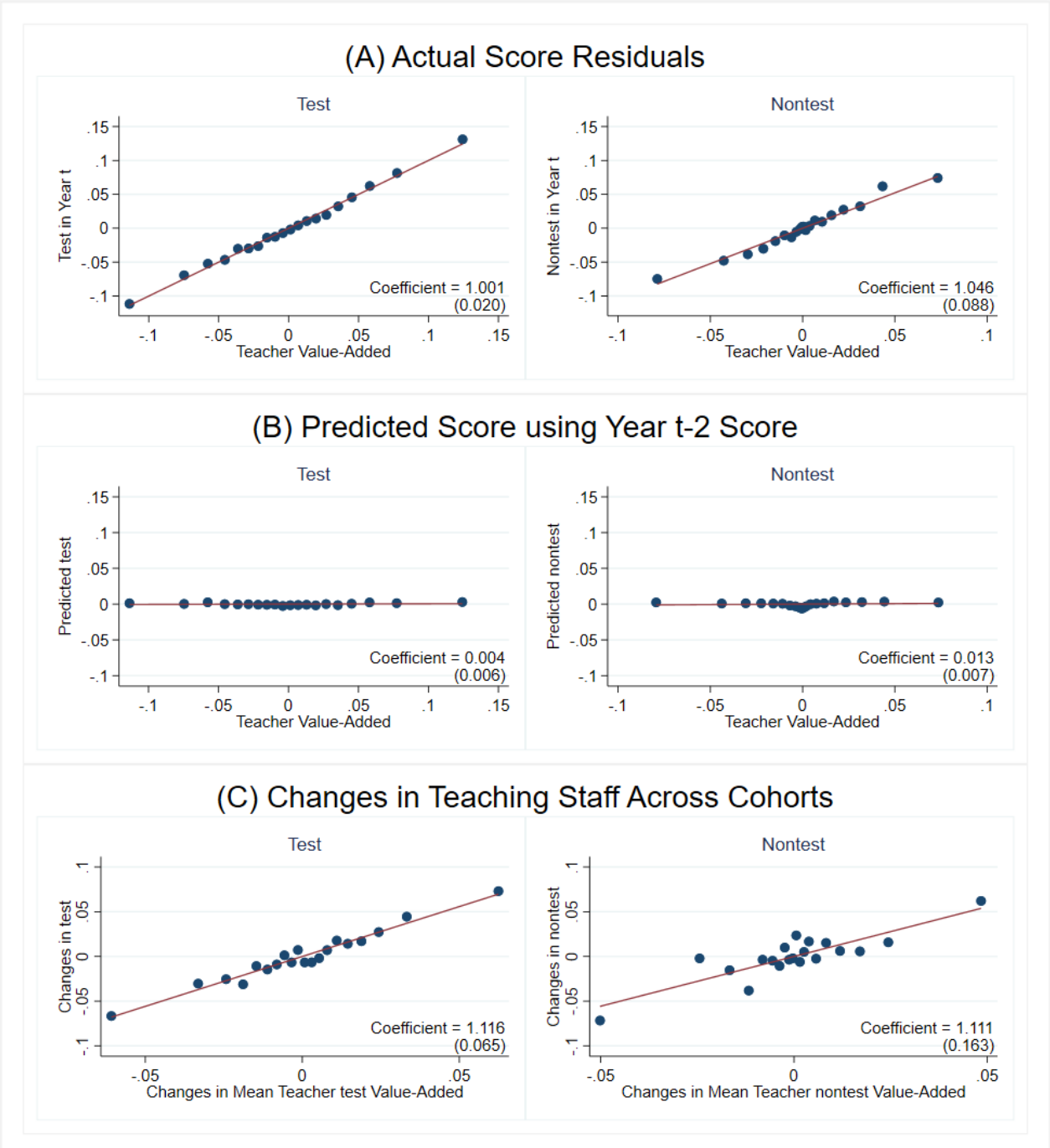
Sorensen, L. C. (2019). "Big data" in educational administration: An application for predicting school dropout risk. *Educational Administration Quarterly*, 55(3), 404–446.

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

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.

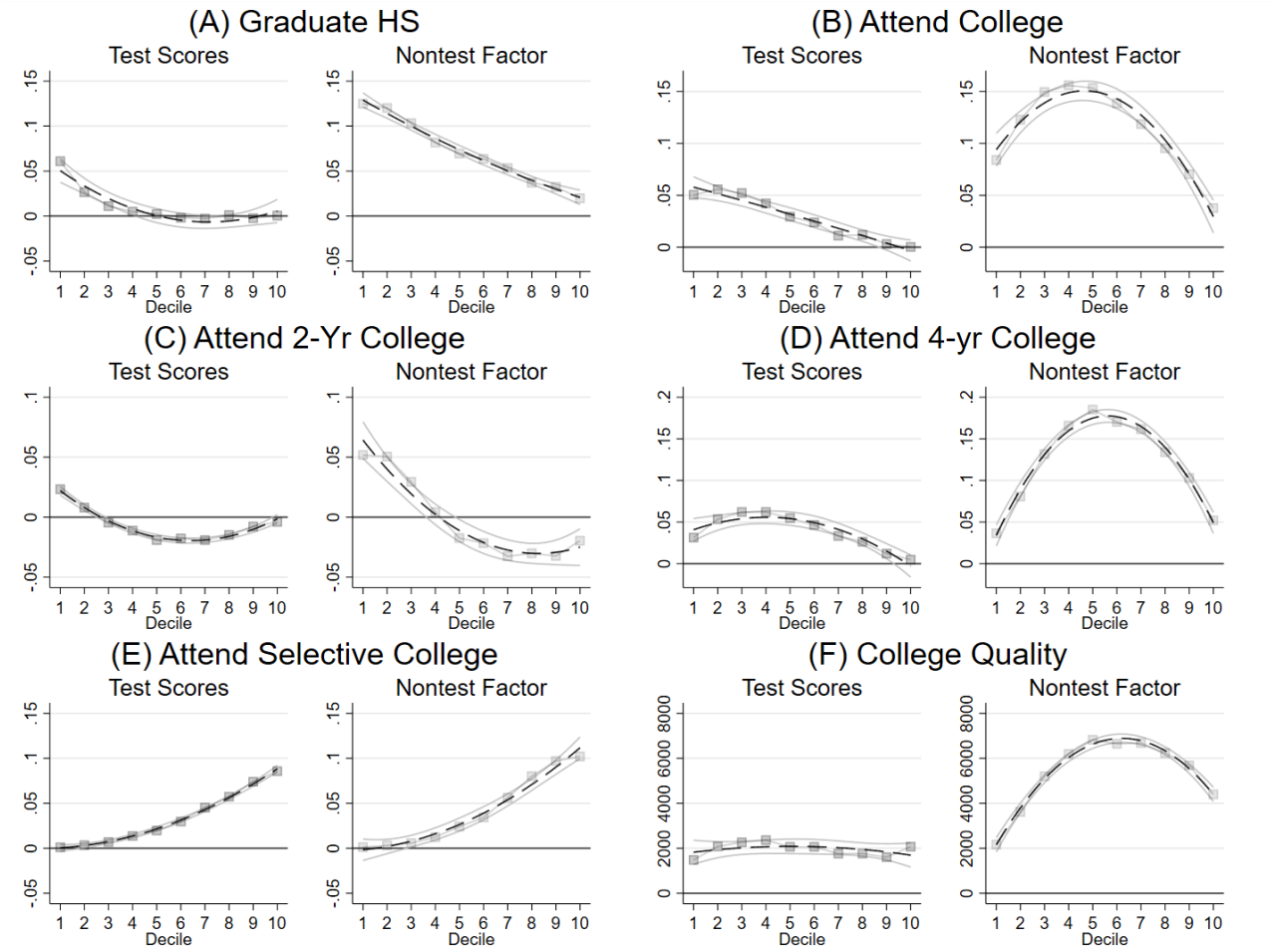
Figures and Tables

Figure 1. Effects of Value-Added on Actual and Predicted Scores



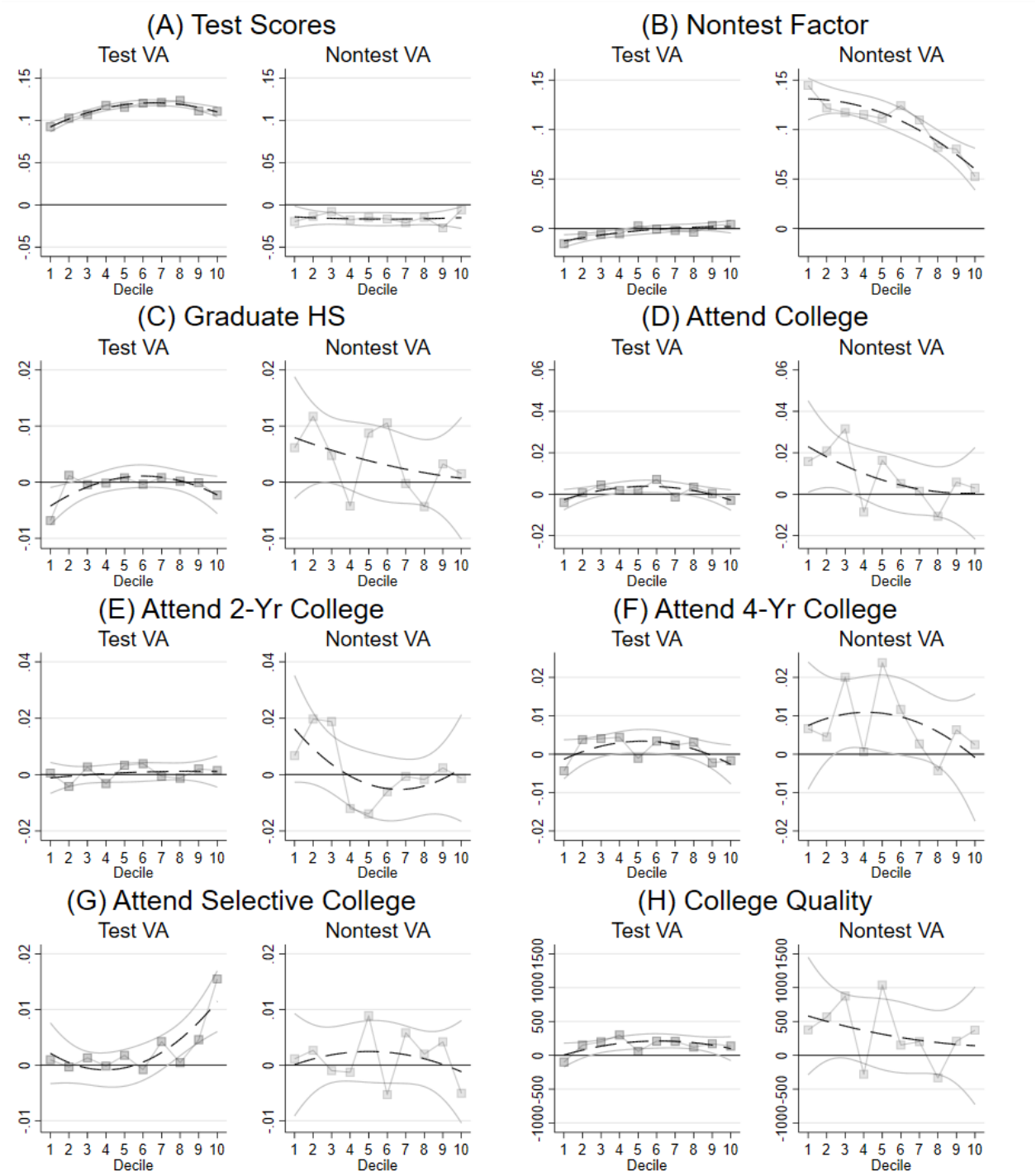
Notes: (A) Binned scatter plots of test score residuals versus test value-added (left) and nontest residuals versus nontest value-added (right). (B) Binned scatter plot of predicted test scores based on twice-lagged test scores versus test value-added (left) and predicted nontest outcomes using twice-lagged nontest factor versus nontest value-added (right). (C) Binned scatterplot of regression of school-grade-subject-year mean test scores (left) and nontest outcome (right) versus mean test value-added (left) and nontest value-added (right) as in Chetty et al. (2014a). This panel corresponds to column 1 in Table C1.

Figure 2. Partial Effect of Increases in Test and Nontest Outcomes by Educational Advantage



Notes: Each panel contains the results from 10 separate regressions, one from each decile of incoming educational advantage (the average of prior test scores and the nontest factor), which includes both test scores and nontest outcomes in year *t* as regressors. Each coefficient represents the impact of a one-standard deviation increase in test scores or the nontest factor on a given student outcome. Regressions are from the specification in Equation (2), which includes student demographic information, prior test and nontest outcomes, class-level averages of each, and school-track-year fixed effects. Each point represents the regression coefficient. In each panel, the 10 points are fitted with a quadratic fit (dashed line), and confidence intervals for the quadratic fit are shown.

Figure 3. Impact of Teacher Value Added by Educational Advantage



Notes: Each panel contains the results from 10 separate regressions, one from each decile of incoming educational advantage (the average of prior test scores and the nontest factor), which includes out-of-sample test value-added and nontest value-added as regressors. Each coefficient represents the impact of a one-standard deviation increase in test scores or the nontest factor on a given student outcome. Regressions are from the specification in Equation (2), which includes student demographic information, prior test and nontest outcomes, class-level averages of each, and school-track-year fixed effects. Each point represents the regression coefficient. In each panel, the 10 points are fitted with a quadratic fit (dashed line), and confidence intervals for the quadratic fit are shown.

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

			Impact of increase of one SD of teacher VA	
Study	Setting	Nontest	Test-based VA	Nontest VA
Chetty et al. (2014b)	Gr. 4-8, NYC, 1989-2009	--	College enroll 0.8 pp College quality $299 High Q college 0.7 pp	--
Jackson (2018) *Test-Nontest VA Correlation*:	Gr. 9, NC, 2005-2012 *0.15*	Factor[1]	HS grad: 0.1 % pts Take SAT: 1.2 % pts SAT score: 0.60 Intend 4yr: 0.1 % pts	HS grad: 1.5 % pts Take SAT: 0.1 % pts SAT score: -0.23 Intend 4yr: 1.3 % pts
Liu and Loeb (2021) *Test-Nontest VA Correlation*:	Gr. 7-11, CA district, 2004-2014 *0.12 (math), 0.08 (ELA)*	Unexc. absences	HS grad: 0.1 % pts AP courses: 0.02 AP credits: 0.11	HS grad: 0.7 % pts AP courses: 0.01 AP credits: 0.08
Mulhern and Opper (2022) *Test-Nontest VA Correlation*:	Gr. 5-7, NYC, 2005-2014 *0.04-0.06*	Attendance, grades[3]	HS grad: 0.2 % pts	HS grad: -.6 to -.8 % pts
Petek and Pope (2021) *Test-Nontest VA Correlation*:	Gr. 3-5, Los Angeles USD, 2003-2015 *0.15*	Factor[1]	Dropout: -0.2 % pts Held back: 0.1 % pts Take SAT: -0.2 % pts SAT score: 6.3 points	Dropout: -0.3 % pts Held back: -0.6 % pts Take SAT: 1.0 % pts SAT score: 2.0 points
Gilraine and Pope (2021) *Test-Nontest VA Correlation*:	Gr. 3-5, 1 large district, 2003-2017 *0.21*	Factor[1]	HS grad: 0.12 % pts Take SAT: 0.05 % pts SAT score: 2.9 points	HS grad: 0.83 % pts Take SAT: 0.33 % pts SAT score: 6.59 points
Rose et al., (2022) *Test-Nontest VA Correlation*:	Gr. 4-8, NC, 1996-2013 *0.06*	Factor[1,2]	HS grad: 0.11 pp Arrested: -0.08 pp	HS grad: 0.20 % pts Arrested: -0.36 % pts
This study *Test-Nontest VA Correlation*:	Gr. 7, 8, 10, MA, 2012-2021 *0.10*	Factor[1]	HS grad: 0.0 % pts Take SAT: 0.1 % pts SAT score: 0.01 sd AP tests passed: 0.03 College enroll: 0.2 % pts 4yr college: 0.1 % pts Selective college: 0.4 % pts College quality: $165	HS grad: 0.5 % pts Take SAT: 0.4 % pts SAT score: 0.00 sd AP tests passed: 0.00 College enroll: 1.1 % pts 4yr college: 0.8 % pts Selective college: 0.0 % pts College quality: $324

(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 in Table 1. The range of correlations listed in this row is for the correlation of math or ELA value-added and absence value-added over both grade levels (elementary or middle).

Table 2. Summary Statistics

	Short Run Measures			Long Run Outcomes		
	Mean	SD	N	Mean	SD	N
ELA test	-0.018	0.904	3031143	-0.028	0.895	1886138
Math test	-0.008	0.913	3035294	-0.017	0.901	1889099
Nontest index	0.036	0.962	2900555	0.050	0.943	1834642
Retained	0.006	0.080	3113737	0.008	0.087	1942322
Absences	8.683	10.266	3112733	8.625	10.326	1941571
Days suspended	0.247	1.897	3113737	0.253	1.973	1942322
GPA	2.961	0.902	2901226	2.904	0.898	1835118
Next year GPA	2.885	0.930	2827362	2.826	0.914	1830041
AP credits				4.273	8.426	1942322
AP tests taken				1.327	2.079	1942322
AP tests passed				0.887	1.783	1942322
Takes SAT				0.689	0.463	1578297
SAT scores (standard deviations)				0.067	1.001	1087702
Graduate				0.898	0.302	1942322
Dropout				0.034	0.182	1942322
Attends college				0.684	0.465	1942322
Attends 2 year college				0.181	0.385	1942322
Attends 4 year college				0.546	0.498	1942322
Median postsecondary income				35996.572	20344.520	1942322
College mobility				0.221	0.173	1942322
Lag math test	-0.010	0.918	2917783	-0.017	0.911	1814965
Lag ELA test	-0.021	0.910	2912736	-0.029	0.901	1810406
Lag retention	0.007	0.080	3038082	0.008	0.088	1900934
Lag absences	7.571	8.485	3027200	7.467	8.469	1893264
Lag days suspended	0.172	1.493	3038083	0.177	1.596	1900934
Lag GPA	3.015	0.879	2655416	2.958	0.875	1708530
Limited English proficient	0.050	0.218	3113737	0.043	0.203	1942322
Male	0.502	0.500	3113737	0.500	0.500	1942322
Free- or reduced-price lunch	0.347	0.476	3113737	0.352	0.478	1942322
Full inclusion special education	0.111	0.314	3113737	0.106	0.308	1942322
Partial inclusion special education	0.021	0.145	3113737	0.023	0.149	1942322
Substantially separate special education	0.006	0.076	3113737	0.006	0.080	1942322
Black student	0.118	0.323	3113737	0.115	0.319	1942322
Asian student	0.084	0.277	3113737	0.079	0.270	1942322
American Indian student	0.029	0.169	3113737	0.028	0.164	1942322
Pacific Islander student	0.010	0.099	3113737	0.010	0.100	1942322
Hispanic student	0.186	0.389	3113737	0.172	0.377	1942322
Takes advanced math	0.265	0.441	3113737	0.271	0.444	1942322
Takes art elective	0.280	0.449	3113737	0.219	0.414	1942322
Takes advanced language	0.105	0.306	3113737	0.143	0.350	1942322
Takes supplemental course	0.096	0.295	3113737	0.090	0.286	1942322

Table 3. Predicting Postsecondary Outcomes with Tests and the Nontest Factor

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: Dropout				Panel B: Graduate HS			
Tests	-0.01	-0.03			0.03	0.08		
Nontest Factor	-0.03		-0.03		0.07		0.08	
Retained				-0.00				-0.04
Log Absences				0.02				-0.04
Log Days Suspended				0.01				-0.02
GPA				-0.03				0.07
R^2	0.22	0.10	0.22	0.25	0.19	0.09	0.18	0.20
	Panel C: AP Tests Passed				Panel D: Take SAT			
Tests	0.94	1.09			0.12	0.20		
Nontest Factor	0.27		0.73		0.15		0.22	
Retained				0.98				-0.06
Log Absences				-0.11				-0.05
Log Days Suspended				0.18				-0.06
GPA				0.90				0.19
R^2	0.31	0.30	0.15	0.21	0.21	0.14	0.18	0.20
	Panel E: Attend College				Panel F: Attend 4-Year College			
Tests	0.10	0.19			0.15	0.26		
Nontest Factor	0.15		0.21		0.19		0.28	
Retained				-0.13				-0.24
Log Absences				-0.05				-0.05
Log Days Suspended				-0.06				-0.09
GPA				0.18				0.26
R^2	0.19	0.12	0.17	0.18	0.25	0.18	0.21	0.24
	Panel G: Attend Selective College				Panel H: College Quality Index			
Tests	0.10	0.14			7.13	11.05		
Nontest Factor	0.07		0.14		7.13		10.67	
Retained				-0.05				2.64
Log Absences				-0.01				-2.33
Log Days Suspended				-0.05				-1.21
GPA				0.19				11.18
R^2	0.25	0.21	0.15	0.21	0.31	0.24	0.25	0.29
Observations	1785735	1887997	1834642	1834642	1785735	1887997	1834642	1834642

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. For binary outcomes, coefficients obtained from taking the marginal effect of a logistic regression. Standard errors suppressed to fit table on one page; due to very large samples, every coefficient is statistically different from zero at the 1 percent level.

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

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Test Scores						
Test VA	0.12***	0.11***			0.12***	0.12***
	(0.00)	(0.00)			(0.00)	(0.00)
Nontest VA			0.00	0.00	-0.02***	-0.02***
			(0.01)	(0.01)	(0.00)	(0.00)
Panel B. Behavioral Factor						
Test VA	0.00	0.00			-0.00	-0.00
	(0.00)	(0.00)			(0.00)	(0.00)
Nontest VA			0.13***	0.12***	0.14***	0.12***
			(0.01)	(0.01)	(0.01)	(0.01)
Panel C. GPA						
Test VA	-0.00	-0.00			-0.01***	-0.01***
	(0.00)	(0.00)			(0.00)	(0.00)
Nontest VA			0.18***	0.17***	0.18***	0.17***
			(0.01)	(0.01)	(0.01)	(0.01)
Panel D. Absences						
Test VA	-0.006***	-0.007***			-0.005**	-0.006***
	(0.002)	(0.002)			(0.002)	(0.002)
Nontest VA			-0.018***	-0.009*	-0.017***	-0.008*
			(0.005)	(0.005)	(0.005)	(0.005)
Panel E. Days Suspended						
Test VA	-0.001	-0.001			-0.001	-0.001
	(0.001)	(0.001)			(0.001)	(0.001)
Nontest VA			-0.007***	-0.007***	-0.007***	-0.006***
			(0.002)	(0.002)	(0.002)	(0.002)
Panel F. Retained						
Test VA	0.02	0.02			0.03	0.03
	(0.02)	(0.02)			(0.02)	(0.02)
Nontest VA			-0.21***	-0.21***	-0.22***	-0.21***
			(0.04)	(0.04)	(0.04)	(0.04)
N	2586301	2535540	2557576	2506967	2555993	2505463
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.

Table 5. Teacher Effects on Students' Secondary Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. AP Credits						
Test VA	0.22***	0.10***			0.21***	0.10***
	(0.07)	(0.03)			(0.08)	(0.04)
Nontest VA			0.06	-0.02	0.02	-0.04
			(0.10)	(0.06)	(0.10)	(0.06)
Panel B. AP Tests Taken						
Test VA	0.06***	0.04***			0.06***	0.04***
	(0.01)	(0.01)			(0.01)	(0.01)
Nontest VA			0.00	-0.01	-0.01	-0.01
			(0.02)	(0.02)	(0.02)	(0.02)
Panel C. AP Tests Passed						
Test VA	0.05***	0.03***			0.05***	0.03***
	(0.01)	(0.01)			(0.01)	(0.01)
Nontest VA			-0.02	-0.02	-0.03	-0.02
			(0.02)	(0.01)	(0.02)	(0.01)
Panel D. Took SAT						
Test VA	-0.02	0.08			-0.03	0.07
	(0.18)	(0.14)			(0.18)	(0.14)
Nontest VA			0.46	0.40	0.47	0.39
			(0.41)	(0.38)	(0.41)	(0.38)
Panel E. SAT Scores (standard deviations)						
Test VA	0.02***	0.01***			0.02***	0.01***
	(0.00)	(0.00)			(0.00)	(0.00)
Nontest VA			-0.01	-0.00	-0.01	-0.00
			(0.01)	(0.01)	(0.01)	(0.01)
Panel F. Graduate HS						
Test VA	0.01	0.02			0.00	0.00
	(0.07)	(0.08)			(0.07)	(0.08)
Nontest VA			0.52**	0.54**	0.52**	0.54**
			(0.22)	(0.24)	(0.22)	(0.24)
Panel G. Dropout						
Test VA	0.03	0.01			0.03	0.03
	(0.04)	(0.05)			(0.04)	(0.05)
Nontest VA			-0.15	-0.26**	-0.16	-0.27**
			(0.12)	(0.13)	(0.12)	(0.13)
N	1621006	1581369	1608889	1569318	1607537	1568043
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.

Table 6. Teacher Effects on Postsecondary Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Enroll in College						
Test VA	0.22*	0.15			0.19	0.14
	(0.13)	(0.12)			(0.13)	(0.12)
Nontest VA			1.15***	1.10***	1.12***	1.08***
			(0.34)	(0.34)	(0.34)	(0.34)
Panel B. Enroll in Two-year College						
Test VA	0.19	0.09			0.19	0.08
	(0.13)	(0.11)			(0.13)	(0.12)
Nontest VA			0.38	0.26	0.35	0.25
			(0.33)	(0.32)	(0.34)	(0.33)
Panel C. Enroll in Four-Year College						
Test VA	0.04	0.11			0.03	0.10
	(0.15)	(0.12)			(0.15)	(0.13)
Nontest VA			0.89**	0.83**	0.89**	0.82**
			(0.41)	(0.38)	(0.41)	(0.38)
Panel D. Enroll in Selective College						
Test VA	0.59***	0.40***			0.59***	0.40***
	(0.14)	(0.11)			(0.14)	(0.11)
Nontest VA			0.04	0.06	-0.04	0.00
			(0.22)	(0.19)	(0.23)	(0.19)
Panel E. College Quality Index ($)						
Test VA	174.8***	164.8***			170.6***	164.5***
	(50.4)	(48.1)			(50.9)	(48.0)
Nontest VA			398.8***	323.8**	374.1**	300.6**
			(152.4)	(133.1)	(153.6)	(132.8)
N	1621006	1581369	1608889	1569318	1607537	1568043
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.

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 A1, 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 A2 through A4. 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 A2 through A4. 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 A2 and A3, 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 A2 and A3, 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 A1. Distribution of the Number of Teachers per Track

Number of Teachers	VA: School-Grade	VA: School-Track	LR: School-Grade	LR: School-Track
1	41582	132502	47606	124155
2	99646	260533	127542	228731
3	179756	342344	179671	288014
4	215080	363909	191448	244810
5	275270	334234	178299	219297
6	248733	298912	160457	175350
7	246530	237419	117284	140951
8	176665	200050	100784	111038
9	183389	194989	81585	90815
10	175829	154790	67230	78657
11	128890	100409	50696	55503
12	129854	93835	57657	42433
13	97499	76039	51485	38259
14	88755	66839	52181	41517
15+	826259	256933	563694	148089

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

Table A2. 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 A3. 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 A4. 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	0.30***	0.30***			0.30***	0.30***
	(0.02)	(0.02)			(0.02)	(0.02)
Nontest VA			0.06	0.08*	0.02	0.03
			(0.04)	(0.04)	(0.04)	(0.04)
Panel B. Performance Rating: Curriculum Planning and Assessment						
Test VA	0.14***	0.15***			0.14***	0.15***
	(0.01)	(0.01)			(0.01)	(0.01)
Nontest VA			0.02	0.03	-0.00	0.00
			(0.02)	(0.02)	(0.02)	(0.02)
Panel C. Performance Rating: Teaching All Students						
Test VA	0.14***	0.14***			0.14***	0.14***
	(0.01)	(0.01)			(0.01)	(0.01)
Nontest VA			0.03	0.04*	0.01	0.02
			(0.02)	(0.02)	(0.02)	(0.02)
Panel D. Performance Rating: Family and Community Engagement						
Test VA	0.05***	0.05***			0.06***	0.06***
	(0.01)	(0.01)			(0.01)	(0.01)
Nontest VA			0.01	0.01	0.00	0.00
			(0.02)	(0.02)	(0.02)	(0.02)
Panel E. Performance Rating: Professional Culture						
Test VA	0.10***	0.10***			0.10***	0.11***
	(0.01)	(0.01)			(0.01)	(0.01)
Nontest VA			0.02	0.03	0.01	0.01
			(0.02)	(0.02)	(0.02)	(0.02)
Panel F. Communications and Literacy Skills						
Test VA	0.07***	0.08***			0.07***	0.08**
	(0.03)	(0.03)			(0.03)	(0.03)
Nontest VA			-0.04	-0.02	-0.05	-0.04
			(0.05)	(0.05)	(0.05)	(0.05)
Panel G. Subject Matter Knowledge						
Test VA	0.18***	0.19***			0.18***	0.19***
	(0.03)	(0.04)			(0.03)	(0.04)
Nontest VA			-0.03	-0.02	-0.06	-0.05
			(0.05)	(0.06)	(0.06)	(0.06)
Sch-Grade-Year FE	Y		Y		Y	
Sch-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.

Appendix C. Robustness Checks and Alternate Specifications

Table C1. Quasi-Experimental Estimates of Forecast Bias

	Δ Score	Δ Score	Δ Score	Δ Other Subj. score	
				Secondary	Elementary
	(1)	(2)	(3)	(4)	(5)
Panel A. Test scores: grades 4-8 *Note: for comparison only. Not used in paper*					
Changes in Mean Teacher VA	1.043	0.871	1.037	0.039	0.212
Across Cohorts	(0.038)	(0.048)	(0.065)	(0.051)	(0.043)
Sch x Grade x Subj x Year Cells	38623	38623	13856	13880	22250
Panel B. Test scores: grades 7-8 and 10					
Changes in Mean Teacher VA	1.116	0.920	0.860	0.064	
Across Cohorts	(0.065)	(0.095)	(0.131)	(0.060)	
Sch x Grade x Subj x Year Cells	14854	14854	9523	13599	
Panel C. Nontest factor: grades 7-8 and 10					
Changes in Mean Teacher VA	1.111	0.076[1]	0.208[1]	0.569	
Across Cohorts	(0.163)	(0.243)	(0.295)	(0.132)	
Sch x Grade x Subj x Year Cells	14007	14007	9027	12823	
Year Fixed Effects	Y			Y	Y
School-year Fixed Effects[1]		Y	Y		
Lagged Score Controls			Y		
Lead and Lag Changes in Teacher VA			Y		
Other-Subject Change in Mean Teacher VA				Y	Y

Notes: Replication of Chetty et al. (2014a) Table 4. Panel A displays results for test scores from grades 4-8 for comparison with prior work. Panels B and C use forecasts of teacher value-added used in this paper.

[1] We include these results for comparison's sake, but school-year fixed effects are conceptually challenging in the case of nontest value-added. This is because, unlike with test scores, nontest outcomes are not subject-specific. Thus, increases in the nontest value-added in one subject mechanically increases the nontest value-added in the other subject in that school (Panel C, Column 4), and thus a school fixed effect model effectively differences out some of the effect of increases in nontest outcomes caused by increases in a new teacher's nontest value-added.

Table C2. Results with Grades 7 and 8 Only

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Graduate High School						
Test VA	0.08	0.03			0.05	-0.00
	(0.10)	(0.10)			(0.10)	(0.11)
Nontest VA			0.57**	0.57**	0.56**	0.57**
			(0.23)	(0.25)	(0.24)	(0.25)
Panel B. College Quality Index						
Test VA	98.8	115.7*			89.1	115.8*
	(64.9)	(65.6)			(66.1)	(65.7)
Nontest VA			406.3**	287.7**	394.1**	271.4*
			(161.2)	(139.3)	(163.3)	(139.5)
Panel C. Enroll in College						
Test VA	0.24	0.15			0.19	0.13
	(0.15)	(0.16)			(0.15)	(0.16)
Nontest VA			1.20***	1.00***	1.17***	0.98***
			(0.36)	(0.35)	(0.36)	(0.35)
Panel D. Enroll in Four-Year College						
Test VA	0.21	0.19			0.17	0.17
	(0.16)	(0.16)			(0.16)	(0.16)
Nontest VA			1.07**	0.82**	1.04**	0.79*
			(0.43)	(0.40)	(0.44)	(0.40)
Panel E. Enroll in Selective College						
Test VA	0.26**	0.28**			0.26**	0.27**
	(0.12)	(0.11)			(0.12)	(0.11)
Nontest VA			-0.14	0.04	-0.18	-0.00
			(0.22)	(0.20)	(0.22)	(0.20)
N	938816	930063	926022	917387	925613	916976
School-Grade-Year FE	Y		Y		Y	
School-Track-Year FE		Y		Y		Y

Notes*:* Coefficients on test value-added and nontest value-added from regressions with student secondary and postsecondary outcomes as the dependent variable. The sample includes all students in the matched long-run sample described in the text with the exception of tenth graders.

Table C3. Results with Nontest Value-Added Excluding GPA

	(1)	(2)	(3)	(4)
Panel A. AP Tests Passed				
Nontest VA	-0.00	-0.02	-0.01	-0.02
	(0.02)	(0.01)	(0.02)	(0.01)
Test VA			0.05***	0.03***
			(0.01)	(0.01)
Panel B. SAT Scores (standard deviations)				
Nontest VA	0.01	0.01	0.00	0.00
	(0.01)	(0.01)	(0.01)	(0.01)
Test VA			0.02***	0.01***
			(0.00)	(0.00)
Panel C. Graduate High School				
Nontest VA	0.47	0.56*	0.48	0.57*
	(0.30)	(0.29)	(0.30)	(0.29)
Test VA			0.01	0.02
			(0.07)	(0.07)
Panel D. College Quality				
Nontest VA	266.3	316.0*	255.1	306.5*
	(163.8)	(168.9)	(163.4)	(167.7)
Test VA			173.7***	163.3***
			(50.4)	(48.0)
Panel E. Enroll in College				
Nontest VA	1.31***	1.39***	1.30***	1.39***
	(0.44)	(0.47)	(0.43)	(0.46)
Test VA			0.21	0.15
			(0.13)	(0.12)
Panel F. Enroll in Four-Year College				
Nontest VA	0.67	0.82*	0.66	0.81*
	(0.42)	(0.44)	(0.42)	(0.44)
Test VA			0.04	0.10
			(0.15)	(0.12)
Panel G. Enroll in Selective College				
Nontest VA	-0.22	-0.28	-0.25	-0.31
	(0.29)	(0.26)	(0.29)	(0.26)
Test VA			0.59***	0.40***
			(0.14)	(0.11)
N	1622478	1582746	1621006	1581369
School-Grade-Year FE	Y		Y	
School-Track-Year FE		Y		Y

Notes*:* Coefficients on test scores and nontest factor from regressions with student secondary and postsecondary outcomes as the dependent variable. The nontest value-added measure is constructed using an alternate measure of the nontest factor that excludes grade point average.

Table C4. Robustness to Alternative Specifications

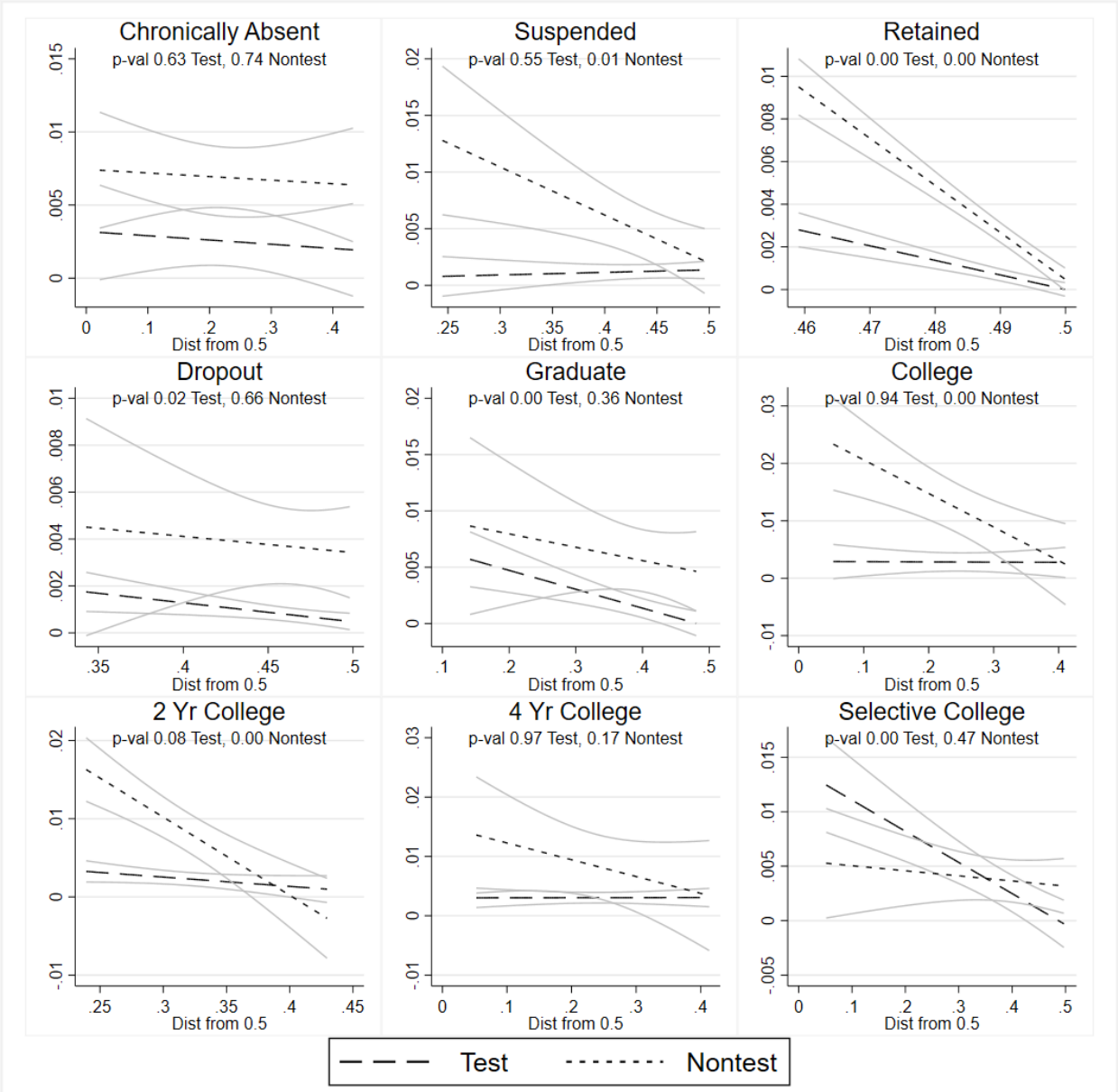
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. AP Tests Passed								
Test VA	0.03***	0.03***	0.03***	0.03				
	(0.01)	(0.01)	(0.01)	(0.02)				
Nontest VA					-0.02	-0.01	-0.01*	-0.04*
					(0.01)	(0.01)	(0.01)	(0.02)
Panel B. SAT Scores (standard deviations)								
Test VA	0.01***	0.01***	0.01***	0.02				
	(0.00)	(0.00)	(0.00)	(0.01)				
Nontest VA					-0.00	0.00	-0.00	0.01
					(0.01)	(0.00)	(0.00)	(0.02)
Panel C. Graduate High School								
Test VA	0.02	0.04	0.02	0.26				
	(0.08)	(0.07)	(0.07)	(0.23)				
Nontest VA					0.54**	0.37**	0.24**	0.06
					(0.24)	(0.17)	(0.12)	(0.42)
Panel D. College Quality								
Test VA	165***	195***	154***	346*				
	(48)	(48)	(42)	(185)				
Nontest VA					324**	296***	159***	173
					(133)	(101)	(59)	(293)
Panel E. Enroll in College								
Test VA	0.15	0.24**	0.10	0.77*				
	(0.12)	(0.12)	(0.10)	(0.42)				
Nontest VA					1.10***	0.96***	0.43***	0.47
					(0.34)	(0.26)	(0.16)	(0.71)
Panel F. Enroll in Four-Year College								
Test VA	0.11	0.15	0.09	1.08**				
	(0.12)	(0.12)	(0.11)	(0.45)				
Nontest VA					0.83**	0.82***	0.45**	0.11
					(0.38)	(0.29)	(0.18)	(0.65)
Panel G. Enroll in Selective College								
Test VA	0.40***	0.42***	0.42***	0.31				
	(0.11)	(0.10)	(0.10)	(0.26)				
Nontest VA					0.06	-0.00	0.00	0.50
					(0.19)	(0.15)	(0.11)	(0.42)
N	1581369	1379647	1581369	8489	1569318	1374526	1569318	8489
Baseline Model	Y				Y			
Other-subj VA		Y				Y		
Resid on Tch			Y				Y	
Switching Design				Y				Y

Notes: Columns 1 and 5 re-produce the baseline models in Tables 5 and 6. Columns 2 and 6 include controls for the test and nontest value-added of student's teacher in other subjects. Columns 3 and 7 use teacher fixed effects rather than school fixed effects in the first stage residualization process to estimate the relationship between outcome (test or nontest) and controls. Columns 4 and 8 use grade-school-year aggregates of value-added and outcomes as in Chetty et al. (2014b).

Appendix D. Tests for Mechanical Heterogeneity

One possible explanation for the impact of nontest value-added being largest on the nontest factor for students at the bottom of the educational advantage distribution is that these students are more likely to be marginal for outcomes like absences, suspensions, and grade repetition. As with Jackson et al. (2022), we find support for this explanation in that the relationship between nontest value added and whether a student is suspended or repeats a grade is strongest for students who are closest to the margin of these binary outcomes (i.e., close to 50 percent probability); results are displayed in Table D1 below. In addition, we see find similar support for the relationship between test value added and selective college-going being largest for those close to the margin, and for nontest value added and college attendance.

Figure D1. Tests for Mechanical Heterogeneity



Notes: Each panel contains the results from 10 separate regressions, one from each decile of incoming educational advantage (the average of prior test scores and the nontest factor), which includes out-of-sample test value-added and nontest value-added as regressors. The y-axis represents estimated impact of a given value-added type on that decile, and the x-axis the distance between the mean of the outcome in that decile and 0.5. Regressions are from the specification in Equation (2), which includes student demographic information, prior test and nontest outcomes, class-level averages of each, and school-track-year fixed effects. The p-values shown in each plot are tests for whether the fitted slope is equal to zero.