Increasing Bias in SAT Test Scores? A Variation on Simpson's Paradox

Zachary Bleemer[a,b]

[a] *Department of Economics, University of California, Berkeley*
[b] *Center for Studies in Higher Education, University of California, Berkeley*

UC-CHP Policy Brief 2020.2
February 2020

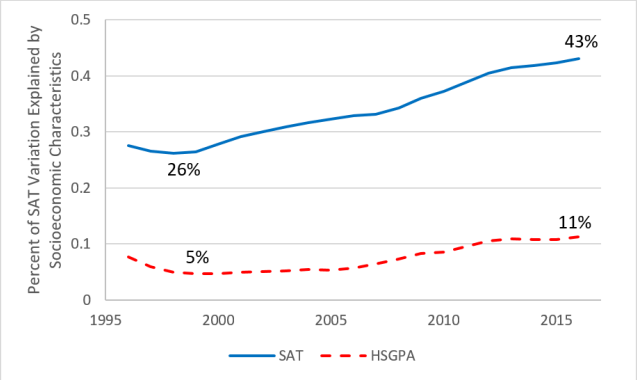
Geiser (2015) documents a potentially-alarming increase in the correlation between University of California (UC) applicants' SAT scores and their socioeconomic characteristics–ethnicity, parental income, and parental education–since 2000. This brief decomposes the increase into three possible explanations: increased SAT testing bias, increased educational stratification, and changes in the composition of UC applicants. About one-third of the increase can be explained by cross-school differences in California high school quality, with the remainder explained by increased UC applicant heterogeneity. The results manifest a variation on Simpson's Paradox: while student socioeconomic characteristics explain an increasing within-high-school share of applicants' SAT scores across the nine UC campuses, there is no such increase among applicants to any one UC campus.

1 Introduction

Standardized tests like the SAT have long been an important component of selective university admissions decisions, but large performance gaps by race and class have provoked substantial controversy, challenging the exams' evaluative validity.[1] As universities across the country have re-examined whether to continue mandating standardized test submission on undergraduate applications, increased attention has focused on the correlation between SAT performance and applicants' socioeconomic characteristics.[2]

Figure 1 displays an apparently disturbing trend in the proportion of SAT score variation that can be ex-

Figure 1: *Demographics' Explanatory Power for SAT and HSGPA*



Note: Two-year moving-average R^2 coefficients from annual OLS regressions of University of California applicants' SAT score or high school GPA on comprehensive parental education indicators, ethnicity indicators, and family income (and an indicator whether family income is reported). Sample restricted to California-resident freshmen. Source: UC Corporate Student System

Thanks to David Card, Tongshan Chang, Charles Masten, and Jesse Rothstein for helpful comments. Much of the analysis discussed in this brief was first published as an institutional research report (University of California, 2020). The author was employed by the University of California in a research capacity throughout the period during which the study was conducted. Remaining errors are my own.

[1] Crouse and Trusheim (1988); Rothstein (2004); Robinson and Monks (2005).
[2] Reardon, Kalogrides, and Shores (2016); Owens (2018).

plained by socioeconomic characteristics among applicants to the nine undergraduate University of California campuses (first identified by Geiser (2015)). Annually-estimated linear regressions modeling SAT score as a function of measures of applicant ethnicity, family income, and parental education suggest that the models' fit has dramatically improved over time; while the socioeconomic characteristics explained 26 percent of variation in applicants' SAT scores in the late 1990s, they explained more than 40 percent of variation in 2016. Meanwhile, the proportion of explained variation of applicants' high school GPAs (HSGPAs) has remained at a far lower level, rising from 5 percent in the 1990s to 12 percent in 2016. When similar models are estimated separately for each socioeconomic characteristic, each of the three explains an additional 10 percentage points of SAT score variation since the 1990s.[3]

There are at least two well-known interpretations of this trend, which I refer to as the "Socioeconomic Testing Trend" (STT). One is that the SAT is increasingly racist, classist, or otherwise biased against students from disadvantaged backgrounds over time, such that despite their unvarying real 'aptitude' over the past 25 years, disadvantaged applicants' SAT performance has deteriorated (while HSGPA may not reflect the same systematic bias). The second is that increased residential segregation by ethnicity and class–combined with increased educational disparities across California secondary schools–have led to actual average declines in college preparation among disadvantaged groups, as accurately reflected by their declining SAT scores (but perhaps not by HSGPAs, which are normed within increasingly-disparate high schools).[4]

A third possible explanation is that the STT reflects changes in the composition of applicants in the estimation sample, most likely as a result of changing UC admissions policies. Expanding admissions policies that favor disadvantaged applicants, for example, could mechanically increase socioeconomic characteristics' explanatory power by growing the number of low-SAT high-disadvantage applicants in the estimation sample.

Arbitrating between these alternative explanations for the STT is important in order to understand how the SAT's informativeness has evolved over the past 25 years. Advocates in favor of the first explanation (the "Bias Explanation") include Geiser (2015, 2016), who argues that the STT invalidates use of the SAT in the absence of race-based affirmative action because it increasingly favors White and Asian applicants. Advocates for the second explanation (the "Real-Disparities Explanation") include the College Board, which argued in a recent meeting with the University of California Academic Council Standardized Testing Task Force that "performance on the SAT differs across subgroups, which largely reflects educational differences in high schools". The third explanation (the "Compositional Explanation") has drawn less vocal support.

This brief presents evidence that approximately one-third of the STT is explained the Real-Disparities Explanation, with two-thirds explained by the Compositional Explanation. The next section discusses the data and empirical methodology used in the brief, with the following section summarizing its results. Figure 2 shows that estimating the STT with high school fixed effects, netting-out the Real-Disparities Explanation, substantially reduces its side and slope. Figures 3 and 4 motivate the Compositional Explanation, showing the magnitude of compositional changes in the UC applicant pool over the past 25 years.

Figure 5 shows that plotting the STT separately by campus, including high school fixed effects to capture changes in Real Disparities, eliminates the trend completely; in fact, socioeconomic characteristics' within-school explanatory power declined among applicants to eight of the nine undergraduate UC campuses between 1996 and 2016. This reversal is reminiscent of Simpson (1951)'s Paradox, with an aggregate trend failing to appear in the subgroups that comprise it.

2 Data and Methodology

The data analyzed in this brief include comprehensive California-resident freshman applications to any University of California campus between 1995 and 2016. I observe each applicant's application year, graduating high school, SAT mathematics and reading comprehension component scores, and self-reported weighted high school GPA along with application and enrollment indicators at each of the nine campuses.[5] I also observe the following socioeconomic indices:

1. Fifteen-category reported ethnicity
2. Family income as reported on the student's federal financial aid application
3. Seven-category reported parental education (from no high school degree to graduate degree), for each

[3] See Appendix A for details on how these estimates differ from those reported in Geiser (2015).

[4] Owens, Reardon, and Jencks (2016) provides evidence of increasing income segregation across US high schools between the 1990s and 2010. Card and Rothstein (2007) show that test score stratification by ethnicity is greater in more-segregated cities; Vigdor and Ludwig (2007) survey evidence that segregation causes testing gaps between white and black SAT-takers.

[5] The SAT made several substantial changes to its test implementation during the sample period, including adding a mandatory writing section in addition to mathematics and reading comprehension. In order to best preserve comparability across years, I focus exclusively on students' performance on the latter two sections, which are consistently scored from 200 to 800 (mean 500, s.d. approximately 100) throughout the period.

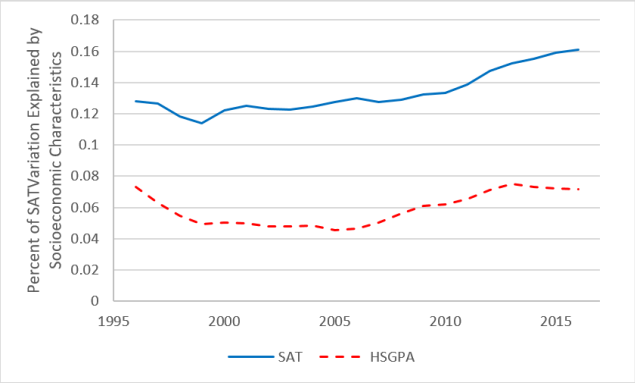
of two parents

I estimate the STT presented in Figure 1 by calculating R^2 coefficients for the following model estimated separately by application year:

$$SAT_{it} = \alpha_t Eth_{it} + \beta_t Inc_{it} + \gamma_t Educ_{it} + \epsilon_{it} \qquad (1)$$

where SAT_{it} is the combined mathematics and reading comprehension SAT score submitted by applicant i in year t; Eth_{it} includes indicators for each observed ethnicity; Inc_{it} includes log family income, and indicator for unobserved family income (if the applicant did not apply for federal financial aid), and an indicator for 0 family income; and $Educ_{it}$ includes indicators for applicants' interacted (and ordered) parental education.[6] Coefficients and standard errors are not reported.

Figure 2: *Within High School STT Trend*



Note: Two-year moving-average R^2 coefficients from annual OLS regressions of applicants' or enrollees' SAT score or high school GPA on comprehensive parental education indicators, ethnicity indicators, and family income (and an indicator whether family income is reported), net of fixed effects by origin high school. Sample restricted to California-resident freshmen. Source: UC Corporate Student System

3 The Real-Disparities Explanation

In order to estimate the degree to which the STT is explained by increasing educational disparities across California high schools, I re-estimate Equation 1 adding high school fixed effects δ_{h_i}, which absorb cross-school variation in SAT scores. If the Real-Disparities Explanation were to fully explain the STT (that is, if the increase in SAT stratification by socioeconomic characteristics was exclusively occurring *across* high schools, without increasing disparities *within* high schools), then the resulting estimates would show no change in demographics' net explanatory power for applicants' or enrollees' SAT scores.

Figure 2 shows that an upward trend in socioeconomic characteristics' explanatory power persists, but less steeply and from a much lower base. While far less variation in SAT scores can be explaned by demographics when only comparing students to others from their same high schools, there is still a notable upward trend in demographics' explanatory power, from 11.4 percent at its trough in 1999 to 16.1 percent in 2016 among UC applicants (representing a 37 percent decline in the proportional increase in explanatory power between 1999's trough and 2016). Socioeconomics' explanatory power for HSGPA increased from 4.9 to 7.2 percent in the same period.[7]

One interesting feature that strengthens in these within-school estimates is the decline in socioeconomic characteristics' SAT explanatory power between 1995 and 1999. This declines likely reflect that period's phasing out of UC's affirmative action program, which decreased underrepresented minorities applicants' UC enrollment by at least 700 students per year across all campuses (Bleemer, 2019) and caused declines in low-SAT URM applications (Card and Krueger, 2005). This observation provides the first evidence supporting the role of student composition in regulating the correlational relationship between socioeconomic characteristics and SAT performance, the subject of the brief's next section.

4 The Compositional Explanation

There is strong reason to suspect that substantial compositional changes in the University of California's applicant pool since the late 1990s have played an important role in producing the STT. Consider Figure 3, which plots the variance of applicants' and enrollees' normalized SAT scores and HSGPAs across all UC campuses annually since the mid-1990s.[8] The figure shows that the amount of variance in SAT scores among UC applicants has been swiftly rising since 2005, while the amount of variance in HSGPA fell in the late 1990s and early 2000s and has persisted at the lower level. The trends among UC enrollees are even more pronounced; variation in SAT scores among UC enrollees has increased by more than 10 percent since 1996, while variation in HSGPAs has fallen by more than 20 percent. These trends likely reflect two important admissions policies—Eligibility in the Local Context and Holistic Review—that have substantially replaced affirmative action since the 1990s in enrolling disadvantaged applicants. As various UC campuses increase

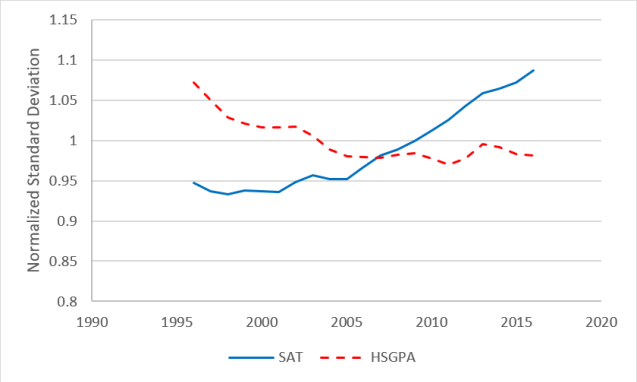
[6] Models are estimated using the *felm* package in *R*.

[7] Figures 2 and 5 present 'projected R^2' measures from the relevant annual linear regression estimates, which excludes variation explained by the high school fixed effects.

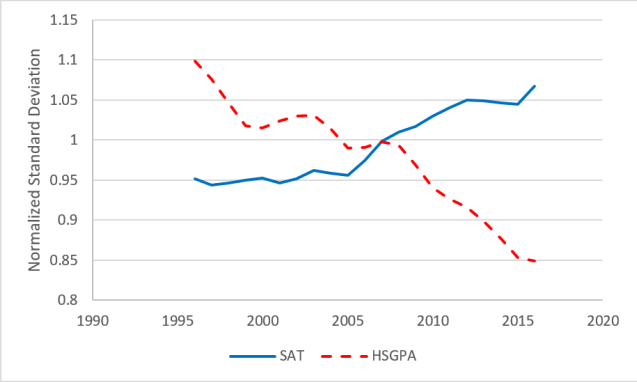
[8] For visualization purposes, SAT scores and high school GPAs are normalized to have standard deviation 1 across the sample period.

Figure 3: *Change in Normalized Annual SAT and HSGPA Standard Deviations Since 1995*

(a) *University of California Applicants*



(b) *University of California Enrollees*

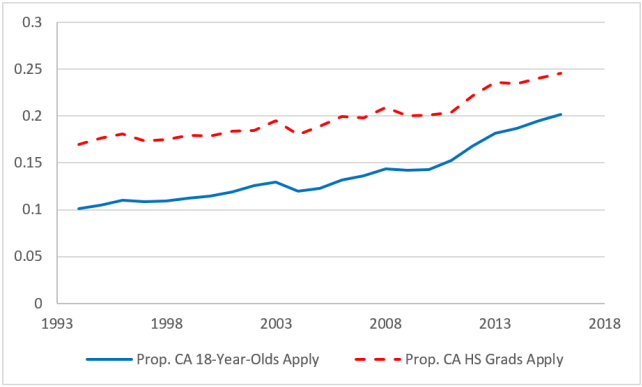


Note: Annual standard deviation in SAT and HSGPA of UC applicants and enrollees. SAT and HSGPA are each normalized to have standard deviation 1 on average across all years. Plot shows two-year moving averages. Source: UC Corporate Student System

their numbers of low-SAT high-HSGPA students targeted by those policies (Bleemer, 2018, 2019), SAT scores are increasingly varying across the campuses' student bodies. In other words, the increased variance is by design, an artifact of admissions policies which intentionally target lower-SAT applicants.

Figure 4 displays the proportion of California 18-year-olds and high school graduates who apply to at least one University of California campus. One of the chief successes–and political challenges–of UC's Comprehensive and Holistic Review admissions programs has been their encouragement of applications from high school graduates who would previously have not applied to UC because of their poor perceived likelihood of admission. Figure 4, which is borrowed from Douglass and Bleemer (2018), shows that the proportion of 18-year-olds in California who apply to at least one UC campus has doubled since 1995, from about 10 percent to about 20 percent. Some of this increase comes from increasing high school graduation rates, but even among graduates the proportion of applicants has increased by about 8 percentage points, to almost 25 percent. This change in application behavior has surely dramatically altered the composition of UC applicants, and is also reflected in UC's students as a result of changing admissions policies.

Figure 4: *Proportion of Young Californians Who Apply to UC*



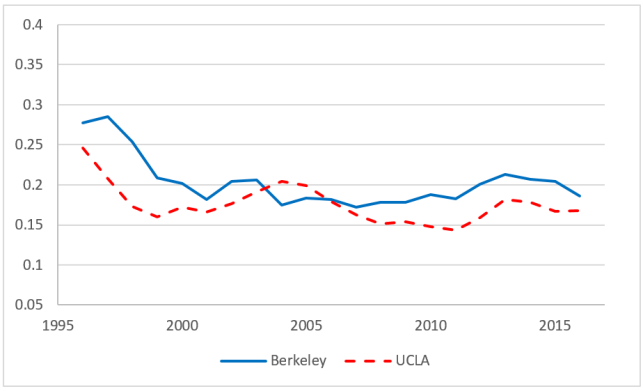
Note: The proportion of California 18-year-olds and California high school graduates who apply to at least one UC campus in each year since 1994. The annual number of California 18-year-olds is as estimated by the California Department of Finance, which also reports the annual number of high school graduates in the state. Source: UC Corporate Student System and California Department of Finance

UC applicants' increasing SAT variation and increasing breadth of California youths strongly suggest that compositional changes in UC applicants and enrollees are central factors in explaining the STT: after all, UC has spent the past 20 years bolstering admissions policies that favor the lower-SAT disadvantaged applicants who would mechanically increase the SAT-demographics correlation.

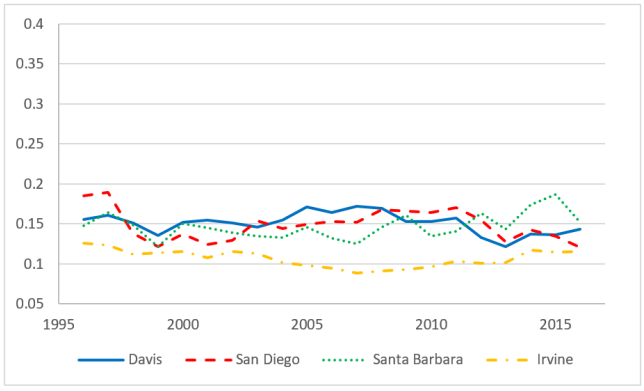
In order to conduct a simple test of the Compositional Explanation, I replicate Figure 2 by UC campus. Under either the Bias Explanation or the Real-Disparities Explanation, one would expect that the predictive power of the SAT has increased consistently at every UC campus, either because of consistent bias or consistently-varying within-school disparities across applicants. In fact, Figure 5 shows a very different pattern. Only a single campus, Santa Barbara, manifests any increase in the correlation between SAT scores and socioeconomic characteristics, and even there the increase in demographics' explanatory power for the SAT is slight (from 14.8 to 16.1 percent). Most other campuses have faced almost no change in demographics' explanatory power since the end of affir-

Figure 5: *Within High School STT Trend by Campus*

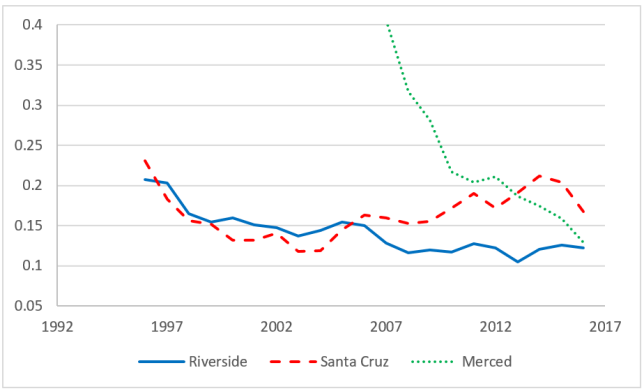
(a) *Most-Selective UC Campuses*



(b) *Middle-Selective UC Campuses*



(c) *Least-Selective UC Campuses*



Note: Two-year moving-average R^2 coefficients from annual OLS regressions of applicants' SAT scores on comprehensive parental education indicators, ethnicity indicators, and family income (and an indicator whether family income is reported), net of fixed effects by origin high school. Estimated separately for each UC campus. Sample restricted to California-resident freshmen. Source: UC Corporate Student System

mative action in the late 1990s (like Irvine, Davis, and UCLA) or have actually experienced declines in demographics' explanatory power (San Diego, Riverside, and most notably Merced).

These patterns are fully in line with the Compositional Explanation. Campuses with high socioeconomic explanatory power—especially Merced, which in the late-2000s had socioeconomic explanatory power in the 30-40 percent range—have grown, in enrollees but especially in applicants (who wouldn't otherwise have applied to UC campuses). The end of affirmative action pushed socioeconomics' explanatory power down, especially at the Berkeley and UCLA campuses where that program was most effective, and the end of the 2001-2011 ELC program appears to have compressed demographics' explanatory power at the campuses where that program was most impactful (San Diego, Davis, and Irvine).[9] Meanwhile, all of the campuses were growing more selective on average, which compressed their HSGPA distributions, but also instituting admissions programs that intentionally targeted students whose low test scores were explicitly offset by measures of socioeconomic disadvantage, mechanically strengthening the correlation between socioeconomic characteristics and the SAT among the applicants that those programs encouraged. These estimates are challenging to reconcile with the Bias Explanation, and strongly suggest that the net STT after the Real-Disparities Explanation can be completely explained by the Compositional Explanation.

A comparison between Figure 2 and Figure 5 presents an interesting statistical anomaly similar to Simpson's Paradox. The classic case of Simpson's Paradox describes 1973 graduate admissions to UC Berkeley: while women were shown to have a small admissions advantage relative to men on average across each of Berkeley's graduate programs, male applicants to *any* Berkeley graduate program were more likely than female applicants to be

[9] See Bleemer (2019).

admitted (Bickel, Hammel, and O'Connell, 1975). The paradox is resolved as a result of female applicants being more likely to apply to more-competitive departments; even a trend that holds in every subgroup that comprises an aggregate may be reversed in the aggregate. In this case, the aggregate UC applicant pool manifests a misleading trend–the STT–despite the trend's not hold at any of the nine UC campuses that comprise the system, resulting from changes in the number of students applying to each campus over time and the various admissions policies implemented at different times by the various campuses.

5 Conclusion

While the Socioeconomic Testing Trend–the increasing proportion of University of California applicants' SAT score variation that can be explained by their ethnicity, family income, and parental education–appears disturbing, in fact it reflects both troubling and praise-worthy changes in California's 21st century education system. On the one hand, one-third of the Trend is explained by increasing socioeconomic and testing stratification across California high schools, a troubling phenomenon to be tackled by K-12 educational policy changes.

But on the other hand, two-thirds of the Trend can be explained by compositional changes in the UC applicant pool that appear to mirror intentional UC admissions policy changes that target disadvantaged applicants. As the university increasingly targets lower-income and otherwise-disadvantaged applicants with lower SAT scores, the relationship between socioeconomic characteristics and SAT performance among UC applicants mechanically strengthens. Given these targeted students' success at as University of California enrollees (Bleemer, 2018), this suggests that there is much to celebrate in a paradoxical Trend that turns out to be far less troubling than it first appears.

References

Bickel, P. J., E. A. Hammel, and J. W. O'Connell. 1975. "Sex Bias in Graduate Admissions: Data from Berkeley." Science 187 (4175):398–404. URL Link.

Bleemer, Zachary. 2018. "Percent Plans and the Return to Postsecondary Selectivity." Manuscript URL Link.

———. 2019. "Diversity in University Admissions: Affirmative Action, Percent Plans, and Holistic Review." Manuscript URL Link.

Card, David and Alan Krueger. 2005. "Would the elimination of affirmative action affect highly qualified minority applicants? Evidence from California and Texas." Industrial & Labor Relations Review 58 (3):416–434. URL Link.

Card, David and Jesse Rothstein. 2007. "Racial segregation and the black–white test score gap." Journal of Public Economics 91 (11-12):2158–2184. URL Link.

Crouse, James and Dale Trusheim. 1988. The Case Against the SAT. Chicago: The University of Chicago Press.

Douglass, John Aubrey and Zachary Bleemer. 2018. Approaching a Tipping Point? A History and Prospectus of Funding for the University of California. Berkeley, CA: Center for Studies in Higher Education. URL Link.

Geiser, Saul. 2015. "The Growing Correlation between Race and SAT Scores: New Findings from California." CSHE Research and Occasional Paper Series 15 (10). URL Link.

———. 2016. "A Proposal to Eliminate the SAT in Berkeley Admissions." CSHE Research and Occasional Paper Series 16 (4). URL Link.

Owens, Ann. 2018. "Income Segregation between School Districts and Inequality in Students' Achievement." Sociology of Education 91 (1):1–27. URL Link.

Owens, Ann, Sean F. Reardon, and Christopher Jencks. 2016. "Income Segregation Between Schools and School Districts." American Educational Research Journal 53 (4):1159–1197. URL Link.

Reardon, Sean F., Demetra Kalogrides, and Kenneth Shores. 2016. "The Geography of Racial/Ethnic Test Score Gaps." American Journal of Sociology 124 (4):1164–1221. URL Link.

Robinson, Michael and James Monks. 2005. "Making SAT scores optional in selective college admissions: a case study." Economics of Education Review 24 (4):393–405. URL Link.

Rothstein, Jesse. 2004. "College performance predictions and the SAT." Journal of Econometrics 121 (1-2):297–317. URL Link.

Simpson, Edward H. 1951. "The Interpretation of Interaction in Contingency Tables." Journal of the Royal Statistical Society, Series B 13:238–241. URL Link.

University of California. 2020. Relationship of the SAT/ACT to College Preparation and Performance at the University of California. Oakland, CA: UC Office of the President. URL Link.

Vigdor, Jacob and Jens Ludwig. 2007. "Segregation and the Black-White Test Score Gap." NBER Working Paper 12988. URL Link.

A Appendix A: Estimation Differences from Geiser (2015)

A number of modeling assumptions are necessary in the production of Figure 1, and my choices differ from those made in Geiser (2015) in several small ways. First, consider the three demographic characteristics analyzed in this brief:

1. Parental Income: Geiser (2015) includes only log CPI-adjusted parental income as his measure of income. This technique implicitly drops two important groups of applicants from the sample: (a) applicants who report 0 parental income, since the log of 0 is nonfinite (about 4 percent of the sample), and (b) applicants who do not report parental income on their applications, usually because they do not intend to receive financial aid, indicating high-income households (about 12 percent of the sample). Omitting these samples may mechanically decrease the correlation between income and SAT, since they represent the two extremes of income where the covariance with test scores may be highest. In order to maintain these samples, this analysis includes three measures of parental income in each regression model: log CPI-adjusted parental income (replaced as 0 when missing or infinite), an indicator for missing income, and an indicator for zero income. This change likely explains the higher proportion of SAT variation explained by the presented estimates.
2. Parental education: It appears that Geiser (2015) included an ordered integer measuring the more-educated parent's highest level of education. This measure simplifies a high-dimensional student feature—the educational level of their parent—into a highly-parametric summary. This analysis includes indicator variables for every possible combination of educational background held by the applicant's parents, using the full available information set. This includes every combination of parent 1's educational background (nine codes), parent 2's educational background, and whether the parent's highest level of education occurred in California. The result includes a total of 576 codes, each of which is included as a separate indicator variable. This change may partly account for the aggregate increase in explanatory power of demographics for applicants' SAT score.

3. Ethnicity: Geiser (2015) includes only an indicator for whether the applicant is from an underrepresented group, including Black, Chicano/a, Latino/a, or Native American. This analysis includes indicators for every observed ethnicity, or 15 in all. This may also contribute to the general increase in demographics' explanatory power for SAT scores.

The added value of including these multi-dimensional measures of students' background characteristics is that they more fully specify each student's background, leading to more explanatory power and avoiding possibly-important model restrictions that could challenge interpretation (especially in the case of parental income). The disadvantage of using multi-dimensional measures is that there is no longer a single standardized regression coefficient associated with each measure, making it impossible to directly compare the degree to which each contributes to their mutual absorption of SAT variation. As a result, rather than presenting regression coefficients, I show the degree to which each individual characteristic (as measured multi-dimensionally) alone can explain variation in applicants' SAT scores.