Within-School Heterogeneity in Quality: Do Schools Provide Equal Value Added to All Students?

Naven, Matthew

Washington and Lee University

5 May 2020

Within-School Heterogeneity in Quality: Do Schools Provide Equal Value Added to All Students?

Matthew Naven*

Washington and Lee University

Draft Date: May 5, 2020

Abstract

Low-socioeconomic status (SES), minority, and male students perform worse than their high-SES, non-minority, and female peers on standardized tests. This paper investigates how within-school differences in school quality contribute to these educational achievement gaps by SES, ethnicity, and sex. Using individual-level data on the universe of public-school students in California, I estimate school quality using a value added methodology that accounts for the fact that students sort to schools on observable characteristics. I run three separate analyses, in which I allow each school to provide a distinct value added to their low-/high-SES, minority/non-minority, and male/female students. I find that there is within-school heterogeneity in value added by SES, ethnicity, and sex, as on average schools provide less value added to their low-SES, minority, and male students. Thus within-school heterogeneity in quality is one factor that contributes to differential outcomes for disadvantaged students. [JEL Codes: **I24, H75, I21, I23, J24**]

1 Introduction

Despite efforts to close them in recent decades, achievement gaps in education by socioeconomic status (SES), ethnicity, and sex have persisted. The test score gap between low- and high-income students has

*Contact the author at mnaven@wlu.edu or visit www.matthewnaven.com for more information. I am grateful to Scott Carrell, Marianne Bitler, Shu Shen, Michal Kurlaender, Paco Martorell, Sherrie Reed, and other members of the California Education Lab for helpful comments and suggestions. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305E150006 to the Regents of the University of California. The opinions expressed are those of the author and do not necessarily represent views of the Institute, the U.S. Department of Education, or of the agencies providing data. I gratefully acknowledge support from the Russell J. and Dorothy S. Bilinski Educational Foundation Dissertation Writing Fellowship. I would like to thank the participants at the University of California, Davis, the University of Nevada, Reno, the All-California Labor Economics Conference, Sonoma State University, APPAM, Analysis Group, Cornerstone Research, Kennesaw State University, IMPAQ International, the Federal Communications Commission, Washington and Lee University, AEFP, and the APPAM California Regional Student Conference seminars for their insights.

increased during the past 50 years (Reardon and Robinson, 2008; Reardon, 2011), as has the college-going gap (Bailey and Dynarski, 2011) and the elite college-going gap (Reardon, Baker and Klasik, 2012) even after accounting for academic ability (Belley and Lochner, 2007; Karen, 2002). Moreover, the low-/high-income test score gap only decreases slightly as students progress through school (Reardon, 2011). The white-black and white-Hispanic test score gap is already 0.5 to 0.8 standard deviations by the time students enroll in kindergarten (Reardon and Portilla, 2016), and on average white students score one and half or more grade levels higher than black and Hispanic students enrolled in socioeconomically similar school districts (Reardon, 2016). While the black-white gap has shrunk over time, it appears to widen during the school years (Reardon and Robinson, 2008). The Hispanic-white gap varies considerably by origin, but narrows for all subgroups in the earliest grades (Reardon and Galindo, 2009). Gender differences in education outcomes are less straightforward. Women perform better on average than males on ELA exams (Chatterji, 2006; Cimpian et al., 2016; Fryer Jr and Levitt, 2010; Husain and Millimet, 2009; Lee, Moon and Hegar, 2011; Penner and Paret, 2008; Robinson and Lubienski, 2011; Sohn, 2012), yet men are more likely to study lucrative STEM fields (Kahn and Ginther, 2017).

While many factors likely play a role in these achievement gaps, one factor that remains unstudied is within-school heterogeneity in the quality that schools provide to their students. Schools are a large component of early human capital formation, and the previous literature on school and teacher quality largely assumes that good schools and teachers are beneficial for *all* students attending that school. By reporting only a single measure for each school or teacher, these studies calculate the average effect of a school or teacher on its students. Although a measure of overall school quality is undoubtedly important, it is not necessarily informative to whether a school will be effective for a *specific type* of student. Average measures ignore the possibility that schools provide a higher-quality education to certain subgroups of students which may in turn accentuate or attenuate existing achievement gaps.

For example, students have been shown to perform better (Dee, 2004, 2007; Egalite, Kisida and Winters, 2015; Gershenson et al., 2017) and are less likely to be disciplined (Dee, 2005; Lindsay and Hart, 2017; Holt and Gershenson, 2017) when assigned to a same-race or same-sex teacher. Likewise, teachers report higher evaluations of students who share the same race or sex (Dee, 2005). Thus students may perform better when they attend a school with faculty that look like them. Low-income students may benefit from attending a school that is familiar with the domestic issues that this population faces, or they may benefit from being surrounded with high-income peers that are more likely to have institutional knowledge on academic resources. Yet there is a dearth of research on the student-school match[1], despite the fact that there is

[1] To my knowledge, only two economics papers explore how value added may differ by subgroup. Carrell, Page and West (2010) find that female students assigned to female professors in STEM classes perform better and are more likely to obtain a STEM degree than their fellow female students who are assigned a male professor. Bau (2015) explores how private schools in

substantial heterogeneity in teacher characteristics and other inputs at the school level.

This paper explores whether school quality differs by SES, ethnicity, and sex within schools. Using individual-level data on the universe of California public school students linked to postsecondary records, I estimate school by subgroup quality by applying the value added with drift methodology, as in Chetty, Friedman and Rockoff (2014), to schools. The value added methodology accounts for the fact that students do not randomly sort to schools and subgroups of students have varying levels of average academic performance. The drift methodology, which allows value added to change from year to year, is particularly suited to the school quality setting, as schools experience faculty and staff turnover that could lead to changes in quality from year to year. I estimate a school's value added on both standardized test scores and postsecondary enrollment. I perform three separate analyses, one each for SES, ethnicity, and sex, and I allow each school to have two value added estimates: one for each dichotomous group (low-/high-SES, minority/non-minority, and male/female). For example, School A will have one estimate of the value added it provides to low-SES students and one estimate of the value added it provides high-SES students. Simulations suggest that assuming homogeneity in school value added can incorrectly rank schools in the presence of within-school heterogeneity in value added. Allowing for this within-school heterogeneity when estimating school value added restores the correct ranking of schools.

Results show that there is indeed within-school heterogeneity in the value added that schools provide to students. On average schools provide less value added to low-SES, minority, and male students. The average school increases the test scores of its low-SES students by 0.02 student-level standard deviations less than it does for its high-SES students. The corresponding minority/non-minority and male/female difference is 0.03 standard deviations. These differences are equivalent to the difference between attending an average school and a school that is between 0.2 and 0.3 standard deviations in school value added below average. The results are even larger when examining within-school heterogeneity in the value added that schools provide on postsecondary enrollment. The average school increases the likelihood that a low-SES students attends a postsecondary institution by four percentage points less than it does for a high-SES student. This is a 7% reduction in the likelihood of attending a college or university relative to the baseline average of 58% for low-SES students. For minorities the average school increases the likelihood of postsecondary enrollment by three percentage points less than for non-minorities and for males the difference is four percentage points less. Interestingly, the gap in school value added on test scores tends to decrease as students age, while the gap in school value added on postsecondary enrollment does not converge between subgroups by the time students reach 11th grade. Thus schools measured as providing equal value added to all subgroups of students on test scores may still be failing their disadvantaged students on measures of long-term success.

Pakistan compete in horizontal quality when students respond differentially to match quality by SES.

This paper adds numerous important contributions to the literature on education quality and achievement gaps. First, this paper provides some of the first evidence on whether schools differentially affect test scores and postsecondary success for low-SES, minority, and male students, which is especially relevant given the large postsecondary enrollment gap by SES and ethnicity. This paper is the first to extend the value added with drift methodology (Chetty, Friedman and Rockoff, 2014) to schools allowing for within-school heterogeneity. Second, this paper proves that if the assumption that schools provide homogenous school value added to their students is violated then a ranking system based on a single value added estimate can incorrectly rank the quality of schools. Finally, as with Naven (2019), this paper is unique in that it links the universe of public school students in California, which has the largest public school population in the United States, to their postsecondary outcomes. California is a particularly relevant state in which to study postsecondary outcomes because California has a robust postsecondary infrastructure that includes two-year community colleges, teaching universities, and globally-ranked research universities.

2 Simulation

A significant advantage of allowing a school's value added to vary by subgroup is that it corrects for the incorrect ranking of schools if the assumption of a homogeneous school effect is violated. In order to demonstrate this, I simulate a dataset of student observations. Let true value added for school s in year t be μ_{sLt} for low-type students and μ_{sHt} for high-type students. Let student test scores be generated according to the true model in equation (1), with school by type by year value added μ_{sdt}, school by year common shocks θ_{st}, and a noise term ε_{isdt} all contributing.

$$z_{isdt} = \mu_{sdt} + \theta_{st} + \varepsilon_{isdt} \tag{1}$$

For simplicity, let $\mu_{sdt} = \mu_{sd}\ \forall t$. Let there be four types of schools with equal probability, A, B, C, and D, with true school value added for low- and high-type distributed as μ_{sLt} and μ_{sHt} in columns 2 and 3 of Table 1. School A and school D are unambiguously the best and worst schools, respectively. School B is particularly effective with low-type students while school C is effective with high-type students. Notice that on average schools provide more value added to high-type students.

Let N_{sLt} and N_{sHt} be the number of students of low- and high-type respectively. I assign each school a baseline value of N_{sLt} and N_{sHt} and allow the yearly population of low- and high-type students to fluctuate within a window of this baseline by multiplying N_{sLt} and N_{sHt} separately by an inflation term j_{sdt}. This allows for yearly variation in the proportion of students that are low-type within a school. The distributions for θ_{st}, ε_{isdt}, N_{sLt}, N_{sHt}, and j_{sdt}, as well as the number of schools and years, are given in Table 2.

Figure 1 plots the value added estimates obtained using the drift methodology in Chetty, Friedman and Rockoff (2014) for the simulated data. Figures 1a, 1b, 1c, and 1d contain both the estimates under the assumption of a homogenous effect for all students and a heterogeneous effect for low- and high-type students for schools of type A, B, C, and D respectively. The horizontal axis plots the proportion of students who are low-type within the school. The horizontal red lines mark the true value added for each school. For each school type allowing for a heterogeneous value added effect produces estimates clustered around the true value regardless of what proportion of students are low-type. Imposing the assumption of a homogeneous value added effect, however, causes the value added estimate to fluctuate according the proportion of low-type students. While this issue is less pronounced when schools have similar value added for both low- and high-type students, such as schools A and D, the value added estimate can vary dramatically according to the proportion of low-type students when schools are more effective with a particular type of student, as with schools B and C.

Even worse, the assumption of homogeneous school value added can incorrectly rank schools when they are at the extreme end of the distribution of proportion low-type students. Figure 1e plots the homogeneous estimate for schools A and C. Despite the fact that school A is better than school C, since it is equally effective with high-type students but better with low-type students, its estimated effect is worse than school C if A enrolls a large proportion of low-type students and C enrolls a large proportion of high-type students. Allowing the schools to have heterogeneous effects, as in figure 1f, preserves the correct ranking of schools. This ranking error illustrates the importance of studying within-school heterogeneity in value added.

3 Data

My study uses individual-level data on the universe of public school students in the state of California. Standardized test score information comes from the California Standards Test (CST). Data from the CST spans the 2002-2003 to 2012-2013 school years[2] and tests students in English language arts (ELA) and math during grades 2-11. This paper focuses on the ELA exam because students take non-comparable math tests starting in the seventh grade. The data also include demographic information on each student, such as sex, race, economic disadvantage status, limited English proficiency status, and whether or not the student has a disability. State student IDs can be used to link students to prior test scores across time. Each cohort consists of about 475,000 students.

[2]Due to the fact that I use test scores from two grades prior as a control variable, I only calculate value added estimates for the years 2004-2005 to 2012-2013.

I assign students to low- or high-SES status based on whether they are defined as socioeconomically disadvantaged[3] by the California Department of Education (CDE). In order to get an idea of the income level of these students, figure 2 plots the distribution of total household income in 2017 dollars by socioeconomic disadvantage status from the Survey of Income and Program Participation (SIPP).[4] Economically disadvantaged students live in households with a median income of about $60,000, which is about $36,000 lower that the median income of students in households that are not economically disadvantaged. However, the peak of the income distribution for economically disadvantaged students occurs much lower at about $38,000. I define minority students as Hispanic, black, Native American and two or more races.

Table 3 gives summary statistics for the CST data by subgroup for the test score value added samples and includes all the dependent and independent variables used in the value added analyses. Appendix Table A.1 shows the limitations that are imposed in order to form the value added sample, which are similar to those made in the teacher value added literature. The vast majority of students in the CST data that cannot be included in the value added estimation are excluded because they lack prior test scores. I exclude grades 2-3 because they lack sufficient prior test scores in order to estimate value added.[5]

Unsurprisingly, SES and minority status are highly correlated. Low-SES students are a little less than 50 percentage points less likely to be white or Asian. Minorities are 50 percentage points more likely to be socioeconomically disadvantaged than their non-minority peers. Low-SES and minority students are also much more likely to be limited English proficient, which is likely due to Spanish-speaking Hispanics. Low-SES and minority students perform much worse on standardized tests, as their average ELA test scores are about 0.75 standard deviations worse than their peers. Males and females are fairly similar demographically with two exceptions: males are slightly more likely to be diagnosed with a disability than females, and females perform about 0.18 standard deviations better on ELA exams than males.[6] As is the case in other value added studies, the value added sample is positively selected on prior test scores, as they score about 0.08 standard deviations above average on their current test scores.[7] Appendix section A gives more information on the data.

Table 4 gives summary statistics for school by grade level peer (jackknife) averages. This table provides evidence of the sorting of students that occurs at the school level. The average low-SES student attends a school where 72% of their peers are also economically disadvantaged and 73% of their peers are minorities.

[3] Defined by the CDE as "a student neither of whose parents have received a high school diploma or a student who is eligible for the free or reduced-price lunch program, also known as the National School Lunch Program (NSLP)."

[4] I exclude observations with a total household income of greater than $250,000 from the figure but not the calculation of the median income, as there is a long low-density right tail in each distribution.

[5] Prior test scores are necessary in order to obtain unbiased estimates when using value added methodologies (Kane and Staiger, 2008; Deming, 2014).

[6] Males perform about 0.02 standard deviations better on math in elementary school, when students take a common test.

[7] Test scores are standardized to have mean zero and standard deviation one at the grade by year level on the entire population of students taking the CST.

In contrast, the average high-SES student attends a school where 35% of their peers are economically disadvantaged and 42% are minorities. These stark differences are similar between minority and non-minority students. The average minority student attends a school with 68% low-SES students and 73% minority students, while the average non-minority student attends a school with 37% low-SES students and 39% minority students. Males and females attend schools with similar peers, as one would expect given that school sorting occurs at the residential level on socioeconomic and ethnic factors but not according to student gender.

Postsecondary data comes from the National Student Clearinghouse (NSC). The NSC data includes enrollment and degree receipt data for the cohorts of students that graduated high school between the spring of 2010 and 2017, inclusive.[8] The NSC provides national enrollment coverage and therefore accounts for both California colleges and universities as well as out-of-state enrollment.

4 School Value Added

4.1 Model

In this section I describe a model of student learning in order to better describe which factors contribute to a school's value added measure. The model follows Naven (2019) with the exception that I allow for within-school heterogeneity in value added. Suppose that the outcome of a student i of subgroup d in grade g of school s in year t is determined according to equation (2), such that a student's endowment ι_i, contemporaneous learning ℓ_{ig}, prior learning ℓ_{ik} depreciated by a factor δ_k, and idiosyncratic school-level shocks θ_{st} all contribute. Assume that students take each grade only once, so that g and t are interchangeable within student.

$$z_{isgdt} = \underbrace{\iota_i}_{\text{Endowment}} + \sum_{k=0}^{g-1} \underbrace{\delta_{kd} \cdot \ell_{ikd}}_{\text{Prior Learning}} + \underbrace{\underbrace{\ell_{igd}}_{\text{Learning}} + \underbrace{\theta_{st}}_{\text{Shocks}} + \underbrace{\varepsilon_{isgdt}}_{\text{Noise}}}_{r_{isgdt}} \tag{2}$$

Assume that the portion of outcome z_{isgdt} that is due to learning is modeled by equation (3) such that teachers τ_{sgdt} and other school factors ψ_{sdt} (such as principals, counselors, curricula, extracurricular activities, and peers) contribute to student learning.

$$\ell_{igd} = \underbrace{\tau_{sgdt}}_{\text{Teachers}} + \underbrace{\psi_{sdt}}_{\text{School Factors}} \tag{3}$$

[8] The cohorts matched were actually spring 2009 to spring 2016 11th grade students, because we do not observe high school graduation data nor the students in 12th grade.

While other studies have investigated the impact of τ_{sgdt} on long-run outcomes, studying school quality allows ψ_{sdt} to also have an impact. This may be particularly important when studying the effects of education on postsecondary enrollment, as high schools are much more likely to have counselors dedicated to the postsecondary decision process and some schools may have better resources on the application process, such as college fairs or mandatory SAT/ACT testing, than others.

Note that by regressing the test score in grade g on the test score in grade $g-1$ it is possible to control for ι_i and $\sum_{k=0}^{g-1} \delta_{kd} \cdot \ell_{ikd}$, the performance a student would achieve even in the absence of school input. This leaves us with the residual term r_{isgdt}, which captures the portion of student performance that is not related to the student's prior achievement.

4.2 Methodology

To estimate ℓ_{ig}, I extend the value added methodology that allows for drift over time described in Chetty, Friedman and Rockoff (2014) to the school level. The value added methodology accounts for the fact that schools receive students of varying backgrounds.[9] Hence, schools that receive only the lowest performing students should not be penalized for the fact that the students they receive will likely have lower outcomes on average. Instead, they should be evaluated on how much they improve the outcomes of those students, regardless of the students' prior achievement. Thus, a school that improves the test scores of the lowest-performing students would be determined to have a higher value added than a school that made no change to the test scores of the highest performing students, even though the latter school's students may perform better on average.

A school's value added is calculated by first removing the portion of each student's test score that is due to non-school factors. To do so, I regress student test scores z_{isgdt} on cubic polynomials in prior test scores $\boldsymbol{z}_{ig}$ and demographic characteristics $\boldsymbol{X}_{it}$ as in equation (4). The cubic polynomials in prior test scores account for mean reversion and the fact that students with low test scores have more room to improve than students with high test scores. I allow the coefficients on prior test scores, $\boldsymbol{\delta}_{gd}$, to vary by subgroup to account for the fact that subgroups may be on different growth trajectories. I also include grade fixed effects γ_g and year fixed effects ψ_t. The demographic characteristics $\boldsymbol{X}_{it}$ contain a linear term for age and fixed effects for sex, ethnicity[10], economic disadvantage, limited English proficiency, and disability status. For each category of subgroup (SES, ethnicity, and sex) I omit the fixed effect for the corresponding category when running equation (4), as including this fixed effect would remove potential differences in the average value added that schools provide to each subgroup. For example, when estimating school value added for

[9] Value added methodologies were first pioneered in estimating school and hospital quality (Willms and Raudenbush, 1989; McClellan and Staiger, 1999, 2000). Meyer (1997) and Everson (2017) provide some background on the methodology.

[10] Asian, Hispanic, black, and other; white is omitted.

low- and high-SES students, I omit the economic disadvantage fixed effect, as including it would force the average school value added for low-SES students to be equal to the average school value added for high-SES students.

$$z_{isgdt} = \boldsymbol{z}_{ig}\boldsymbol{\delta}_{gd} + \boldsymbol{X}_{it}\boldsymbol{\beta}_X + \gamma_g + \psi_t + r_{isgdt} \tag{4}$$

Because there could be idiosyncratic shocks that are uncorrelated with school quality but influence the performance of all students within a school in each year, such as the proverbial dog barking outside of the school on the day of the test, the residual term r_{isgdt} will contain school value added μ_{sdt}, idiosyncratic shocks θ_{st}, and a student-level error term ε_{isgdt} as in equation (5).

$$r_{isgdt} = \mu_{sdt} + \theta_{st} + \varepsilon_{isgdt} \tag{5}$$

Under the assumptions that ε_{isgdt} is a mean zero random error term, students do not sort to schools in each year on unobservable characteristics, and subgroups are not unobservably different from each other, the student-level error terms have expected value zero conditional on school, year, and subgroup, which gives us equation (6).

$$\mathbf{E}[r_{isgdt}|s, d, t] = \mu_{sdt} + \theta_{st} \tag{6}$$

Note that the assumption on the error term here is stronger than in Naven (2019), which only requires that students do not sort to schools in each year on unobservable characteristics. When allowing for within-school heterogeneity, it must be the case that on average there are no unobservable differences between students in different subgroups. If this assumption is violated, then there will appear to be differences in the average value added that schools provide to each subgroup when in fact this is just picking up differences in student performance due to unobserved factors that are correlated with subgroup. This is an especially strong assumption due to the fact that I drop the fixed effect that would account for unobservable differences between subgroups, as this fixed effect is collinear with the existence of differences in average school value added by subgroup. Nevertheless, Kane and Staiger (2008) and Deming (2014) provide evidence that demographic characteristics are essentially irrelevant after conditioning on prior test scores when estimating value added, so the vector of prior test scores will likely be sufficient to satisfy this assumption.

I therefore average the residual r_{isgdt} to the school by subgroup by year level in order to eliminate the student-level error term. However, because value added and idiosyncratic shocks are the same for all students at this level, the average residual will contain both school value added and the school-level idiosyncratic shock as in equation (7).

$$\bar{r}_{sdt} = \mu_{sdt} + \theta_{st} \tag{7}$$

In order to reduce the variation from the idiosyncratic shocks while retaining the variation in school value added, I project the average residual in year t onto the residuals in all other years t' (jackknife projection) as in equation (8).

$$\bar{r}_{sdt} = \bar{\boldsymbol{r}}_{sdt'}\boldsymbol{\beta}_{\bar{r}t'} + \epsilon_{sdt} \tag{8}$$

The value added estimates that I use in this paper are the predicted values from equation (8), $\hat{\mu}_{sdt} = \bar{\boldsymbol{r}}_{sdt'}\hat{\boldsymbol{\beta}}_{\bar{r}t'}$. However, I rescale the estimates so that they have mean zero, thus, schools with positive value added are above average and vice versa.[11] The appendix of Naven (2019) outlines additional methodological details under the assumption of homogenous school quality.

This projection strategy has several advantages. Under the assumptions that school value added is correlated across years ($\text{cov}(\mu_{sdt}\mu_{sdt'}) \neq 0$), the school-level common shocks are uncorrelated across years ($\text{cov}(\theta_{st}\theta_{st'}) = 0$), and the school-level common shocks are not correlated with school value added across years ($\text{cov}(\mu_{sdt}\theta_{st'}) = 0$), the projection will retain variation from school value added and remove variation from the common shocks. In practice, the finite sample size in the number of years may lead to violations of the last two assumptions regarding θ_{st}, which is why the projection will reduce the variation from the idiosyncratic shocks instead of completely eliminating it.[12]

4.3 Test Score Results

Figure 3 shows the distributions of school value added by subgroup. Overall, the standard deviation of school value added is about 0.1, which indicates that a one standard deviation increase in school value added increases the average test score of its students by 0.1 standard deviations. The magnitudes are within the range of those found for the distribution of school value added in Deming (2014) and Angrist et al. (2017) and for the distribution of teacher value added in Kane and Staiger (2008) and Chetty, Friedman and Rockoff (2014). Each graph shows the kernel density of value added separated by subgroup. The bottom of each graph gives the subgroup mean and standard deviation as well as the difference in average school value

[11] This rescaling has no impact on the results to follow.

[12] If the common shocks are truly idiosyncratic, then the last two assumptions regarding θ_{st} are likely to hold as the number of years goes to infinity. Furthermore, to the extent that good or bad events happen continuously at the same schools, these should be considered part of a school's value added, which further reinforces that the common shocks are idiosyncratic. As for the first assumption, schools will experience some faculty and staff turnover, but school value added is likely to be correlated from year to year as the majority of the personnel will remain in the same school from one year to the next.

added between subgroups and a Kolmogorov-Smirnov (Kolmogorov, 1933; Smirnov, 1939) test for equality of distribution. Panel 3a gives the results by SES, panel 3b gives the results by ethnicity, and panel 3c gives the results by sex.

On average, schools provide less value added to low-SES (0.02 student-level standard deviations), minority (0.03 student-level standard deviations), and male (0.03 student-level standard deviations) students. The differences in value added that schools provide to these subgroups amount to the difference between attending an average school and a school that is 0.2 (= 0.022/0.096), 0.3 (= 0.031/0.099), and 0.3 (= 0.032/0.099) standard deviations of school value added below average for SES, ethnicity, and sex respectively. In all cases the Kolmogorov-Smirnov test rejects the null hypothesis that the school value added distributions are equal by subgroup. Figure 3 gives the results pooling all grades, while figure B.1 in appendix section B.1 shows how the heterogeneity in school value added evolves by grade. For all subgroups the difference in average school value added decreases as students age, and in fact the average high school provides more value added to low-SES students than high-SES students and equal value added to minority students as non-minority students.

While figure 3 shows that the average school provides less value added to low-SES, minority, and male students, it doesn't provide any information on how exactly these differences occur. It could be that some schools specialize in certain populations, or it could be that good schools tend to be good for all students but that overall schools provide less value added for certain subgroups of students. Figure 4 provides some insight. Each point in figure 4 represents an observation for a school in a specific year, with the horizontal axis giving the school's value added for high-SES/non-minority/female students and the vertical axis giving the school's value added for low-SES/minority/male students. The figures indicate that while value added is highly correlated between subgroups, with the correlation ranging from 0.846 for ethnicity to 0.934 for sex, there is also substantial within-school heterogeneity in value added. The red line gives the points at which value added is equal for each subgroup, and while the value added estimates trend parallel to this line, there is also substantial deviation. Interestingly, schools tend to provide more homogenous value added added by sex than they do by SES or ethnicity, as the scatter plot of school value added by sex has the least amount of deviation from the 45 degree line and highest correlation between the value added for each subgroup. Nevertheless, figure 4 provides evidence that while some schools in quadrants two and four are particularly effective with a certain type of student, the vast majority of schools are either good or bad overall but simply tend to provide less quality to low-SES students, minorities, and men.

Thus within-school heterogeneity is one factor that contributes to the existing achievement gaps by SES, ethnicity, and sex. If the average school provided equal value added on test scores between all subgroups, this would eliminate 3% (= 0.022/[0.495 - -.246]) of the SES achievement gap, 4% (= 0.031/[0.517 - -.218]) of the ethnicity achievement gap, and 18% (= 0.032/[0.177 - 0]) of the sex achievement gap.[13] These achievement gaps have important consequences. Students must compete on achievement in a variety of educational settings. Class rank, college admission, and scholarships are all competitions in which high-achieving students can reap large benefits, which could in turn prevent low-performing students from enrolling in college. More importantly, achievement gaps may contribute to the existence of poverty traps. Because cognitive gaps are highly predictive of wage gaps (Neal and Johnson, 1996; Bollinger, 2003; Carneiro, Heckman and Masterov, 2005), achievement gaps may perpetuate the cycle of poverty by ensuring that low-performing students obtain low-paying jobs and, in turn, have children who will suffer similar consequences from growing up in a low-income family. Therefore schools that close the achievement gap for vulnerable students could be a valuable mechanism for reducing poverty.

4.4 Postsecondary Results

Much more important than test scores, however, is the effect of school value added on postsecondary enrollment, because attending college has proven to be a worthwhile investment for both the average and marginal student (Oreopoulos and Petronijevic, 2013). Hoekstra (2009) finds that attending a flagship university increases the earnings of white men by 20%, while Zimmerman (2014) shows that admission to a 4-year university for the marginal student gives a wage premium of 22% and bachelor's degree receipt for the marginal admission increases wages by 90%. I define postsecondary enrollment as enrolling in any institution in the NSC data within one year of high-school graduation.[14] I code two-year and four-year enrollment as mutually exclusive, so if students enroll in both a two-year and four-year institution within a year of graduating high school (such as if they take a summer course at a community college) then I code them as only enrolling in a four-year institution.

In order to estimate a school's value added on postsecondary enrollment directly, I reestimate equation (4) with an indicator for postsecondary enrollment as the dependent variable instead of a student's test score.[15] It should be noted that the assumptions to obtain unbiased estimates of school value added on postsecondary enrollment are stronger than those for school value added on test scores. Value added on test scores relies upon the assumption that prior test scores and demographic characteristics are sufficient to predict how a student would perform on the current year's test, such that any differences in test scores after controlling for these variables are attributable to schools. Prior research shows that this is a valid assumption (Kane and

[13] These calculations come from dividing the difference in the mean school value added by subgroup from Figure 3 by the difference in the mean current test score by subgroup from Table 3.

[14] I also include any student who enrolls in a CSU or a CCC within one year of high-school graduation in order to account for any missing data in the NSC data. The sample consists solely of students who could potentially be matched to the NSC data, as students who did not enroll in a CSU or CCC and could not be potentially matched to the NSC data may still have enrolled in a postsecondary institution, such as a UC, but I would not observe this.

[15] Because each student's enrollment outcome is invariant across grades, I only use observations from 11th grade.

Staiger, 2008; Deming, 2014). Estimating value added on postsecondary enrollment, however, relies upon the assumption that prior test scores and demographic characteristics are sufficient to predict the likelihood that a student will attend a postsecondary institution. This assumption may not hold, especially given that Abdulkadiroğlu et al. (2017) finds that the bias of value added estimates on postsecondary enrollment is larger than the bias of value added estimates on test scores at the high school level. Thus the results for school value added on postsecondary enrollment should be interpreted keeping these caveats in mind.

Table 5 gives summary statistics for the school postsecondary enrollment value added sample. As with test scores and demographic characteristics, there are substantial differences in postsecondary enrollment by SES, ethnicity, and sex. Low-SES students are 17 percentage points less likely to attend a postsecondary institution than high-SES students. The entirety of this difference comes from differences in four-year university enrollment, as low-SES students are actually slightly more likely to enroll in a community college than their high-SES peers. Low-SES students are also more likely to attend in-state and public institutions, which is intuitive given that these options tend to be cheaper than out-of-state and private institutions. Similar differences exist between minority and non-minority students. Differences between males and females are less stark, although males are still about eight percentage points less likely to attend a four-year university than females.

Figure 5 shows the distributions of the estimated school value added on postsecondary enrollment. As with the results using value added on test scores, schools provide less value added on postsecondary enrollment to their low-SES, minority, and male students. These gaps range from 2.5 to 4.3 percentage points. The differences in school postsecondary enrollment value added by subgroup are even larger in magnitude than the differences in school test score value added. Schools provided 0.23 (= 0.022/0.096) standard deviations less value added on test scores to low-SES students, but they provide 0.48 (= 0.041/0.086) standard deviations less value added on postsecondary enrollment. For minorities, schools provide 0.31 (= 0.031/0.099) and 0.29 (= 0.025/0.087) standard deviations less value added on test scores and postsecondary enrollment respectively, and for males 0.32 (= 0.032/0.099) and 0.52 (= 0.043/0.082) standard deviations. If the average school provided equal value added on postsecondary enrollment to low- and high-SES students, it would close 24% (= 0.041/[0.755 - 0.581]) of the college-going gap between low- and high-SES students. Similarly, if the average school provided equal value added to minorities it would close 16% (= 0.025/[0.749 - 0.597]) of the enrollment gap by ethnicity and if the average school provided equal value added to males it would close 52% (= 0.043/[0.703 - 0.62]) of the enrollment gap by sex.

These figures provide further evidence that test scores should not be the only measure for determining whether schools perpetuate inequality (Jennings et al., 2015). Furthermore, because schools have a significant impact on whether students enroll in a postsecondary institution, the public school system could be a valuable asset for increasing the postsecondary outcomes of low-SES and minority students. Given that students' income ranks are highly inherited by their parents' income rank (Chetty et al., 2014*a*) and that intergenerational mobility in the U.S. has stagnated in recent years (Chetty et al., 2014*b*) but that a college education of any level flattens the relationship between student income rank and parent income rank (Chetty et al., 2017), improving primary and secondary school quality for low-SES and minority students may be an effective way to reduce poverty and increase intergenerational mobility.

Figure 6 is analogous to figure 4 except that I plot school by subgroup value added on postsecondary enrollment instead of on test scores. Here we see a little more within-school heterogeneity as evidenced by the wider cloud of points in figure 6 relative to figure 4. While within-school value added on postsecondary enrollment is still highly positively correlated, with correlations ranging from 0.775 to 0.858, the correlations are lower than they were for value added on test scores. In particular, there are many schools in quadrant four that provide above average value added to high-SES/non-minority/female students but below average value added to low-SES/minority/male students. In contrast there are essentially no schools in quadrant two that provide relatively more value added on postsecondary enrollment to their low-SES, minority, and male students. Nevertheless, as with test scores, it's largely the case that schools that are good at increasing the likelihood of postsecondary enrollment will do so for all students attending the school.

5 Value Added Characteristics

5.1 School Characteristics

Finally, I explore what school characteristics are correlated with school value added. While these regressions are not causal, they provide a description of what high value added schools have in common and whether there are differences in the characteristics of high value added schools by subgroup. This analysis may therefore provide clues of some effective characteristics that could be explored in a causal framework in future studies.

I run regressions of school by subgroup value added on school-level inputs as in equation (9). I cluster the standard errors at the school level. The school characteristics included in $\boldsymbol{X}_{st}$ are the number of full-time equivalent (FTE) teachers per student, FTE pupil services staff[16] per student, English-learner staff per student, proportion teachers with three years or less experience, proportion teachers with full credentials, proportion male teachers, and proportion minority teachers. I drop the top and bottom 2.5% of each independent variable in order to account for outliers and potential errors in the data that schools report.

[16]This includes counselors, psychologists, librarian/library/media teachers, social workers, nurses, and speech/language/hearing specialists.

Note that I allow $\boldsymbol{\beta}_d$ to vary by subgroup, so that each component of $\boldsymbol{X}_{st}$ can contribute differentially to the value added provided to each subgroup. This is implemented by additionally interacting each element in $\boldsymbol{X}_{st}$ with a subgroup fixed effect.

$$\hat{\mu}_{sdt} = \boldsymbol{X}_{st}\boldsymbol{\beta}_d + \phi_d + \varepsilon_{st} \tag{9}$$

Table 6 shows the correlations between school by subgroup value added and school characteristics. The left three columns give value added on test scores, while the right three columns give value added on postsecondary enrollment. Here we see that more teachers per student is correlated with higher value added on test scores, although not postsecondary enrollment. Pupil services per student, having a high proportion of inexperienced teachers, and having more male teachers are consistently negatively correlated with school value added on test scores. Interestingly, English learner staff per student are positively correlated with test score value added but negatively correlated with postsecondary enrollment value added. Fully credentialed teachers are associated with higher value added on both test scores and postsecondary enrollment.

There are some interesting differential correlations by subgroup. The positive correlation between teachers per student and school value added on test scores is negated for the value added that schools provide to low-SES students. Having more English learner staff is even more positively correlated with the value added that schools provide on test scores to low-SES and minority students, which makes sense given that most of California's low-SES and minority students are Hispanic. This is suggestive that staff dedicated to improving the language skills of English as a second language (ESL) students could have positive effects on this population of students. A higher proportion of male teachers has a less negative effect on low-SES, minority, and male students than it does on high-SES, non-minority, and female students. In addition, a higher proportion of minority teachers is positively correlated with the value added that schools provide to low-SES students. Overall, there is suggestive evidence that the inputs necessary to provide high value added to low-SES, minority, and male students are potentially different than those required to provide value added to high-SES, non-minority, and female students. These differential correlations could partially explain the ambiguous results found in Naven (2019) when correlating school characteristics with homogenous school value added.

5.2 Spatial Correlations

Given that there is substantial sorting of students to schools based on the cost of housing in the neighborhoods attending schools and that SES is highly correlated with ethnicity, another important way in which to look at the characteristics of high value added schools is where high value added schools are located. Figure 7 shows the average school test score value added within each zip code in California. As with Naven (2019), I find that high value added schools tend to be clustered in the large metropolitan areas of Los Angeles and the Bay Area.

However, I also find some heterogeneity by SES and ethnicity. Figures 7a and 7b, which show where high value added schools are located allowing for within-school heterogeneity by SES, show that the value added that schools provide to low-SES students is higher in the Central Valley than for high-SES students. A similar story emerges for minorities compared to non-minorities in figures 7c and 7d. Unsurprisingly, as seen in figures 7e and 7f, schools that provide high value added to men and women tend to be in the same area, as students don't sort to schools based on gender.

The spatial correlation results are somewhat intuitive. Schools have resource constraints and must maximize some objective function subject to their budget constraint. For example, there is evidence that when faced with school accountability measures based on the number of students who pass a proficiency standard schools respond by focusing their efforts on "bubble students" — those students who would meet proficiency standards given sufficient investment but who would otherwise score below proficient (Booher-Jennings, 2005; Krieg, 2008; Neal and Schanzenbach, 2010; Krieg, 2011). Another objective function that schools may have is to maximize the average gain of their students regardless of where their final test scores fall compared to the proficiency cutoff. Given that California's accountability system from 2002-2003 to 2012-2013 was based on the Academic Performance Index (API), which is calculated via the number of students in each of five performance bands and rewards schools for having more students in the higher bands, schools may have had an incentive to adopt teaching strategies that were particularly effective with the typical (i.e. majority) student so that they could increase the band in which the majority of their students fell, even if the majority of their students still fell below a proficiency cutoff. The Central Valley is poorer and has a higher proportion of minorities than Los Angeles or the Bay Area, thus it makes sense that schools would work better with these populations of students when they are the typical student that a school enrolls.

Figure 8 shows the same information for school value added on postsecondary enrollment. Here a similar pattern emerges as to that with value added on test scores. High postsecondary enrollment value added schools tend to be clustered in the population centers of the state, but schools in the lower-income and higher-minority Central Valley tend to be better for low-SES and minority students.

6 Conclusion

Achievement gaps in education by SES, ethnicity, and sex continue to persist despite efforts to close them. This paper explores to what extent within-school heterogeneity in school quality by SES, ethnicity, and sex contribute to the existence of these achievement gaps. I investigate this issue by performing three separate analyses in which I allow each school to have distinct measures of school quality for low-/high-SES, minority/non-minority, and male/female students. I therefore do not impose the assumption that schools have an equal impact on all students enrolled in the school but instead allow for the possibility that schools provide more quality to certain subgroups of students.

Using data on the universe of public school students in California, I estimate school by subgroup quality on both standardized test scores and postsecondary enrollment by applying the value added with drift methodology, as in Chetty, Friedman and Rockoff (2014), to schools. The value added methodology accounts for the fact that students do not randomly sort to schools and subgroups of students have varying levels of average academic performance. Allowing for within-school heterogeneity by subgroup relaxes the assumption schools must have the same impact on all types of students enrolled in the school. Simulations suggest that assuming homogeneity in school value added incorrectly ranks schools in the presence of within-school heterogeneity in value added. Allowing for this within-school heterogeneity when estimating school value added restores the correct ranking of schools. To my knowledge this paper is the first to allow for within-school heterogeneity in how much schools increase the postsecondary enrollment of their students.

Results show that there is indeed within-school heterogeneity in value added by SES, ethnicity, and sex on both test scores and postsecondary enrollment. The difference in the average test score value added provided to low- and high-SES students is equivalent to attending a school that is 0.2 standard deviations below average in the value added that it provides to students. The corresponding differences for minority and male students are 0.3 standard deviations. If schools provided equal test score value added on average to their low-SES, minority, and male students as they did their high-SES, non-minority, and female students, then it would close 3%, 4%, and 18%, of the test score achievement gap respectively. Within-school heterogeneity in the value added that schools provide on postsecondary enrollment is even larger. If schools provided equal postsecondary enrollment value added on average to their low-SES, minority, and male students as they did their high-SES, non-minority, and female students, then it would close 24%, 16%, and 52%, of the postsecondary enrollment gap respectively.

I then correlate school value added measures with school characteristics and location. I find that more English learner staff are positively correlated with value added for low-SES and minority students, which could be due to the fact that the vast majority of low-SES and minority students are Hispanic. Having more minority teachers is also positively correlated with the value added that schools provide to low-SES students. Schools that are beneficial to low-SES and minority students are also geographically located in different areas than schools that provide high value added to high-SES and non-minority students. While schools in populous areas tend to be good for all types of students, schools in the more sparsely-populated

Central Valley tend to only be good with low-SES and minority students. Low-SES and minority students students disproportionately live in the Central Valley, so schools may be more effective with these types of students if they are constantly in contact with them.

While within-school heterogeneity in quality is not sufficient to explain the entirety of the achievement gap by SES, ethnicity, or gender, it is a previously unexplored mechanism that contributes to the persistence of these gaps. Within-school differences in how much schools increase the likelihood that their students attend college are particularly large, especially relative to differences in how much schools increase test scores. Given that students of varying backgrounds will have different learning styles and differences in home life, schools should work to assure that disadvantaged students are not left behind by catering to more advantaged students who are already succeeding.

References

Abdulkadiroğlu, Atila, Parag A Pathak, Jonathan Schellenberg, and Christopher R Walters. 2017. "Do Parents Value School Effectiveness?" National Bureau of Economic Research.

Angrist, Joshua D, Peter D Hull, Parag A Pathak, and Christopher R Walters. 2017. "Leveraging lotteries for school value-added: Testing and estimation." *The Quarterly Journal of Economics*, 132(2): 871–919.

Bailey, Martha J, and Susan M Dynarski. 2011. "Gains and gaps: Changing inequality in US college entry and completion." National Bureau of Economic Research.

Bau, Natalie. 2015. "School competition and product differentiation." mimeo, Harvard University.

Belley, Philippe, and Lance Lochner. 2007. "The changing role of family income and ability in determining educational achievement." *Journal of Human capital*, 1(1): 37–89.

Bitler, Marianne P, Sean P Corcoran, Thurston Domina, and Emily K Penner. 2014. "Teacher Effects on Student Achievement and Height: A Cautionary Tale." *Society for Research on Educational Effectiveness.*

Bollinger, Christopher R. 2003. "Measurement error in human capital and the black-white wage gap." *Review of Economics and Statistics*, 85(3): 578–585.

Booher-Jennings, Jennifer. 2005. "Below the bubble:"Educational triage" and the Texas accountability system." *American educational research journal*, 42(2): 231–268.

Carneiro, Pedro, James J Heckman, and Dimitriy V Masterov. 2005. “Labor market discrimination and racial differences in premarket factors.” *The Journal of Law and Economics*, 48(1): 1–39.

Carrell, Scott E, Marianne E Page, and James E West. 2010. “Sex and science: How professor gender perpetuates the gender gap.” *The Quarterly Journal of Economics*, 125(3): 1101–1144.

Chatterji, Madhabi. 2006. “Reading achievement gaps, correlates, and moderators of early reading achievement: Evidence from the Early Childhood Longitudinal Study (ECLS) kindergarten to first grade sample.” *Journal of educational psychology*, 98(3): 489.

Chetty, Raj, John N Friedman, and Jonah E Rockoff. 2014. “Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates.” *The American Economic Review*, 104(9): 2593–2632.

Chetty, Raj, John N Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan. 2017. “Mobility report cards: The role of colleges in intergenerational mobility.” National Bureau of Economic Research.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014*a*. “Where is the land of opportunity? The geography of intergenerational mobility in the United States.” *The Quarterly Journal of Economics*, 129(4): 1553–1623.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, Emmanuel Saez, and Nicholas Turner. 2014*b*. “Is the United States still a land of opportunity? Recent trends in intergenerational mobility.” *American Economic Review*, 104(5): 141–47.

Cimpian, Joseph R, Sarah T Lubienski, Jennifer D Timmer, Martha B Makowski, and Emily K Miller. 2016. “Have gender gaps in math closed? Achievement, teacher perceptions, and learning behaviors across two ECLS-K cohorts.” *AERA Open*, 2(4): 2332858416673617.

Dee, Thomas S. 2004. “Teachers, race, and student achievement in a randomized experiment.” *Review of Economics and Statistics*, 86(1): 195–210.

Dee, Thomas S. 2005. “A teacher like me: Does race, ethnicity, or gender matter?” *American Economic Review*, 95(2): 158–165.

Dee, Thomas S. 2007. “Teachers and the gender gaps in student achievement.” *Journal of Human Resources*, 42(3): 528–554.

Deming, David J. 2014. “Using school choice lotteries to test measures of school effectiveness.” *The American Economic Review*, 104(5): 406–411.

Egalite, Anna J, Brian Kisida, and Marcus A Winters. 2015. "Representation in the classroom: The effect of own-race teachers on student achievement." *Economics of Education Review*, 45: 44–52.

Everson, Kimberlee C. 2017. "Value-Added Modeling and Educational Accountability: Are We Answering the Real Questions?" *Review of Educational Research*, 87(1): 35–70.

Fryer Jr, Roland G, and Steven D Levitt. 2010. "An empirical analysis of the gender gap in mathematics." *American Economic Journal: Applied Economics*, 2(2): 210–40.

Gershenson, Seth, Cassandra Hart, Constance Lindsay, and Nicholas W Papageorge. 2017. "The long-run impacts of same-race teachers."

Hoekstra, Mark. 2009. "The effect of attending the flagship state university on earnings: A discontinuity-based approach." *The Review of Economics and Statistics*, 91(4): 717–724.

Holt, Stephen B, and Seth Gershenson. 2017. "The Impact of Demographic Representation on Absences and Suspensions." *Policy Studies Journal*.

Husain, Muna, and Daniel L Millimet. 2009. "The mythical 'boy crisis'?" *Economics of Education Review*, 28(1): 38–48.

Jennings, Jennifer L, David Deming, Christopher Jencks, Maya Lopuch, and Beth E Schueler. 2015. "Do differences in school quality matter more than we thought? New evidence on educational opportunity in the twenty-first century." *Sociology of Education*, 88(1): 56–82.

Kahn, Shulamit, and Donna Ginther. 2017. "Women and STEM." National Bureau of Economic Research.

Kane, Thomas J, and Douglas O Staiger. 2008. "Estimating teacher impacts on student achievement: An experimental evaluation." National Bureau of Economic Research.

Karen, David. 2002. "Changes in access to higher education in the United States: 1980-1992." *Sociology of Education*, 191–210.

Kolmogorov, Andrey. 1933. "Sulla determinazione empirica di una lgge di distribuzione." *Inst. Ital. Attuari, Giorn.*, 4: 83–91.

Krieg, John M. 2008. "Are students left behind? The distributional effects of the No Child Left Behind Act." *Education Finance and Policy*, 3(2): 250–281.

Krieg, John M. 2011. "Which students are left behind? The racial impacts of the No Child Left Behind Act." *Economics of Education Review*, 30(4): 654–664.

Lee, Joohi, Sungseek Moon, and Rebecca L Hegar. 2011. "Mathematics skills in early childhood: Exploring gender and ethnic patterns." *Child Indicators Research*, 4(3): 353–368.

Lindsay, Constance A, and Cassandra MD Hart. 2017. "Exposure to same-race teachers and student disciplinary outcomes for Black students in North Carolina." *Educational Evaluation and Policy Analysis*, 39(3): 485–510.

McClellan, Mark, and Douglas Staiger. 1999. "The quality of health care providers." National bureau of economic research.

McClellan, Mark, and Douglas Staiger. 2000. "Comparing the quality of health care providers." In *Frontiers in Health Policy Research, Volume 3*. 113–136. MIT Press.

Meyer, Robert H. 1997. "Value-added indicators of school performance: A primer." *Economics of Education Review*, 16(3): 283–301.

Naven, Matthew. 2019. "Human Capital Formation During Childhood and Adolescence: Evidence from School Quality and Postsecondary Success in California." Washington and Lee University.

Neal, Derek A, and William R Johnson. 1996. "The role of premarket factors in black-white wage differences." *Journal of political Economy*, 104(5): 869–895.

Neal, Derek, and Diane Whitmore Schanzenbach. 2010. "Left behind by design: Proficiency counts and test-based accountability." *The Review of Economics and Statistics*, 92(2): 263–283.

Oreopoulos, Philip, and Uros Petronijevic. 2013. "Making college worth it: A review of research on the returns to higher education." National Bureau of Economic Research.

Penner, Andrew M, and Marcel Paret. 2008. "Gender differences in mathematics achievement: Exploring the early grades and the extremes." *Social Science Research*, 37(1): 239–253.

Reardon, Sean F. 2011. "The widening academic achievement gap between the rich and the poor: New evidence and possible explanations." *Whither opportunity*, 91–116.

Reardon, Sean F. 2016. "School district socioeconomic status, race, and academic achievement." *Stanford Center for Educational Policy Analysis. Retrieved from.*

Reardon, Sean F, and Claudia Galindo. 2009. "The Hispanic-White achievement gap in math and reading in the elementary grades." *American Educational Research Journal*, 46(3): 853–891.

Reardon, Sean F, and Joseph P Robinson. 2008. "Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps." *Handbook of research in education finance and policy*, 497–516.

Reardon, Sean F, and Ximena A Portilla. 2016. "Recent trends in income, racial, and ethnic school readiness gaps at kindergarten entry." *AERA Open*, 2(3): 2332858416657343.

Reardon, Sean F, Rachel Baker, and Daniel Klasik. 2012. "Race, income, and enrollment patterns in highly selective colleges, 1982-2004." *Center for Education Policy Analysis, Stanford University. Retrieved from http://cepa. stanford. edu/content/race-income-and-enrollmentpatterns-highly-selective-colleges-1982-2004.*

Robinson, Joseph Paul, and Sarah Theule Lubienski. 2011. "The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings." *American Educational Research Journal*, 48(2): 268–302.

Rothstein, Jesse. 2017. "Measuring the Impacts of Teachers: Comment." *American Economic Review*, 107(6): 1656–84.

Smirnov, Nikolai V. 1939. "Estimate of deviation between empirical distribution functions in two independent samples." *Bulletin Moscow University*, 2(2): 3–16.

Sohn, Kitae. 2012. "A new insight into the gender gap in math." *Bulletin of Economic Research*, 64(1): 135–155.

Willms, J Douglas, and Stephen W Raudenbush. 1989. "A longitudinal hierarchical linear model for estimating school effects and their stability." *Journal of educational measurement*, 26(3): 209–232.

Zimmerman, Seth D. 2014. "The returns to college admission for academically marginal students." *Journal of Labor Economics*, 32(4): 711–754.

A Data

Table A.1 gives the number of observations in the CST data conditional on a set of restrictions implemented in order to form the value added sample. The rows are additive, such that the first row contains all observations, the second row imposes one restriction, the third row imposes two restrictions, etc. The first row denotes the total number of observations in the CST dataset. The second row keeps students who have information on test scores, as opposed to just demographic characteristics. The third row keeps only the first time that a student attempted a grade, and thus drops observations in which a student is repeating a grade. I impose this restriction because students repeating a grade are tested on material for which they have already been tested at least once. The fourth row keeps only students at "conventional" schools. This includes schools in the following categories defined by the CDE: Preschool, Elementary School (Public), Elementary School in 1 School District (Public), Intermediate/Middle Schools (Public), Junior High Schools (Public), K–12 Schools (Public), High Schools (Public), and High Schools in 1 School District (Public).[17] The fifth row drops any schools that enroll 10 students or fewer in a given year. The sixth row drops students who are missing a test score in the specific subject for which value added is calculated. The seventh row drops students who are missing any of the demographic controls. The eighth and ninth rows drop students who are missing test scores from one grade and two grades prior, respectively. This restriction is the cause of the vast majority of observations which are excluded from the value added estimates. The tenth row drops observations for which peer averages of the control variables could not be calculated. The eleventh row drops students if fewer than seven observations can be used to estimate value added for their school by subgroup by year cell, which insures that all value added estimates are based on at least seven observations.

B School Value Added

B.1 Test Score Results

Figure B.1 shows how the heterogeneity in school value added evolves by grade. In general, the difference in value added that schools provide between low-SES/minority/male students and high-SES/non-minority/female students is highest in elementary school and shrinks as students progress towards high school. In fact, in grade 11 schools provide more value added on average to low-SES students than they do to high-SES students and equal value added on average between minority and non-minority students.

B.1.1 Validity Tests

There are three potential concerns regarding the validity of the value added estimates. The first is that the estimates may be picking up noise due to sampling error and small sample variability. This would be the case if test scores are sufficiently noisy that student-level residual test scores, ε_{isgdt}, do not average out to

[17]This drops students in the following categories: Special Education Schools (Public), County Community, Youth Authority Facilities (CEA), Opportunity Schools, Juvenile Court Schools, Other County or District Programs, State Special Schools, Alternative Schools of Choice, Continuation High Schools, District Community Day Schools, Adult Education Centers, and Regional Occupational Center/Program (ROC/P).

zero in each school by subgroup cell, even when schools have no effect on student performance (Bitler et al., 2014). If this were the case, we would attribute value added to schools when we were in fact just observing sampling error.

In order to measure how much of the estimated variation in school value added is due to noise, I calculate school value added estimates after randomly assigning students to schools. I call these value added estimates permuted value added, as I permute the school assignment vector within a grade by year cell. Figure B.2 shows the distributions of permuted value added, and I plot the distributions on the same axes as figure 3 so that the variability can be directly compared. Unlike in Naven (2019), there remains a distribution of estimated school value added even after randomly assigning students to schools. While Naven (2019) found a variance of essentially zero after randomly assigning students to schools, I find standard deviations of school value added as large as 0.015. Assuming that this variation is purely due to measurement error, then 15% (= 0.014/0.096), 15% (=0.015/0.099), and 12% (= 0.012/0.099) of the standard deviation in estimated school value added is due to noise for SES, ethnicity, and sex respectively.

It is intuitive that I will find noisier estimates when estimating school by subgroup value added as opposed to school value added as in Naven (2019) because the cell sizes here are smaller. Because each school is effectively cut in half, so that there are twice as many schools but with fewer students, the law of large numbers is less likely to kick in regarding $\mathbf{E}[\varepsilon_{isgdt}|s,d,t] = 0$. In simpler terms, there is more likely to be noise in an average calculated using a subsample than in an average calculated using the entire sample. Nevertheless, the vast majority of the variation in school value added is still due to signal as opposed to noise, so these results alleviate concerns that the value added estimates are purely an artifact of noisy test score measures or small sample variability.

Figure B.2 also shows differences in the average permuted value added between subgroups. This, however, is expected given that we observe differences in the actual value added that schools provide to subgroups. For example, let the subscript p denote a student's permuted school (i.e. the school to which they are randomly assigned). If the distribution of school value added to high-type students has mean μ and the distribution of school value added for low-type students has mean $\mu - x$, then we have

$$\begin{aligned}
\mathbf{E}[r_{ispgHt}|p,H,t] &= \mathbf{E}[\mu_{sHt}|p,H,t] + \mathbf{E}[\theta_{st}|p,H,t] + \mathbf{E}[\varepsilon_{isgHt}|p,H,t] \\
&= \mathbf{E}[\mu_{sHt}|H] + \mathbf{E}[\theta_{st}] + \mathbf{E}[\varepsilon_{isgHt}] \\
&= \mu
\end{aligned} \tag{10}$$

$$
\begin{aligned}
\mathbf{E}[r_{ispgLt}|p, L, t] &= \mathbf{E}[\mu_{sLt}|p, L, t] + \mathbf{E}[\theta_{st}|p, L, t] + \mathbf{E}[\varepsilon_{isgLt}|p, L, t] \\
&= \mathbf{E}[\mu_{sLt}|L] + \mathbf{E}[\theta_{st}] + \mathbf{E}[\varepsilon_{isgLt}] \qquad (11) \\
&= \mu - x
\end{aligned}
$$

where the second line in equations 10 and 11 follow because μ_{sdt}, θ_{st}, and ε_{isgdt} are independent of p by definition of random assignment and independent of t by assumption of stationarity, and θ_{st}, and ε_{isgdt} are independent of d by assumption. Therefore even after randomly assigning students to schools we would expect to see a difference in the average permuted value added provided to each subgroup, although in theory we would expect no variation in permuted value added.

Another concern is that the value added estimates are the incorrect magnitude. Specifically, the issue is whether a one unit increase in school value added actually is associated with a one standard deviation increase in student test scores. In order to test this, I run a bivariate regression of residualized test scores r_{isgdt} on the school value added estimates μ_{sdt}, where the residualized test scores are calculated using equation (4). This follows the procedure used in Chetty, Friedman and Rockoff (2014) and Rothstein (2017) and calculates by how much a school's estimated value added actually increases the test scores of its students. We expect the coefficient to equal one, which would indicate that a one unit increase in school value added increases student test scores by one standard deviation on average.

The first row of Table B.1 provides this estimate along with its 95% confidence interval. The coefficient estimates range from 1.000 to 1.003, which are statistically and economically indistinguishable from one. Chetty, Friedman and Rockoff (2014) obtain a coefficient estimate of 0.998. This gives evidence that the school value added estimates have the correctly-sized effect on student test scores. Furthermore, figure B.3 graphs the relationship between r_{isgdt} and μ_{sdt} in 20 equally sized bins by subgroup. Results show that the value added estimates and test score residuals have an almost perfectly linear relationship throughout the value added distribution and that there are essentially no differences by subgroup.

The final concern, and potentially most problematic, involves the potential sorting of students to schools based on unobserved ability. If students with high unobserved ability sort to specific schools, such that $cov(\varepsilon_{isgdt}, \mu_{sdt}) \neq 0$, then these schools' estimated value added will be higher than their true value added. However, this is only an issue if the sorting occurs on *unobserved* ability. Hence, there is no issue if students sort to schools on observed ability, because this will be controlled for with the inclusion of prior test scores and demographic controls. For example, if students with high test scores tend to attend the same schools, as occurs in practice, then we can still obtain unbiased estimates of school value added as long as prior test scores are included in the control vector so that $\mathbf{E}[\varepsilon_{isgdt}|s, d, t] = 0$. In fact, research comparing value added estimates to estimates obtained using random assignment to schools (Deming, 2014; Angrist et al., 2017) or teachers (Kane and Staiger, 2008) shows that once you control for prior test scores, even the inclusion of demographic characteristics in the control vector is essentially irrelevant because prior test scores are a sufficient statistic for student ability.

The primary threat to this assumption would be if students or parents *changed* their level of input into academic preparation between the student's prior grade and current grade and students sorted to schools based on this change in behavior. For example, if all students of parents who received an increase in income between grades, where the extra income was used to purchase academic assistance, attended the same school, then the estimated value added of this school would be positively biased. This is due to the fact that the prior test scores and demographic controls of those students would not control for this change in academic assistance, so $\mathbf{E}[\varepsilon_{isgdt}|s, d, t] > 0$. If students whose parents *consistently* have high income sort to the same schools there would not be the same issue, because the students' prior test scores would also reflect their high SES.

The issue in determining to what degree students sort to schools on unobserved ability is that, by definition, we have no measures of unobserved ability. However we can approximate unobserved ability using variables in our data that likely would be correlated with ability but that were not included as a control variable in equation (4). Given the available data, the best possible measure of unobserved student ability is an additional prior test score. Under the assumption that this omitted variable is the only component of ε_{isgdt} that is correlated with student test scores, we can then obtain an estimate of $\frac{cov(\varepsilon_{isgdt}, \hat{\mu}_{sdt})}{var(\hat{\mu}_{sdt})}$. Chetty, Friedman and Rockoff (2014) call this value forecast bias, which gives an estimate of what proportion of the variation in school value added is due to sorting on unobserved ability.[18]

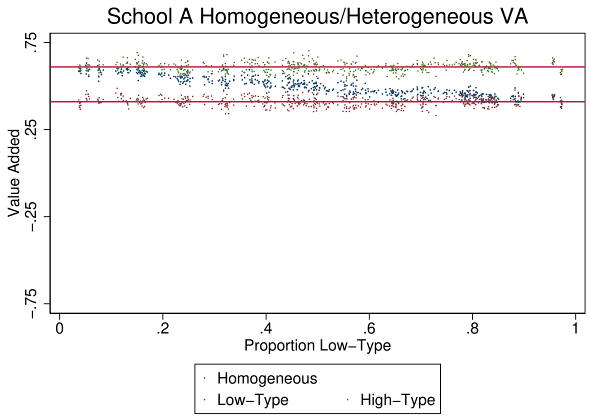
The second row of Table B.1 provides the estimate of forecast bias along with its 95% confidence interval. Here we expect an estimate of zero, which would give evidence that there is no sorting of students to schools on unobservable characteristics. The estimates suggest that between 0.2% (ethnicity) and 0.9% (sex) of the variance in school value added is due to sorting on unobserved ability, thus selection on unobservables does not appear to be a large issue. Chetty, Friedman and Rockoff (2014) estimate forecast bias of 2.2%. Given that the forecast bias estimates are all negative, this would suggest that students who are unobservably worse tend to attend schools with higher value added. This would result in value added estimates that are closer to zero, thus the value added estimates are slightly conservative if anything. Figure B.4 shows that this relationship holds throughout the distribution of school value added and that there are no differences

[18]Similar to Chetty, Friedman and Rockoff (2014), I estimate this using the following steps. First I obtain the portion of contemporaneous test scores that projects onto three-grade prior test scores by adding three-grade prior test scores to equation (4). The projection is equal to the predicted value using only the test score from three grades prior. I then regress this projection on school value added.

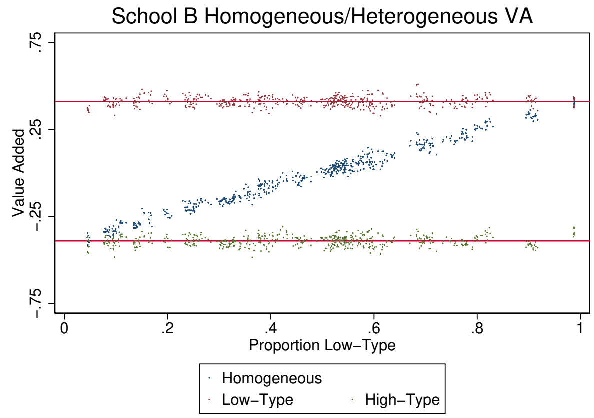
in the relationship by subgroup.

B.2 Postsecondary Results

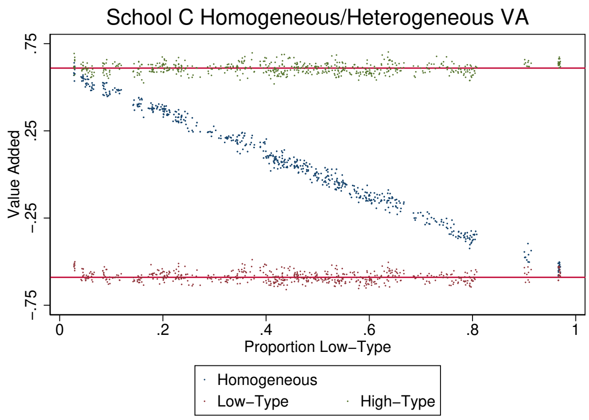
Figure B.5 shows how the heterogeneity in school postsecondary enrollment value added evolves by grade. As with value added on test scores, the difference in value added on postsecondary enrollment that schools provide between low-SES/minority/male students and high-SES/non-minority/female students is highest in elementary school and shrinks as students progress towards high school (with the exception of middle school for minorities). However, the differences do not shrink to zero by the time students reach high school as they do for value added on test scores.



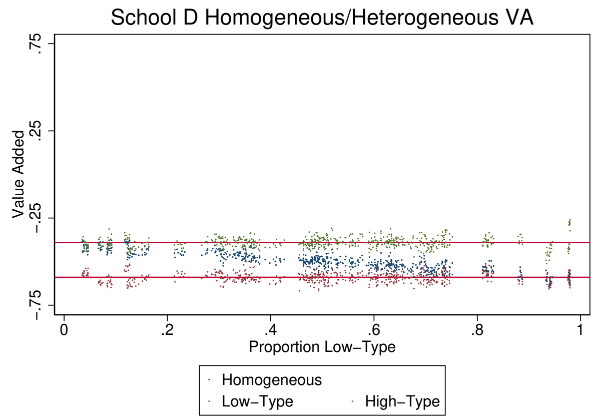
(a) $\mu_{AL} = 0.41$, $\mu_{AH} = 0.61$



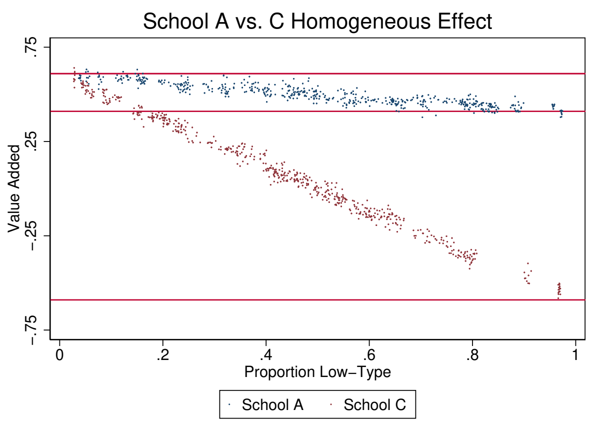
(b) $\mu_{BL} = 0.41$, $\mu_{BH} = -0.39$



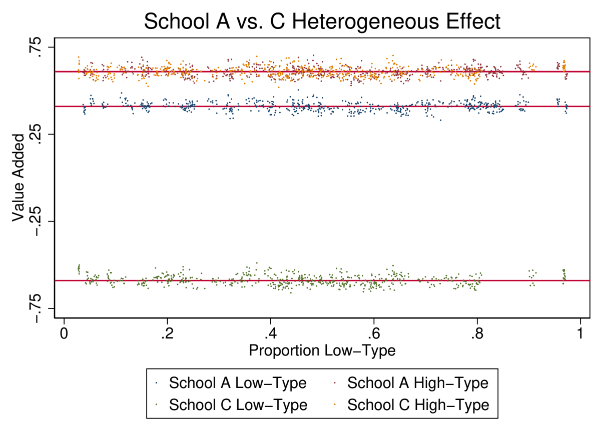
(c) $\mu_{CL} = -0.59$, $\mu_{CH} = 0.61$



(d) $\mu_{DL} = -0.59$, $\mu_{DH} = -0.39$



(e) $\mu_{AL} = 0.41$, $\mu_{CL} = -0.59$, $\mu_{AH} = \mu_{CH} = 0.61$



(f) $\mu_{AL} = 0.41$, $\mu_{CL} = -0.59$, $\mu_{AH} = \mu_{CH} = 0.61$

Figure 1: Value Added Heterogeneity Simulation

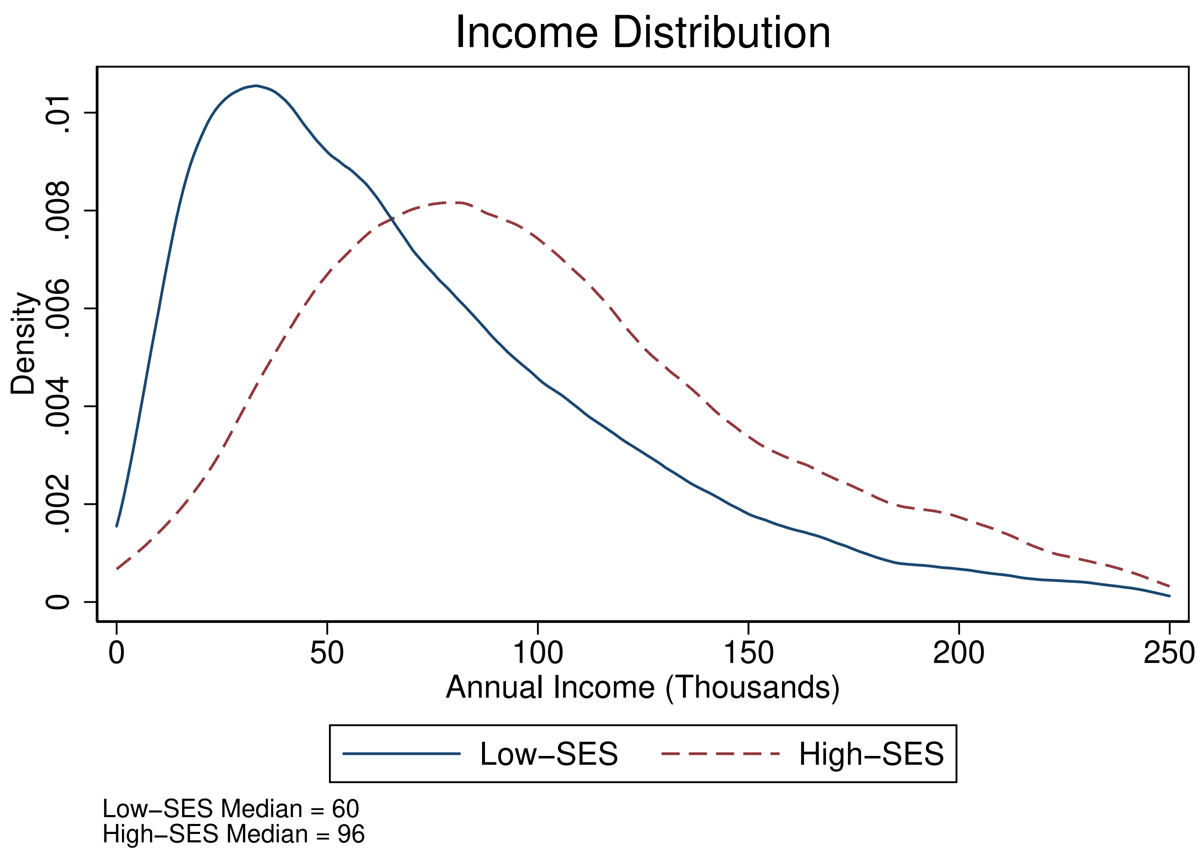
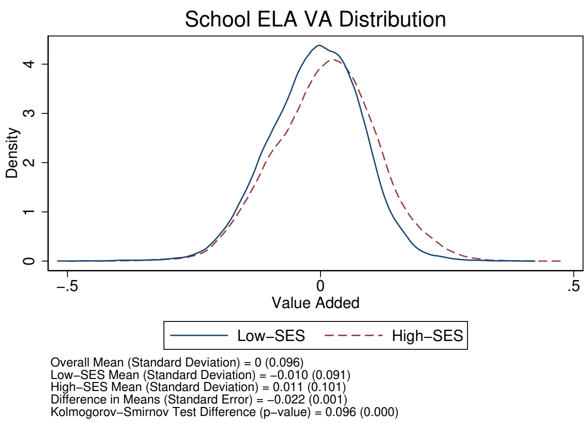
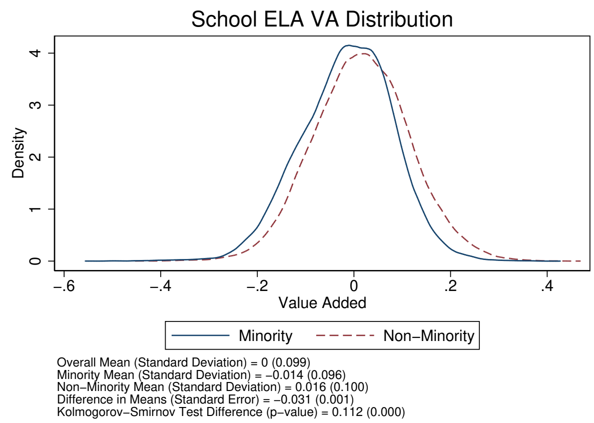


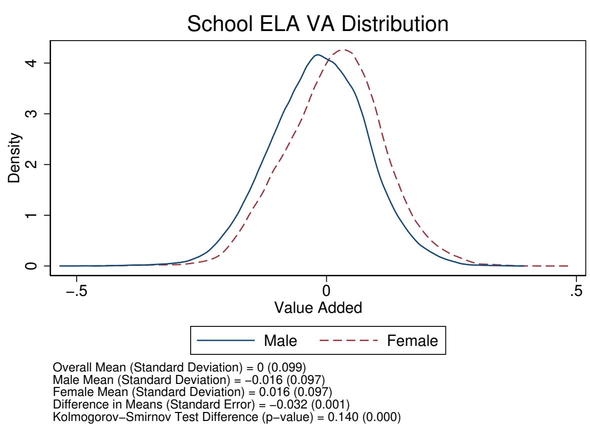
Figure 2: Income Distribution by Economic Disadvantage



(a) SES

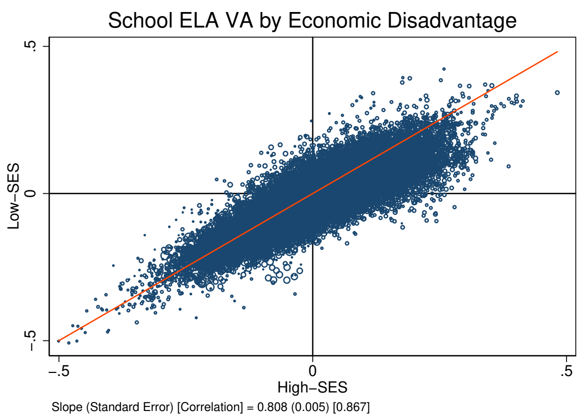


(b) Ethnicity

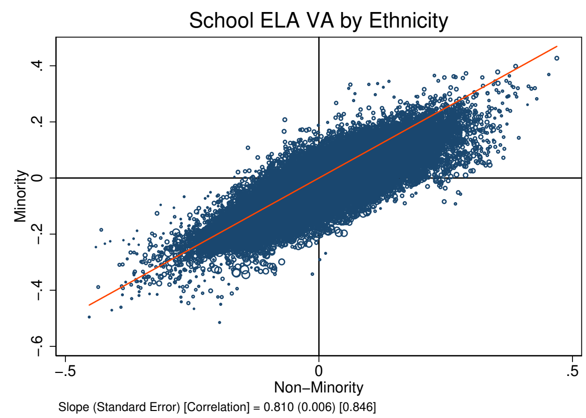


(c) Sex

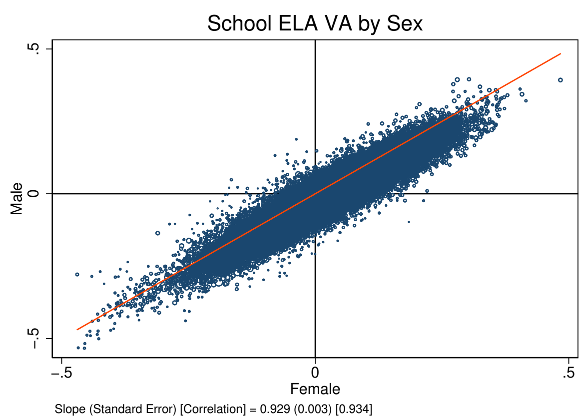
Figure 3: School by Subgroup Test Score Value Added Distributions



(a) SES

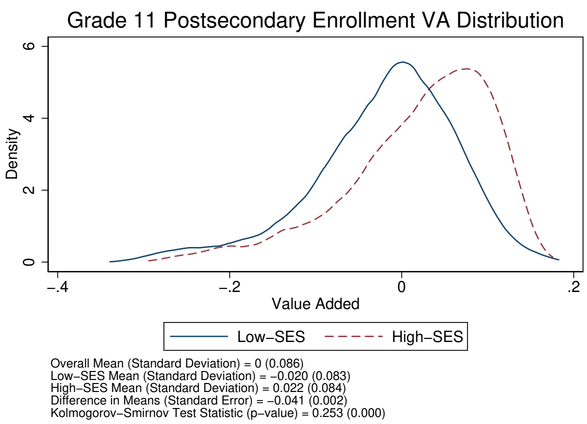


(b) Ethnicity

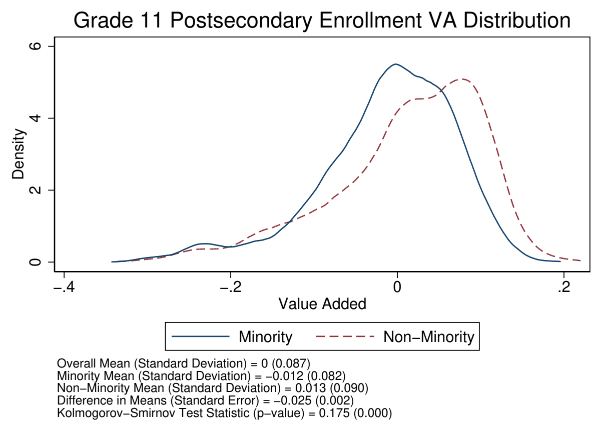


(c) Sex

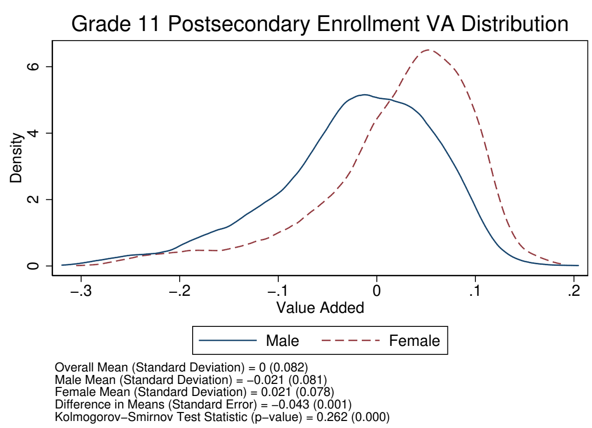
Figure 4: School Test Score Value Added by Subgroup



(a) SES

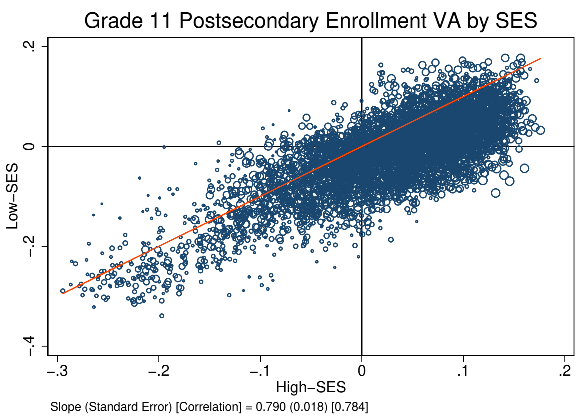


(b) Ethnicity

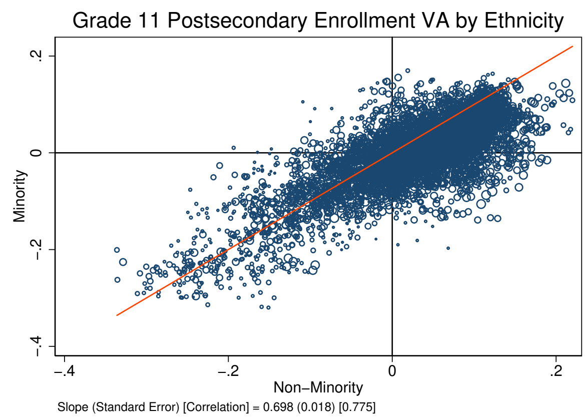


(c) Sex

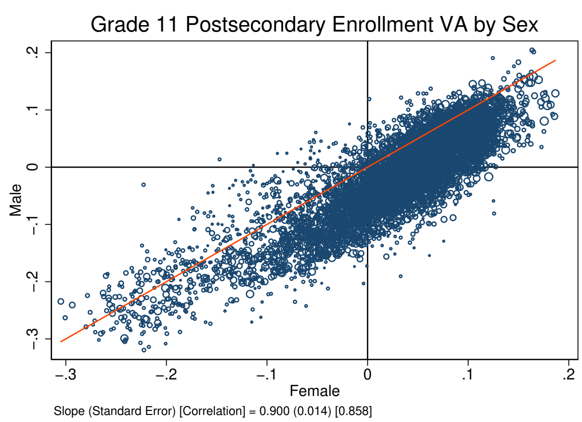
Figure 5: School by Subgroup Postsecondary Enrollment Value Added Distributions



(a) SES

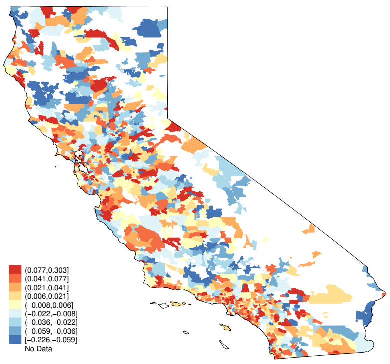


(b) Ethnicity

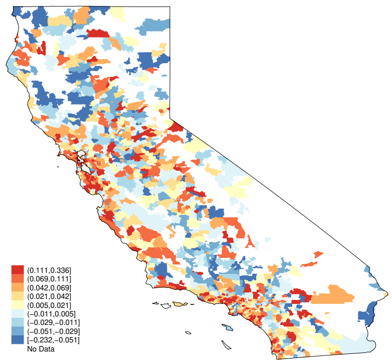


(c) Sex

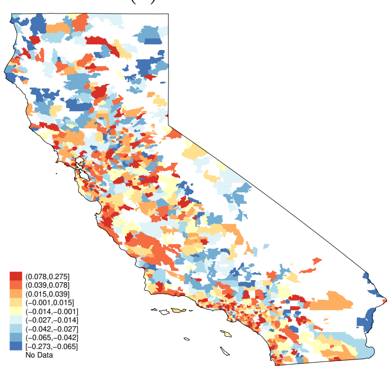
Figure 6: School Postsecondary Enrollment Value Added by Subgroup



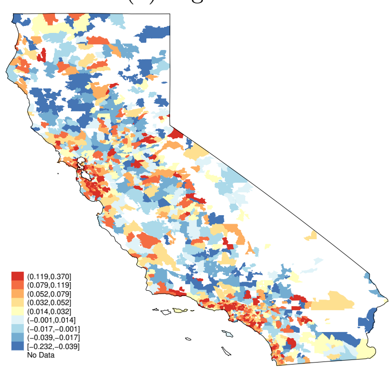
(a) Low-SES



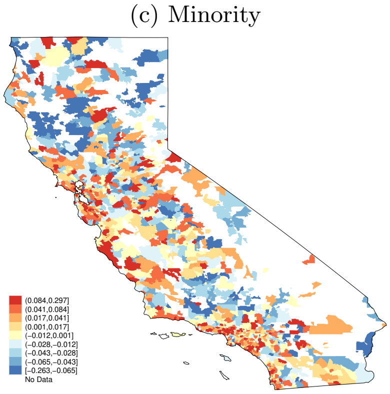
(b) High-SES



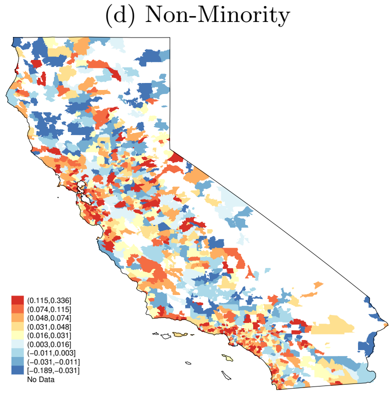
(c) Minority



(d) Non-Minority

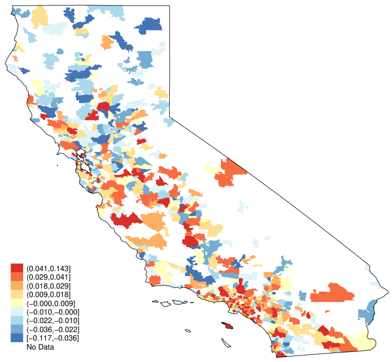


(e) Male

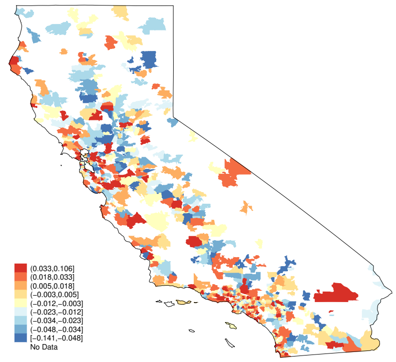


(f) Female

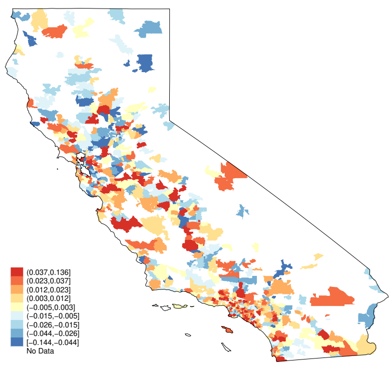
Figure 7: Map of School Test Score Value Added by Zip Code



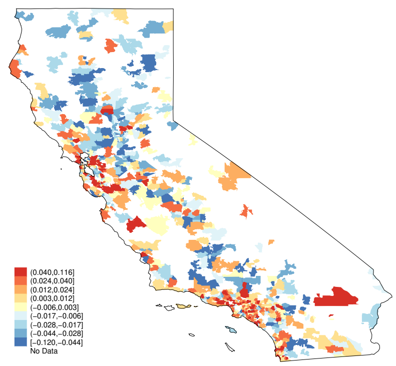
(a) Low-SES



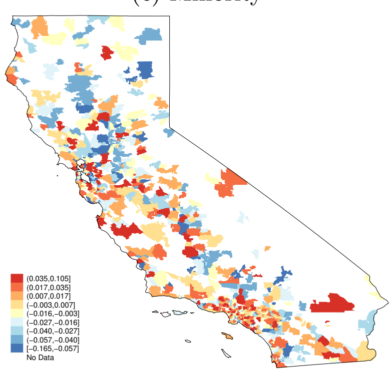
(b) High-SES



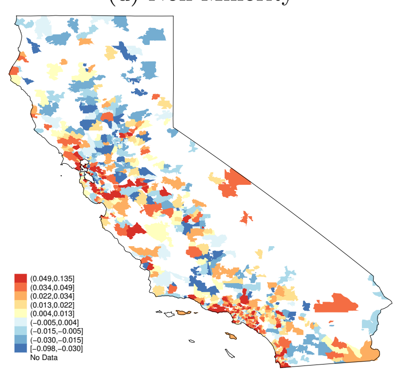
(c) Minority



(d) Non-Minority

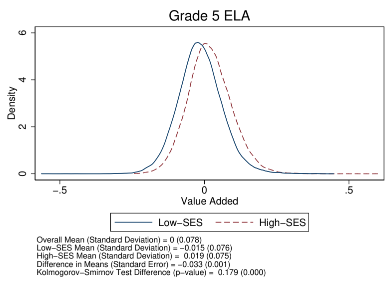


(e) Male

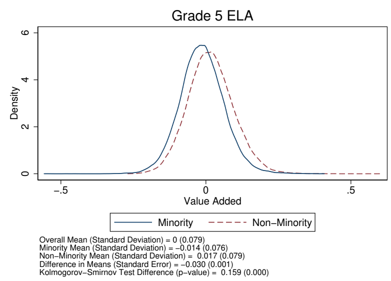


(f) Female

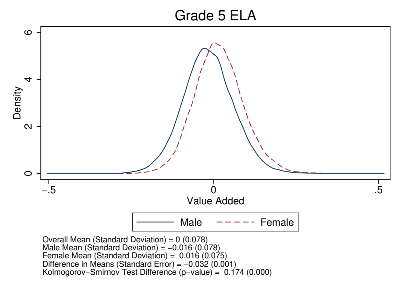
Figure 8: Map of School Postsecondary Enrollment Value Added by Zip Code



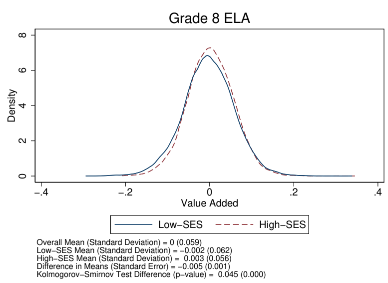
(a) SES: Grade 5



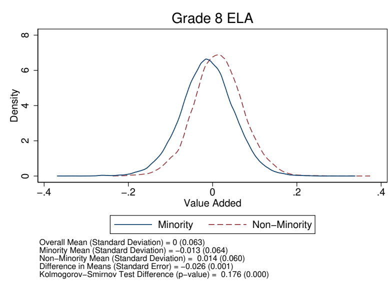
(b) Ethnicity: Grade 5



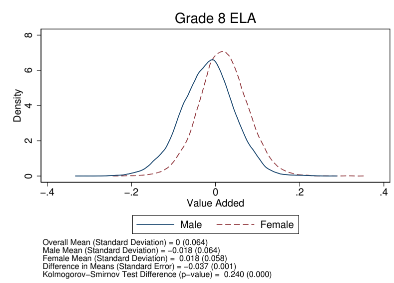
(c) Sex: Grade 5



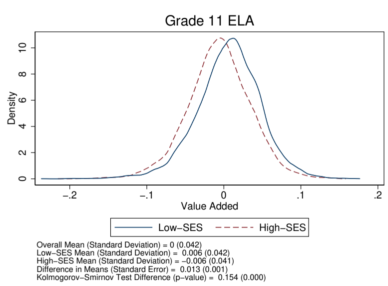
(d) SES: Grade 8



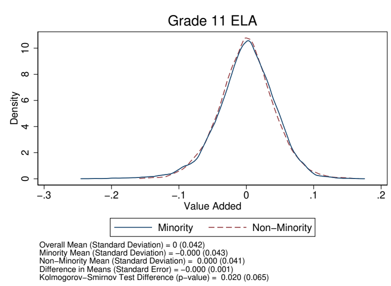
(e) Ethnicity: Grade 8



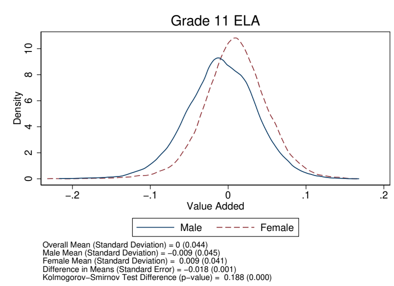
(f) Sex: Grade 8



(g) SES: Grade 11

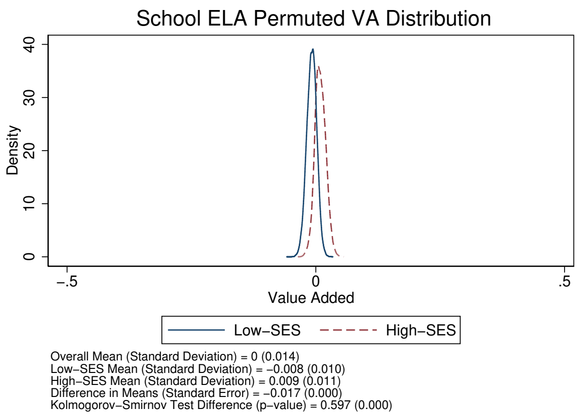


(h) Ethnicity: Grade 11

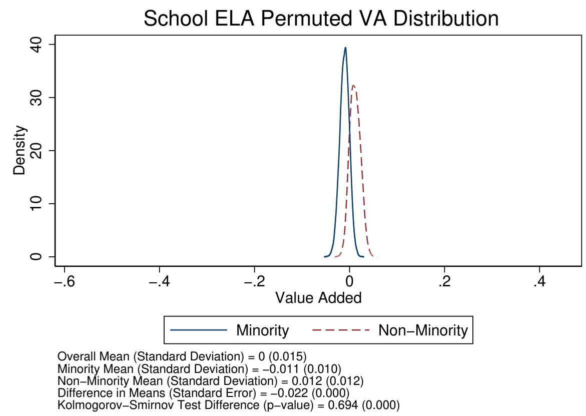


(i) Sex: Grade 11

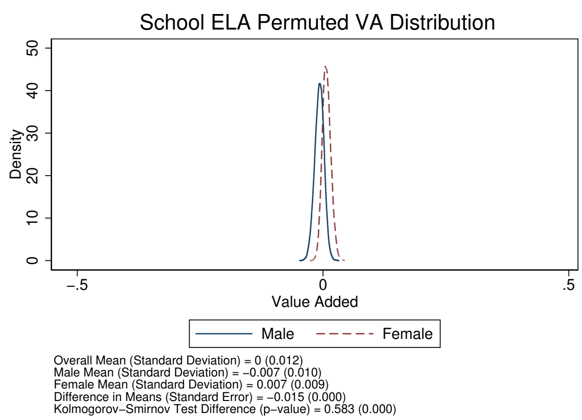
Figure B.1: School by Subgroup Test Score Value Added Distributions



(a) SES

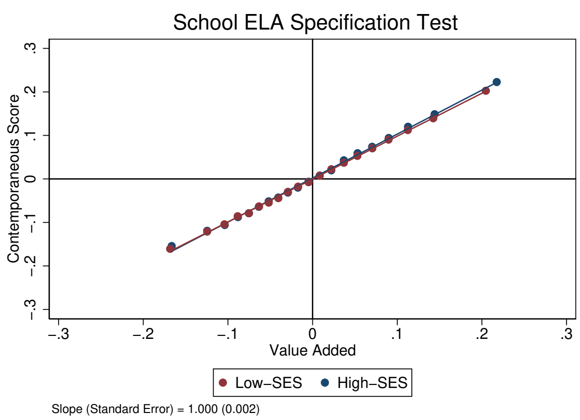


(b) Ethnicity

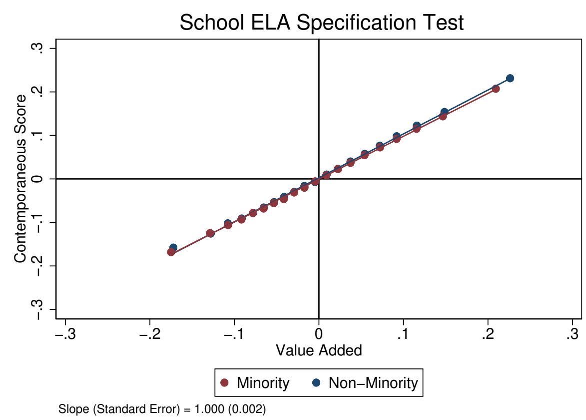


(c) Sex

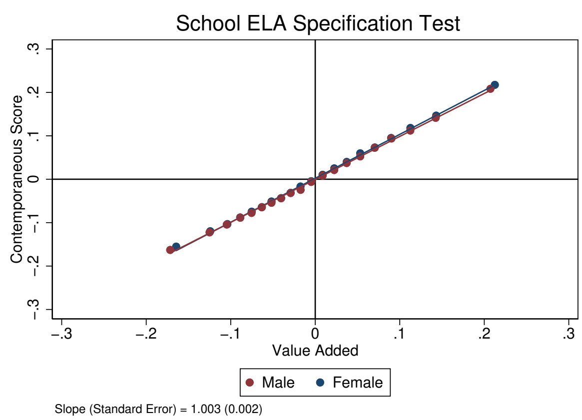
Figure B.2: School by Subgroup Test Score Permuted Value Added Distributions



(a) SES

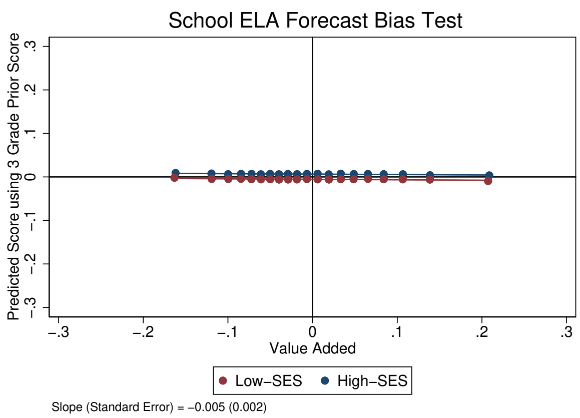


(b) Ethnicity

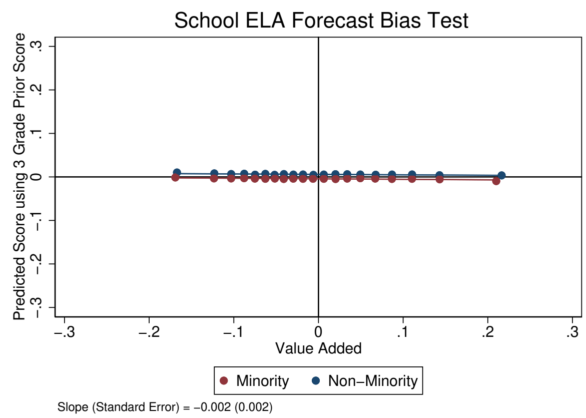


(c) Sex

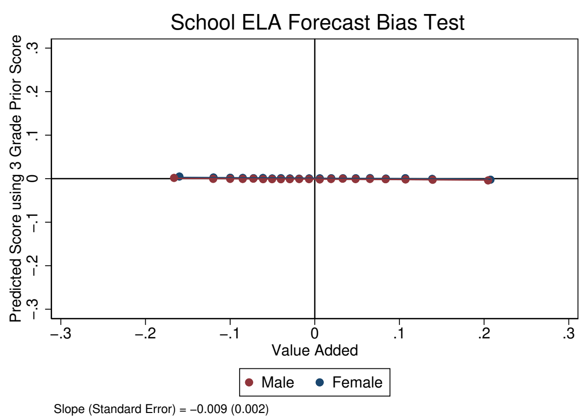
Figure B.3: School Test Score Value Added Specification Tests



(a) SES

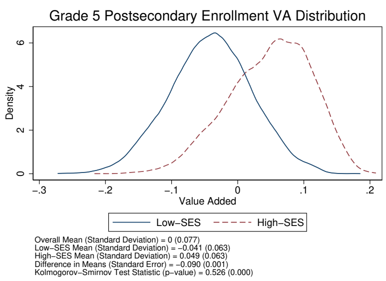


(b) Ethnicity

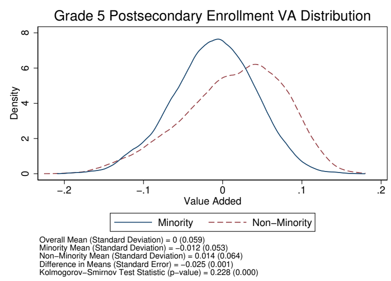


(c) Sex

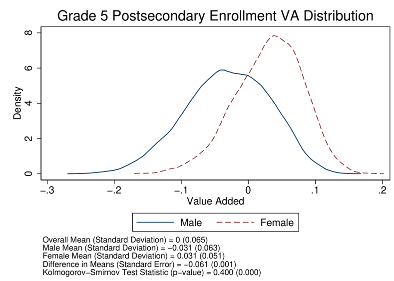
Figure B.4: School Test Score Value Added Forecast Bias Tests



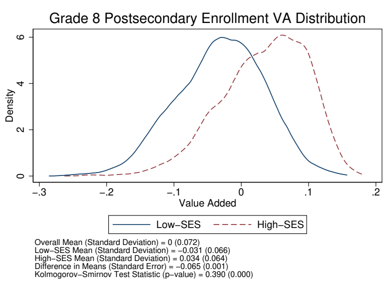
(a) SES: Grade 5



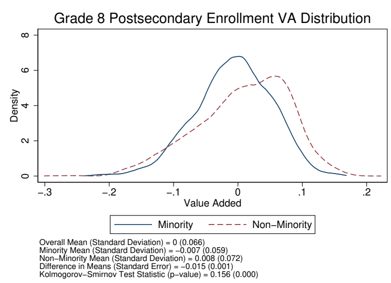
(b) Ethnicity: Grade 5



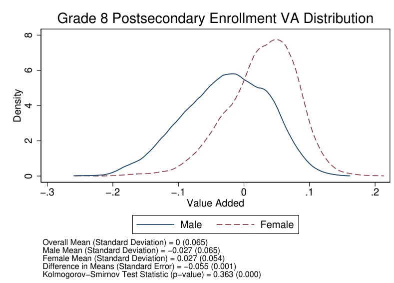
(c) Sex: Grade 5



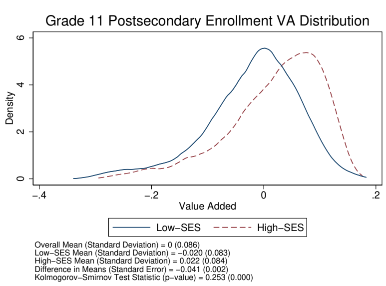
(d) SES: Grade 8



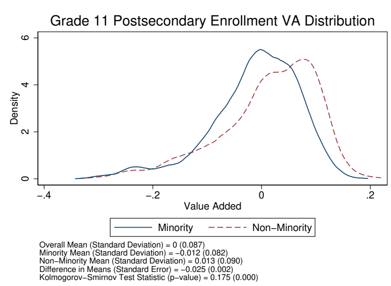
(e) Ethnicity: Grade 8



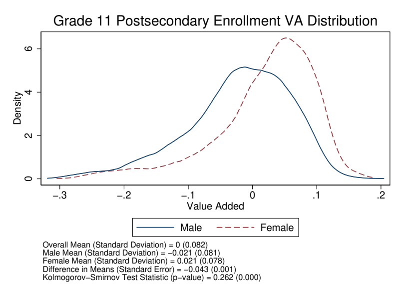
(f) Sex: Grade 8



(g) SES: Grade 11



(h) Ethnicity: Grade 11



(i) Sex: Grade 11

Figure B.5: School by Subgroup Postsecondary Enrollment Value Added Distributions

Table 1: Rank Example Value Added

s	μ_{sLt}	μ_{sHt}	$p(s = S)$
A	0.41	0.61	0.25
B	0.41	-0.39	0.25
C	-0.59	0.61	0.25
D	-0.59	-0.39	0.25

Table 2: Rank Example Parameters

$\theta_{st} \sim \mathcal{N}(0, 0.05)$
$\varepsilon_{isdt} \sim \mathcal{N}(0, 0.5)$
$N_{sLt}, N_{sHt} \sim U[7, 1007]$
$j_{sdt} \sim U[1, 1.1]$
of Schools = 250
of Years = 9

Table 3: K–12 Summary Statistics

	Low-SES	High-SES	Minority	Non-Minority	Male	Female
Demographic Controls						
Age in Years	12.97	13.35	13.05	13.27	13.17	13.11
	[2.294]	[2.328]	[2.303]	[2.330]	[2.320]	[2.315]
Male	0.493	0.501	0.491	0.504	1	0
	[0.500]	[0.500]	[0.500]	[0.500]	[0]	[0]
Hispanic or Latino	0.718	0.236	0.860	0	0.496	0.504
	[0.450]	[0.424]	[0.347]	[0]	[0.500]	[0.500]
White	0.110	0.509	0	0.694	0.294	0.287
	[0.313]	[0.500]	[0]	[0.461]	[0.455]	[0.452]
Asian	0.0867	0.179	0	0.306	0.132	0.125
	[0.281]	[0.383]	[0]	[0.461]	[0.338]	[0.331]
Black or African American	0.0712	0.0514	0.107	0	0.0602	0.0643
	[0.257]	[0.221]	[0.309]	[0]	[0.238]	[0.245]
Other Race	0.0139	0.0252	0.0327	0	0.0189	0.0192
	[0.117]	[0.157]	[0.178]	[0]	[0.136]	[0.137]
Economic Disadvantage	1	0	0.757	0.257	0.544	0.551
	[0]	[0]	[0.429]	[0.437]	[0.498]	[0.497]
Limited English Proficient Status	0.276	0.0413	0.256	0.0497	0.182	0.157
	[0.447]	[0.199]	[0.436]	[0.217]	[0.386]	[0.364]
Disabled	0.0424	0.0393	0.0420	0.0396	0.0549	0.0272
	[0.201]	[0.194]	[0.201]	[0.195]	[0.228]	[0.163]
Test Scores						
Current Test Score	-0.246	0.495	-0.218	0.517	0.000492	0.177
	[0.868]	[0.943]	[0.875]	[0.947]	[0.990]	[0.953]
1 Grade Prior Test Score	-0.237	0.521	-0.206	0.539	0.0224	0.188
	[0.859]	[0.933]	[0.868]	[0.938]	[0.980]	[0.953]
2 Grade Prior Test Score	-0.233	0.532	-0.199	0.547	0.0342	0.191
	[0.858]	[0.931]	[0.867]	[0.938]	[0.976]	[0.956]
Observations	12,171,238	10,044,078	12,922,696	9,288,668	11,041,063	11,189,152
R^2						

Values are means and standard deviations [in brackets] of the dependent and independent variables used in the test score value added estimation. Only students included in the test score value added sample are included this table. Data comes from grades 4-11 in public schools in the state of California between the 2004-2005 and 2012-2013 school years.

Table 4: K–12 Peer Summary Statistics

	Low-SES	High-SES	Minority	Non-Minority	Male	Female
Peer Controls						
Peer Avg. Age in Years	12.99	13.39	13.08	13.30	13.17	13.18
	[2.260]	[2.298]	[2.271]	[2.300]	[2.284]	[2.288]
Peer Avg. Male	0.510	0.509	0.509	0.510	0.510	0.509
	[0.0451]	[0.0437]	[0.0440]	[0.0451]	[0.0442]	[0.0448]
Peer Avg. Hispanic or Latino	0.633	0.336	0.634	0.311	0.497	0.500
	[0.258]	[0.231]	[0.257]	[0.212]	[0.288]	[0.287]
Peer Avg. White	0.178	0.420	0.182	0.434	0.288	0.286
	[0.201]	[0.242]	[0.198]	[0.245]	[0.252]	[0.251]
Peer Avg. Asian	0.0958	0.161	0.0886	0.177	0.126	0.125
	[0.130]	[0.171]	[0.113]	[0.184]	[0.154]	[0.152]
Peer Avg. Black or African American	0.0767	0.0579	0.0782	0.0542	0.0678	0.0686
	[0.101]	[0.0760]	[0.104]	[0.0647]	[0.0902]	[0.0914]
Peer Avg. Other Race	0.0202	0.0305	0.0208	0.0305	0.0249	0.0248
	[0.0369]	[0.0413]	[0.0412]	[0.0357]	[0.0394]	[0.0392]
Peer Avg. Economic Disadvantage	0.720	0.351	0.684	0.371	0.552	0.555
	[0.234]	[0.243]	[0.257]	[0.260]	[0.301]	[0.301]
Peer Avg. Limited English Proficient Status	0.259	0.113	0.247	0.117	0.193	0.193
	[0.171]	[0.110]	[0.171]	[0.115]	[0.164]	[0.164]
Peer Avg. Disabled	0.0855	0.0752	0.0840	0.0764	0.0813	0.0804
	[0.0548]	[0.0528]	[0.0543]	[0.0536]	[0.0540]	[0.0542]
Peer Avg. 1 Grade Prior Test Score	-0.143	0.325	-0.113	0.323	0.0701	0.0675
	[0.351]	[0.408]	[0.369]	[0.413]	[0.445]	[0.443]
Peer Avg. 2 Grade Prior Test Score	-0.125	0.345	-0.0960	0.343	0.0889	0.0862
	[0.351]	[0.411]	[0.370]	[0.415]	[0.446]	[0.445]
Observations	12,171,238	10,044,078	12,922,696	9,288,668	11,041,063	11,189,152
R^2						

Values are means and standard deviations [in brackets] of the peer (jackknife) averages of the independent variables used in the test score value added estimation. Peer averages are calculated at the school by grade by year level and include students omitted from the value added sample. Only students included in the test score value added sample are included this table. Data comes from grades 4-11 in public schools in the state of California between the 2004-2005 and 2012-2013 school years.

Table 5: Postsecondary Summary Statistics

	Low-SES	High-SES	Minority	Non-Minority	Male	Female
Enrolled at a Postsecondary Institution	.581	.755	.597	.749	.62	.703
	[.493]	[.43]	[.49]	[.434]	[.485]	[.457]
Enrolled at a 2-Year College	.385	.37	.396	.355	.38	.376
	[.487]	[.483]	[.489]	[.478]	[.485]	[.484]
Enrolled at a 4-Year University	.196	.385	.202	.394	.24	.326
	[.397]	[.487]	[.401]	[.489]	[.427]	[.469]
Enrolled at a Public Institution	.551	.658	.56	.657	.573	.628
	[.497]	[.474]	[.496]	[.475]	[.495]	[.483]
Enrolled at a Private Institution	.03	.0971	.0379	.0923	.0471	.0746
	[.171]	[.296]	[.191]	[.289]	[.212]	[.263]
Enrolled at a CA Institution	.558	.645	.565	.643	.563	.632
	[.497]	[.478]	[.496]	[.479]	[.496]	[.482]
Enrolled at an Out-of-State Institution	.0232	.11	.0322	.106	.0563	.0705
	[.151]	[.313]	[.176]	[.308]	[.23]	[.256]
Observations	8,124,601	7,017,998	8,696,875	6,443,249	7,476,349	7,674,020
R^2						

Values are means and standard deviations [in brackets]. Only students included in the postsecondary value added sample are included this table. Data comes from grades 4-11 in public schools in the state of California between the 2004-2005 and 2012-2013 school years.

Table 6: School Value Added Characteristics

	Test Scores			Enrollment		
	SES	Ethnicity	Sex	SES	Ethnicity	Sex
FTE Teachers per Student	9.75***	5.07***	5.11***	-17.6***	-8.73	-7.33
	(1.79)	(1.84)	(1.42)	(5.1)	(6.28)	(4.82)
FTE Pupil Services per Student	-35***	-31.3***	-41.2***	-3.25	-11.6	3.2
	(3.91)	(4.02)	(3.31)	(14.9)	(17.2)	(13.9)
English Learner Staff Per Student	3.25***	3.24***	4.38***	-5.51***	-5.33***	-4.25***
	(.447)	(.468)	(.417)	(1.3)	(1.44)	(1.15)
Proportion $\leq$ 3 Years Experience Teachers	-.486***	-.368***	-.3***	-.162	.26	-.217
	(.0971)	(.1)	(.0745)	(.272)	(.336)	(.277)
Proportion Full Credential Teachers	.411***	.195	.391***	.955**	1.21**	.715*
	(.152)	(.168)	(.0977)	(.448)	(.544)	(.366)
Proportion Male Teachers	-4.16***	-4***	-3.64***	.489	.345	.866**
	(.0764)	(.0808)	(.0616)	(.426)	(.546)	(.437)
Proportion Minority Teachers	-.422***	.13	.0274	-2.23***	-.383	-.554***
	(.082)	(.0952)	(.0513)	(.224)	(.326)	(.195)
Low-SES/Minority/Male $\times$ FTE Teachers per Student	-8.75***	-2.24	.977	23***	8.91	5.19*
	(1.98)	(2.02)	(.626)	(5.31)	(6.14)	(3)
Low-SES/Minority/Male $\times$ FTE Pupil Services per Student	-24.4***	-29***	-2.74*	18.9	28.6	6.59
	(4.72)	(4.51)	(1.53)	(17)	(18.5)	(9.65)
Low-SES/Minority/Male $\times$ English Learner Staff Per Student	3.95***	3.68***	.139	2.04	.827	-2.68**
	(.656)	(.565)	(.191)	(1.68)	(1.56)	(1.1)
Low-SES/Minority/Male $\times$ Proportion $\leq$ 3 Years Experience Teachers	.203**	.0887	-.126***	.216	-.271	.539***
	(.103)	(.11)	(.0328)	(.323)	(.36)	(.172)
Low-SES/Minority/Male $\times$ Proportion Full Credential Teachers	.176	.47***	.0348	-.243	-.719	.338
	(.149)	(.167)	(.0492)	(.49)	(.549)	(.242)
Low-SES/Minority/Male $\times$ Proportion Male Teachers	1.3***	.989***	.312***	-.302	-.208	-1.13***
	(.0817)	(.0853)	(.0286)	(.407)	(.466)	(.212)
Low-SES/Minority/Male $\times$ Proportion Minority Teachers	.732***	.0863	.0759***	2.39***	.0542	-.63***
	(.0864)	(.102)	(.0232)	(.267)	(.358)	(.12)
Observations	56,355	57,232	61,322	1,968	1,963	2,081
R^2	.538	.529	.53	.302	.198	.185

Each column represents a separate regression of value added on school characteristics. Standard errors clustered at the school level are presented in parentheses.

Table A.1: K–12 Counts

	Low-SES	High-SES	Minority	Non-Minority	Male	Female
All Students	19,075,808	15,610,340	20,311,735	14,211,766	17,805,722	16,917,263
+ Nonmissing Test Score	18,075,617	15,155,345	19,295,255	13,776,157	16,863,634	16,392,791
+ First Test Score for Grade	17,292,950	14,564,130	18,413,136	13,297,435	16,110,124	15,761,440
+ Conventional School	16,517,224	14,008,368	17,570,516	12,820,318	15,361,367	15,176,934
+ School Size > 10	16,516,147	14,007,304	17,569,766	12,818,982	15,360,248	15,175,897
+ Nonmissing Subject Test Score	16,379,207	13,906,140	17,419,734	12,720,672	15,234,127	15,066,203
+ Nonmissing Demographic Controls	15,792,205	13,420,922	16,839,919	12,367,166	14,561,263	14,664,021
+ Nonmissing 1 Grade Prior Test Score	13,946,145	11,666,105	14,810,002	10,797,373	12,738,844	12,883,208
+ Nonmissing 2 Grade Prior Test Score	12,185,012	10,060,598	12,935,577	9,306,276	11,053,300	11,200,792
+ Nonmissing Peer Controls	12,183,634	10,059,133	12,934,328	9,304,715	11,051,847	11,199,483
+ School VA Sample Size $\geq$ 7	12,171,238	10,044,078	12,922,696	9,288,668	11,041,063	11,189,152

Values are counts of the number of observations in each sample. Each row is additive, so the restrictions from all prior rows are also present in the current row. Data comes from grades 4-11 in public schools in the state of California between the 2004-2005 and 2012-2013 school years.

Table B.1: School Test Score Value Added Specification/Forecast Bias Tests

	SES	Ethnicity	Sex
VA Specification Test: Contemporaneous Score	1.000	1.000	1.003
	(0.002)	(0.002)	(0.002)
	[0.997,1.004]	[0.997,1.003]	[1.000,1.006]
VA Forecast Bias Test: Prior Score	-0.005	-0.002	-0.009
	(0.002)	(0.002)	(0.002)
	[-0.009,-0.002]	[-0.005,0.002]	[-0.012,-0.006]

Each cell represents a separate regression. The first row contains the coefficient for a bivariate regression of test score residuals r_{isgdt} on school value added $\hat{\mu}_{sdt}$. Statistical inference is conducted under the null hypothesis that the coefficient equals 1. The second row contains the coefficient for a regression of the projection of test scores onto three grade prior test scores on school value added $\hat{\mu}_{sdt}$. Statistical inference is conducted under the null hypothesis that the coefficient equals 0. Standard errors cluster bootstrapped at the school level are presented in parentheses. The 95% confidence intervals are presented in brackets.