Who Wins and Who loses Under Common Educational Approaches?

By

KATHERINE ANKETELL KRAMER
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Economics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

_____

Scott Carrell, Chair

_____

Michal Kurlaender

_____

Marianne Bitler

Committee in Charge

2020

Contents

Abstract

Chapter 1: I quantify the predictive power of three common assessments, high school grades, SAT scores, and 11th grade achievement test scores on first-year college outcomes using both traditional validity models and newer statistical learning methods. I use a novel dataset that allows me to track California public high school 11th grade students into their first year either at a University of California or California State University campus. I also create back-of-the-envelope calculations to give a sense of how the assessments used in the college admissions process affect which students are admitted. I find that all three common assessments are similar in their modest levels of predictive power. However, I find large differences in which students are admitted to college when different sets of assessments are chosen.

Keywords: Education; Education Attainment; College; Equality of Opportunity; Education Policy.

JEL Classification Numbers: I21, I23, I24, I28, J15, J18.

Chapter 2: I use a novel dataset that allows me to track five cohorts of first-time freshmen for six years in order to to quantify the predictive power of four assessments, high school grades, class rank, SAT scores, previous student performance of students from a given high school, on two long term college outcomes, degree attainment and final cumulative GPA. I quantify the predictive power of these assessments using statistical learning methods. I also create back-of-the-envelope calculations to give a sense of how the assessments used in the college admissions process affect which students are admitted. I find that all four common assessments are similar in their levels of predictive power. However, I find large differences in which students are admitted to college when different sets of assessments are chosen.

Keywords: Education; Education Attainment; College; Equality of Opportunity; Education Policy.

JEL Classification Numbers: I21, I23, I24, I28, J15, J18.

Chapter 3: In this paper we exploit a unique natural experiment to estimate the impacts of school tracking on test score achievement and student behavior. Our data comes from a school district that places 4th grade students whose score on a standardized test of cognitive ability falls in the top 10% of the national distribution into self-contained "gifted" classrooms. We identify a causal relationship between academic track and these students' achievement by exploiting a discontinuity in entrance exam scores, and find that there is no discernable effect of being placed in the high ability classroom. We find some evidence that students who do not qualify for the GATE program have lower math test scores than they would in the absence of the program.

Keywords: Education; Equality of Opportunity; Education Policy.

JEL Classification Numbers: I21, I23, I24, I28, J15, J18.

Chapter 1

Predicting Short Term College Outcomes

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. This work was done as part of a larger research partnership between the California Department of Education and UC Davis (Michal Kurlaender, PI). I thank the California Department of Education, the California State University Chancellor's Office, and the University of California Office of the President for providing data access and expertise. I also thank researchers at the College Board for their assistance with restricted range adjustments, and Policy Analysis for California Education for dissemination of earlier drafts of this work. The author is grateful to Michal Kurlaender, Scott Carrell, Marianne Bitler, Marianne Page, Paco Martorell, Shu Shen, Sherrie Reed, and all the members of the California Education Lab for advice and suggestions. I would like to thank seminar participants at the University of California, Davis and the UC Office of the President for their helpful comments and insight. The opinions expressed are those of the author alone and do not represent the views of the Institute or the U.S. Department of Education, or of the agencies providing data.

1.1 Introduction

In June of 2018, the University of Chicago announced it would no longer require either the SAT or the ACT for admission. In doing so, it become the most high-profile example of a growing trend to deemphasize the use of standardized tests in the college admissions process (Anderson, 2018). That same summer California's Legislature passed A.B. 1951. A.B. 1951 would have allowed school districts to substitute the SAT for the state's 11th grade achievement tests. If it had been enacted, California would have joined 12 other states that use either the SAT or ACT to satisfy federal accountability measures. Instead, Governor Brown vetoed the bill and recommended that the University of California (UC) and the California State Universities (CSU) investigate using the California's 11th grade accountability tests in place of the SAT or ACT for admission to California's public universities (AB-1951; Gewertz, 2018).

These conflicting trends are the result of admissions offices and policy makers struggling with two questions. The first is, what predicts college success? The relationship between common assessments, like high school grades and standardized tests, and college performance is not as well understood as is often believed. Current methods used to measure the relationship between assessments and outcomes use in-sample methods to make out-of-sample predictions. This can lead to misleading results. Further, the statistics most often used to quantify these relationships, correlation coefficients, are opaque and difficult to interpret. As researchers at the College Board stated:

> Educational researchers, including the authors of this chapter, should strive to do a better job of communicating validity results in a more effective way. Correlation coefficients are probably not the best choice. Even those with doctorates in educational psychology or measurement have a hard time communicating what a correlation of a specific magnitude means beyond being small, moderate, or large. (Mattern et al, 2009)

The second question both admissions offices and policy makers would like an answer to is: "How does requiring different assessments during the application process affect who attends

college?" Choosing one combination of assessments over another in admissions will directly affect which applicants are accepted to college. But assessments also affect who applies to college. It is easier to acquire some assessments than others. For example, virtually all high school students have a GPA whether they want one or not, whereas most students have to select into taking the SAT or ACT. Many otherwise qualified low-income students do not take the SAT or ACT and therefore never join the college applicant pool. (For an overview of the research, see Dynarski, 2018.) Therefore, choosing to use an assessment in the college admissions process should take into consideration both the assessment's predictive power and the effect the assessment will have on which applicants are admitted.

This paper advances the conversation about the use of assessments in the college admissions process in five key ways. First, I have access to a novel dataset due to a unique partnership between researchers at UC Davis, the California Department of Education (CDE), the CSU Chancellor's Office, and the UC Office of the President and funded by the U.S. Department of Education Institute of Education Sciences. The data allows me to follow 11th grade California public high school students during the 2014-2015 academic year to either the UCs or the CSUs. Second, because of this partnership, I have access to assessments that are normally unavailable to the researcher. Specifically, I can compare the efficacy of California's 11th grade achievement test scores in predicting college outcomes to both high school grades and SAT scores. Third, I reexamine the limitations of predictive validity models and propose new estimators based on statistical learning methods that both address the tendency of previous models to produce distorted results and lack of interpretability of the results produced. Fourth, I use these new methods to show how much uncertainty remains in any prediction of college outcomes even after common explanatory factors are controlled for. Fifth, I create back-of-the-envelope calculations to give a sense of how the assessments used in the college admissions process affect which students are admitted.

The rest of this paper proceeds as follows. Section 2 describes the data. Section 3 begins by performs a validity study, the most common methodology used to judge the

worth of assessments in predicting college outcomes. This is followed by a discussion of the shortcomings of this methodology. Section 4 models college outcomes using a loss function (root means squared error) that compares actual college outcomes to predicted college outcomes within a statistical learning framework. Section 5 creates back-of-the-envelope calculations to estimate how using different sets of assessments affect the acceptance pool at the UCs and the CSUs. Specific methodologies are included at the beginning of section 3 through Section 5. Section 6 discusses the policy implications of my findings and Section 7 concludes.

1.2 Data

My sample consists of all California public high school students who took the 11th grade Smarter Balanced Summative Assessments (SBAC) in 2014-2015, whom I was able to match to a UC or CSU application for the Fall of 2016, and whose application contained both their high school GPA and SAT scores. Students who did not take the SAT are excluded from this analysis.[1] The SBACs are an annual statewide assessment administered to most California high school students in the 3-8 and 11th grades.[2] They are a criterion-referenced test designed to measure a student's progress towards college and career readiness under the Common Core standards.[3] The 2014-2015 school year was the first year that California public schools administered the SBACs. The SAT is a norm-referenced exam designed to predict college outcomes.[4] The SAT was recently redesigned and the new exam was first offered for testing in March 2016. Hence, this is the last cohort of California public high school students to be admitted to college exclusively using the old SAT format. Further research will be needed to confirm that the results of this study persist under the new SAT format, but preliminary investigations give no indication that the new SAT format would

[1] 15% of UC applicants do not take the SAT while 20% of CSU applicants do not take the SAT.

[2] Students who either take part in alternative assessments or who are in their first 12 months of school in the United States do not have to participate

[3] For more information, see: https://www.cde.ca.gov/ta/tg/sa/sbacsummative.asp

[4] For more information, see: https://research.collegeboard.org/programs/sat/data/validity-studies

change the results reported here in any substantive way.

Currently California does not have a system that tracks K-12 students into college. To create my sample, I matched students from the California Department of Education SBAC files to students in the UC and CSU application and enrollment files using name, date of birth, gender, and high school. I was able to match 85% of the first-time freshman in the UC files and 83% of the first-time freshmen coming from California public high schools in the CSU files to SBAC takers. For more information on match methodology and the match rate, please see Appendix 1. The SBAC files contain both ELA and math SBAC scale scores, along with demographic information including gender, race, and whether or not a student was socioeconomically disadvantaged (SED) at the time of the test.[5] The college level files contain the student's high school GPA, SAT scores, first-year college GPA, and whether the student persisted to the second year. A drawback of these data are that I can only observe first-year GPA if the student persisted to the second year, as the fall term of the second year is when first-year GPA is reported. Summary statistics are contained in Table 1.1.

Table 1.1: Summary Statistics

	UC		CSU	
	Applicants	Enrollees	Applicants	Enrollees
Female	.5792	.5838	.5815	.5876
	[.4937]	[.4929]	[.4933]	[.4923]
Asian/P.I.	.2908	.37	.1972	.183
	[.4541]	[.4828]	[.3979]	[.3867]
Black	.0435	.03	.0541	.0508
	[.2039]	[.1705]	[.2262]	[.2195]
Hispanic	.3969	.3518	.4881	.501
	[.4893]	[.4775]	[.4999]	[.5]
White	.2347	.216	.2293	.2352
	[.4238]	[.4115]	[.4204]	[.4241]
SED	.4699	.4611	.536	.5406
	[.4991]	[.4985]	[.4987]	[.4984]
HS GPA	3.453	3.64	338	339.1
	[.4048]	[.3009]	[55.16]	[43.84]
SAT Verbal	542.6	575.8	492.7	483.5
	[109.5]	[100]	[103.3]	[90.64]
SAT Math	561.8	600.1	505.6	496.4
	[115.6]	[105.1]	[109.9]	[95.62]
SAT Writing	542.3	578.2	488.3	476.9
	[110.8]	[102.3]	[102.2]	[86.88]
SBAC ELA	2686	2709	2651	2649
	[77.76]	[67.54]	[83.38]	[76.83]
SBAC Math	2681	2714	2630	2625
	[100]	[89.02]	[101.9]	[92.16]
1st Year GPA		3.09		2.973
		[.5906]		[.6138]
Persist to Year 2		.9271		.8382
		[.26]		[.3683]
N	73,804	28,544	121,606	45,002

Standard deviations in square brackets.

[5]Socioeconomically disadvantaged students either qualify for free or reduced-price lunch or do not have a parent who has graduated from high school.

1.3 Validity Study

1.3.1 Methodology

The earliest research examining the relationship of high school assessments to college outcomes are referred to as "validity studies". Validity studies grew out of the College Board's need to demonstrate the "validity" of their test, the SAT (see Kobrin et al., 2008, Mattern et al., 2008, Shaw et al., 2016, and Kurlaender and Cohen, 2019, for recent examples). A standard validity study compares the in-sample predictive strength of models without SAT scores to models with SAT scores. The difference in predictive strength between these models is called the "incremental validity". The predictive strength of a model is judged by the correlation, R, between the students' predicted outcomes and their actual outcomes. R is sometimes called a "multiple correlation coefficient" as there are often multiple predictors in a given model. More formally, R is defined as:

$$R = \sqrt{R^2} = (P_{yx} P_{xx}^{-1} P_{xy})^{\frac{1}{2}} \tag{1.1}$$

where P_{yx}, P_{xx}, and P_{xy} are partitions of the correlation matrix P such that:

$$P = \begin{bmatrix} P_{xx} & P_{xy} \\ P_{yx} & P_{yy} \end{bmatrix} \tag{1.2}$$

where x is the set of explanatory variables in the model and y is the outcome being predicted. It can be shown that R is also the square root of the R^2 of a regression of y on x. A detailed explanation of why using R this purpose is problematic is included in section 3.3 "Criticisms."

Unfortunately, the researcher can only observe outcomes for students who were admitted and chose to attend a given school. Because SAT scores, SBAC scores, and high school grades are used for admissions either directly (as in the case of SAT scores and high school grades) or indirectly (as in the case of SBAC scores), the correlations estimated using this selected sample will underestimate the strength of the relationship between the predictors

and the outcomes. To develop intuition, consider the relationship between the age of primary school children and their height. If we were to sample children from an elementary school, we would expect the age of a child and their height to be highly correlated. Fifth graders are invariably taller than kindergarteners. However, within each grade we would expect that age and height would be much less highly correlated. It is likely that some of the younger children in a grade would be taller than some of the older children in a grade. As age is used to select students into a grade level, observing the heights of the children within a grade restricts the range of the age of the children under observation. This restricted range then biases the observed correlation towards zero. As a reminder, R being biased due to restriction of range issues is a separate concern than regression coefficients being biased due to selection on x. In the regression context, selection on x only biases regression coefficients if x is correlated with unobserved factors that affect the outcome in question. In contrast, selection on x will bias R even when the explanatory variables included in a model are uncorrelated with unobserved factors.

To ameliorate the bias due to this restriction of range, validity studies traditionally make a restriction of range correction as first proposed by Pearson (1903); developed in Lawley (1943); and then outlined in Gulliksen (1950), Lord and Novick (1968), and Lewis (2006). Following Lewis's notation, let x and y be vectors of random variables with covariance matrix Σ for the full population.[6] Partition Σ such that:

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \tag{1.3}$$

Since y is not observed for the full population, Σ_{xy}, Σ_{yx}, and Σ_{yy} cannot be estimated directly from the data. However, if we let s be a selection variable such that $s_i = 1$ if an individual from the full population is included in the selected population while $s_i = 0$ if the individual is not included, the covariance matrix Σ for the selected population is:

[6]Due to data considerations, I take the population of applicants to the CSU or UC as my full population.

$$\Sigma_s = \begin{bmatrix} \Sigma_{xx|s} & \Sigma_{xy|s} \\ \Sigma_{yx|s} & \Sigma_{yy|s} \end{bmatrix} \tag{1.4}$$

which we estimate using our selected sample.

Under the assumptions of linearity and homoscedasticity it can be shown that:

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xx}\Sigma_{xx|s}^{-1}\Sigma_{xy|s} \\ \Sigma_{yx|s}\Sigma_{xx|s}^{-1}\Sigma_{xx} & \Sigma_{yy|s} - \Sigma_{yx|s}(\Sigma_{xx|s}^{-1} - \Sigma_{xx|s}^{-1}\Sigma_{xx}\Sigma_{xx|s}^{-1})\Sigma_{xy|s} \end{bmatrix} \tag{1.5}$$

In effect, this correction reduces the observed sample variance of the outcome to what we would theoretically see if we were able to observe outcomes for the full population. It is traditional to use the full set of explanatory variables in Σ_s of the most fully specified model. Then in order to calculate the correlation matrix P, where:

$$P = (diag\,(\Sigma))^{-\frac{1}{2}}\,(diag\,(\Sigma))^{-\frac{1}{2}} \tag{1.6}$$

I delete the rows and columns from Σ which contain the variances and covariances for explanatory variables not included in the model being estimated. For additional derivations and further discussion, see Lawley (1943), Lord & Novick (1968), and Lewis (2006).

1.3.2 Results

Tables 2 through 5 present the results for a validity study performed on the UC and CSU systems using my sample. Tables 2 and 3 present results for models that predict whether a student will persist to the second year at the UCs and CSUs respectively. Tables 4 and 5 present results for models that predict a student's first-year GPA. In all four tables, I have bolded the columns that report the most policy relevant models, i.e. the models that include only high school GPA as an assessment, the models that includes high school GPA and SAT scores, and the models that includes high school GPA and SBAC scores.

Correlation coefficients with the restriction of range correction are reported first followed by raw correlation coefficients in square brackets.

Table 1.2: Multiple Correlation Coefficient, Persistence to Second Year (University of California)

	(1)	(2)	(3)	**(4)**	**(5)**	(6)
Included: campus fixed effects.						
Multiple Corr. Coeff.	**.1748 [.1343]**	.1965 [.1654]	.1894 [.1587]	**.2171 [.1802]**	**.2120 [.1735]**	.2198 [.1831]
Included: campus fixed effects, gender, SED, and race.						
Multiple Corr. Coeff.	**.1986 [.1647]**	.2050 [.1745]	.2004 [.1716]	**.2228 [.1873]**	**.2219 [.1843]**	.2268 [.1899]
Included Covariates:						
HSGPA	**X**			**X**	**X**	X
SAT		X		**X**		X
SBAC			X		**X**	X

Note: Coefficience corrected for restriction of range [uncorrected coefficient].

Table 1.3: Multiple Correlation Coefficient, Persistence to Second Year (California State University)

	(7)	(8)	(9)	**(10)**	**(11)**	(12)
Included: campus fixed effects.						
Multiple Corr. Coeff.	**.2262 [.1844]**	.1919 [.1488]	.1978 [.1599]	**.2383 [.1956]**	**.2406 [.1991]**	.2418 [.1999]
Included: campus fixed effects, gender, SED, and race.						
Multiple Corr. Coeff.	**.2397 [.1976]**	.2070 [.1662]	.2131 [.1755]	**.2461 [.2045]**	**.2526 [.2078]**	.2530 [.2082]
Included Covariates:						
HSGPA	**X**			**X**	**X**	X
SAT		X		**X**		X
SBAC			X		**X**	X

Note: Coefficience corrected for restriction of range [uncorrected coefficient].

Table 2 reports correlation coefficients of models that predict persistence to the second year of college for students at the UCs. The top row reports the correlation between the predicted results and the actual outcomes for models that do not include demographics while the second row includes demographics as controls. The cell in the top row, first column containing a correlation of .1748 indicates that the correlation between predicted outcomes and actual outcomes for a model that does not include demographics and only uses high school grades to assess college readiness is .1748. When using only SAT scores in a model without demographics, actual and predicted outcomes have correlation of .1965 as can be seen in the top row of column (2).[7] The third column reports the results when SBAC

[7]The SAT scores are entered separately in all models. Without the restriction of range adjustment, this

scores are the only assessments included in the model.[8] In this case the correlation between predicted and actual outcomes is .1894. The fourth and fifth column report the results for models that include HSGPA and SAT scores and HSGPA and SBAC scores respectively with corresponding correlation coefficients of .2171 and .2120 for models that do not include demographics. This leads to an incremental validity of .0423 for SAT scores above high school grades alone and .0372 for SBAC scores. Finally, the correlation between predicted and actual outcomes in models that include all three assessments is .2198. The take home from these models is that both SAT scores and SBAC scores increase the predictive accuracy of a model above grades alone. Whether these increases are substantial or trivial is hard to judge. When demographics are added to these models, the same patterns hold as above, though the correlation coefficients increase slightly and the incremental validities for SAT scores and SBAC scores shrink to .0242 and .0129 respectively. Interestingly, the incremental validity for SAT scores shrinks more when demographics are added to the model than the incremental validity for SBAC scores. Though again, it is hard to judge whether these differences are meaningful.

Table 3 reports results from the same models predicting persistence to year two as in Table 2, but here they are estimated on the sample of CSU students. In models without demographics that only use high school grades as an assessment, the correlation between actual and predicted outcomes is .2262. Adding SAT scores or SBAC scores to the model increases the correlation to .2383 and .2405 respectively. Doing so results in incremental validities of .0121 for SAT scores and .0144 for SBAC scores. When demographics are added to the models, the incremental validities shrink to .0064 and .0129 for SAT and SBAC scores respectively.

Table 4 reports results from models that predict first-year GPA at the UCs. The first thing a reader might notice is that these correlations are much higher than the correlations

would be analogous to regressing an outcome on a student's SAT verbal score, math score, and writing score.

[8]Similarly to SAT scores, SBAC scores are entered separately in all models. Without the restriction of range adjustment, this would be analogous to regressing an outcome on a student's SBAC ELA score and math score.

Table 1.4: Multiple Correlation Coefficient, First-Year GPA (University of California)

	(13)	(14)	(15)	**(16)**	**(17)**	(18)
Included: campus fixed effects.						
Multiple Corr. Coeff.	**.4742 [.3410]**	.5748 [.4975]	.5117 [.4218]	**.6186 [.5299]**	**.5738 [.4648]**	.6203 [.5321]
Included: campus fixed effects, gender, SED, and race.						
Multiple Corr. Coeff.	**.5559 [.4594]**	.5864 [.5110]	.5516 [.4754]	**.6260 [.5392]**	**.6037 [.5079]**	.6292 [.5420]
Included Covariates:						
HSGPA	**X**			**X**	**X**	X
SAT		X		**X**		X
SBAC			X		**X**	X

Note: Coefficience corrected for restriction of range [uncorrected coefficient].

Table 1.5: Multiple Correlation Coefficient, First-Year GPA (California State University)

	(19)	(20)	(21)	**(22)**	**(23)**	(24)
Included: campus fixed effects.						
Multiple Corr. Coeff.	**.4601 [.3804]**	.3680 [.2776]	.3572 [.2734]	**.4786 [.4043]**	**.4747 [.3990]**	.4815 [.4077]
Included: campus fixed effects, gender, SED, and race.						
Multiple Corr. Coeff.	**.4832 [.4087]**	.4035 [.3218]	.4005 [.3251]	**.4927 [.4216]**	**.4976 [.4204]**	.5011 [.4246]
Included Covariates:						
HSGPA	**X**			**X**	**X**	X
SAT		X		**X**		X
SBAC			X		**X**	X

Note: Coefficience corrected for restriction of range [uncorrected coefficient].

from models predicting persistence to year two. This is to be expected as persistence to year two is a binary outcome whereas first-year GPA is continuous. Here the correlation between predicted and actual outcomes for models without demographics and with high school grades as the only assessment is .4742. The analogous model with high school grades and SAT scores has a correlation of .6186 while the model with high school grades and SBAC scores has a correlation of .5738. These results show an incremental validity of .1444 for SAT scores and .0996 for SBAC scores. When demographics are added to the models the incremental validities shrink to .0701 and .0478 for SAT and SBAC scores respectively. Table 5 reports the corresponding results for the CSUs. The incremental validities are .0185 and .0146 in models with SAT or SBAC scores respectively with no demographics while they shrink to .0095 and .0144 for SAT and SBAC scores in models with demographics.

There are a few patterns to point out. The first is that in models that only use a single

assessment, the best predictor of outcomes at the UCs is SAT scores. This is in contrast to the CSUs where the best single predictor of outcomes is high school grades. It might be tempting to attribute this difference to a ceiling effect for grades at the UCs, but there is no evidence that this is the case. Instead, I would point the reader to the relative size of the standard deviation of grades and SAT scores in each system. High school grades have a larger standard deviation in the sample of CSU students while SAT scores have a larger standard deviation in the sample of UC students. The relative strength of high school grades and SAT scores as predictors may well be a mechanical relationship due to selection into each sample. At the UCs, SAT scores also have the highest incremental validity. At the CSUs, in three out of four models, SBAC scores have the highest incremental validity. At the same time, as best as it can be judged, all of these differences in validity, incremental or not, are relatively small.

1.3.3 Criticisms

The criticisms of validity studies can be grouped into three main categories. The most common is that the assumptions underlying the restriction of range correction do not hold. Since these assumptions do not hold, any estimate of an assessment's incremental validity made using this methodology will be biased in unpredictable directions (Rothstein 2002, 2004, & Black et al, 2016), This assertion, while correct, is less important than its frequency would indicate due to the other two criticisms being so noteworthy. First, it is not clear what standard measures of incremental validity are actually measuring. Is it actually some measure of ability or is it a proxy for demographics like race and socioeconomic status (Rothstein 2004, & Black et al, 2016)? Second, incremental validity as measured in a traditional validity study is virtually impossible to interpret. But these three criticisms are missing a much more fundamental problem. In any set of linear models where one fully contains the other, e.g.:

$$\begin{aligned} (A) \quad & \hat{y}_i = \alpha_0 + \alpha_1 x_{1i} \\ (B) \quad & \hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \end{aligned} \tag{1.7}$$

$corr_B(\hat{y}_i,\ y_i) \geq corr_A(\hat{y}_i,\ y_i)$ for in-sample estimation. Further, while it is theoretically possible for x_2 to contain no explanatory power and for $corr_B(\hat{y}_i,\ y_i) = corr_A(\hat{y}_i,\ y_i)$, even if x_2 is unrelated to y in actuality, adding x_2 would almost certainly increase the in-sample correlation coefficient due to some "explanatory" noise. So one would expect to always find $corr_B(\hat{y}_i,\ y_i) > corr_A(\hat{y}_i,\ y_i)$. This is equivalent to the more familiar reminder often given in introductory regression classes, "If you add an explanatory variable to your model, R^2 will increase." This means traditional validity tests are structured so a positive, non-zero incremental validity will always be found. Restriction of range corrections do not change this. Moving from a multiple correlation framework to a regression framework in order to ease significance testing does not solve the problem. Due to selection, OLS will produce biased and inconsistence estimates of all parameters in the model and tests of significance will be incorrect.

While this is a much more fundamental issue than the restriction of range correction being incorrect, it is worth spending time reviewing why the restriction of range correction produces biased results. To begin with, the assumptions of linearity and homoscedasticity that underlie the correction are not as innocuous as the reader might assume. With OLS, we do not need homoscedasticity to hold to produce unbiased and/or consistent parameter estimates. And with Eicker-Huber-White standard errors, bootstrap methods, and randomization inference, today's practitioner often does not spend overly much time worrying about correcting their standard errors for significance testing. Nonlinearity is addressed using nonlinear transformations and by stating that all continuous correspondences in a small enough range can be approximated by linear functions.

Unfortunately, the assumptions of linearity and homoscedasticity are not innocuous in the context of a restriction of range correction. In order to solve for the for the unrestricted

correlation matrix, we have to equate the parameters of the regression of y on x from the restricted population and from the unrestricted population. It is the assumptions of linearity and homoscedasticity that allow us to do so. Without these two assumptions, the correlation matrix of the unrestricted population cannot be derived.

Yet for many of our models we know that the assumption of homoscedasticity will not hold. All linear probability models are heteroscedastic. As such, the assumption of homoscedasticity will not hold for models that predict persistence to the second year. Further, as first-year GPA is constrained between zero and four, heterosedasticity would be a reasonable expectation for models predicting first-year GPA. And in fact, the non-constant variance of the error term can be easily observed in Figure 1.3.

Further, as Rothstein (2002, 2004) points out, estimates for Σ are only consistent when we would expect the slope parameters from a corresponding regression model to be consistent. However, assuming that we expect students to be selected into college based on both their HSGPA and SAT scores, we would expect any model without SAT scores to be biased and inconsistent. This means that all of our validity estimates for HSGPA are wrong. The intuition is as follows. Assume you have two groups of students, A students and C students, who apply to college. The A students will be admitted given a wide range of SAT scores. As such, they are generally representative of the SAT scores for A students in the general population. In contrast, C students will only be admitted if they have high SAT scores. This will mean the sample of C students will have higher expected SAT scores than C students in the general population. As Rothstein shows, this means estimates of the validity of HSGPA will be smaller than its true validity. Since our incremental validities are calculated by subtracting the validity of HSGPA from the validities of models that contain SAT/SBAC scores and HSGPA, this will cause us to overestimate our incremental validities.[9]

There is also a great deal of disagreement as to what the different assessments actually measure. Some argue that assessments are relatively clean measures of college readiness,

[9]For a more detailed and technical discussion see Rothstein (2002) and (2004).

often using validity studies as proof. Others point to the high correlation between assessments, especially SAT scores, and demographics like race and class. Rather than being an assessment of college readiness, the argument goes, these measures proxy for societal disadvantages students face. To investigate each argument's relative worth, Rothstein (2004) decomposes the predictive power of SAT scores and high school GPA into parts that can be explained by individual and school levels demographics (race, socioeconomic status, and gender) and what cannot be explained by demographics. He finds that student demographics do a much better job of predicting SAT scores than they do of predicting high school grades. Further when he breaks SAT scores into two parts, one part that can be explained by student and school characteristics and another part that those characteristics do not explain, the part of SAT scores that is explained by student and school characteristics explain a much larger proportion of first-year GPA than the part not determined by student characteristics. Since admissions offices are often prohibited from taking race into account, Rothstein (2004) points out that using an assessment that has its predictive power largely derived from race is a back-door way of including race in the college admissions process.

Finally, interpreting incremental validities is not a trivial task. When judging correlation coefficients, Cohen's (1988) rule of thumb is that a correlation coefficient of below .1 is trivial, .1-.3 is small, .3-.5 is moderate, and .5 and above is large.[10] "Small", "moderate", and "large" are not the most illuminating of categories. However, this would indicate the incremental validities found here for SAT and SBAC scores are at best "small" and are often "trivial" and do not represent substantive increases of predictive accuracy. Cohen's rule of thumb also suggests that the predictive power of SAT and SBAC scores do not differ in any meaningful way. Still, Cohen was discussing judging the size of individual correlations and not the difference between two correlations. At best, his framework serves as a rough guide.

[10]Cohen suggested many rules of thumb. These are the rules for correlation coefficients, not effect sizes.

1.4 Predictive Accuracy As Measured By The Root Mean Squared Error and Statistical Learning

1.4.1 Methodology

Any framework used to judge the predictive accuracy of a model should meet two criteria. It should produce results that are accurate and it should produce results that can be understood. Validity studies meet neither of these criteria. Validity frameworks are not accurate. They will always indicate that any additional factor added to a model will improve results. This is the case even if the additional factor worsens the ability of the model to predict outcomes. Validity frameworks also do not produce results that are easy to interpret. Explaining what a .05 change in a multiple correlation coefficient means is a non-trivial task. To solve the accuracy problem, I move from an in-sample framework to an out-of-sample framework, i.e. I move from judging a model's predictive power by comparing a model's predictions to outcomes used to estimate the model to comparing a model's predictions to outcomes that were not used to estimate the model. In an out-of-sample framework, increases in measures of fit mean that one can expect out-of-sample predictions to improve. If goodness of fit statistics worsen, then one would expect the out-of-sample predictions of the model to also worsen.

It is possible to calculate correlation coefficients using an out-of-sample framework. Measuring the correlation between a model's prediction of an outcome and outcomes of individuals not used to construct the model would do this. As I am already changing the framework, however, there is no reason not to move to a more intuitive measure of goodness of fit. To that end I propose using the root mean squared error (rMSE).

The root mean squared error (rMSE) is defined as the sum of the difference between the all predicted values and their corresponding actual values, squared. This sum is then divided by n and the square root is taken. Formally, the rMSE is defined as:

$$rMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{1.8}$$

This can be thought of as the standard deviation of the model's error, or the average distance between actual and predicted outcomes.

Figure 1.1 is an example that allows the reader to compare validity studies to this new framework based on statistical learning. Imagine there are four possible measures that could explain some outcome, y, like a test score. For simplicity, and in order to be able to graph this in two dimensions, let the four measures equal x, x^2, x^3, x^4. Now assume that we collect a sample of data that includes a student's score on all four measures and the outcome of interest. Using a statistical learning framework, this is called the training data because it is the data that is used to train a model to make predictions. I graph the training data using the first measure (x) and the outcome (y). These are the solid circles plotted on the graphs in the first column of Figure 1.1. Ordinary least squares is used to create (train) models to predict the outcome. These models are represented by the solid line on the graphs. As we move down the first column of graphs, the dots do not change because the same sample is used each time. However, the line changes because the values predicted by the model changes as more explanatory measures are added. Using the first measure as the only explanatory variable, the predicted values have are linear and correlation with the training sample of r = .7217. When we add a second predictive measure to the model, the in-sample predictions improve. We can see this visually as the predicted values now curve to follow the training data. Mathematically we can see this as the correlation coefficient increases to r = .8010. As we add measures three and four respectively, we can see that the curve increasingly wiggles and bends to produce a better fit to the training data. The correlation coefficients also continue to increase reflecting the improved fit

When it comes to assessments and college outcomes, we are not particularly interested in how well a model can predict outcomes that are already known. This would be akin to looking at the high school transcripts of a college sophomore to learn about their first-

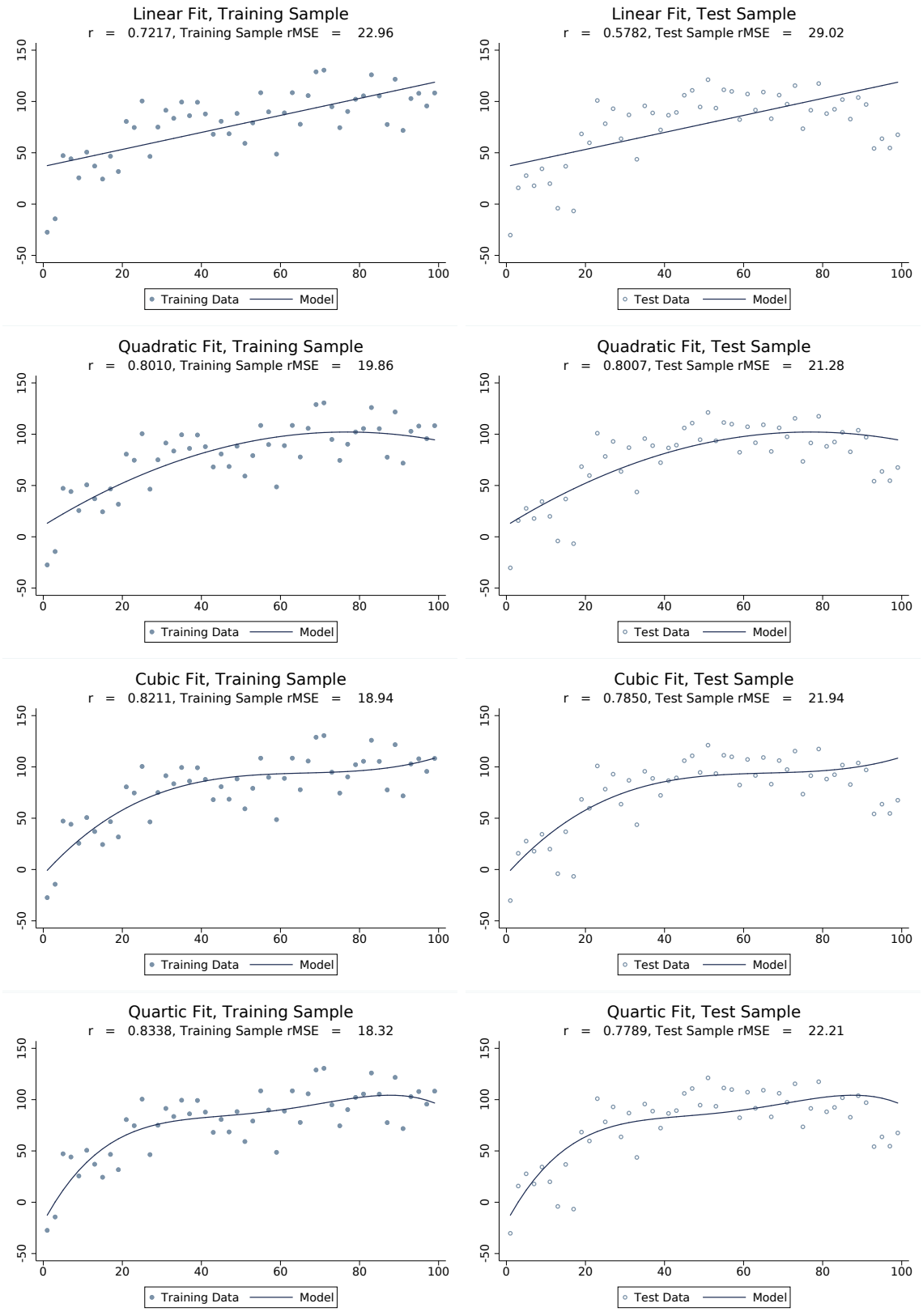


Figure 1.1: In-Sample vs. Out-of-Sample Prediction.

year college grades. Instead, we would like any model selected to do a good job predicting outcomes that we do not yet know. To learn if a model can do this, we have to test the model on data that was not used to train the model.

Looking across the top row of graphs, the blue dots on the graphs on the left represent the data used to construct (train) the model. The open circles on the graphs to the right represents data that was generated by the same process but that was not used to construct the model. Since we did not use this data to build the models, we can instead use it to test the model's ability to make out-of-sample predictions. As such, it is called the test data. The line on the graphs in the right column represents the models' predicted values, which are the same on the left and right in each row. Starting with the top right graph, when we compare our predicted values to the test data, we get a correlation coefficient of .5782. Just as before, this increases when a second explanatory measure is added to the model, which can be seen in the second row, both by visual inspection and because the correlation coefficient has now increased to .8007. When a third explanatory measure is added, however, the predicted values of the model do a worse job of predicting the test data. This is reflected with a decreased correlation coefficient of .7850. Adding a fourth measure further worsens our model fit and now the correlation between the predicted values and the test data is .7789.

While moving from judging a model using in-sample data to judging a model using out-of-sample data produces much more accurate judgements, by continuing to use correlation coefficients it is hard to say more than that one model predicts outcomes better or worse than another model. Ideally, we would like to understand the magnitude of these changes. Using the rMSE instead of correlation coefficients make it much easier to judge how much a model is improving (or worsening) as explanatory variables are added. Again, looking at the right column where model predictions are compared to test data, using one measure to predict outcomes leads to a rMSE is 29.02. If our outcome was a test score this would mean that the average distance between the actual outcomes in the test data and the outcomes our model would predict would be about 29 points. Adding a second explanatory measure causes

the rMSE to drops almost 8 points to 21.28. This is much easier to understand compared to changes in the correlations. Adding a third explanatory worsens model fit and the rMSE increases to 21.94. A fourth explanatory measure further decreases the expected explanatory power of the model further as indicated by the increase of the rMSE to 22.31.

It may seem counter intuitive that adding more explanatory variables to a model could worsen the model's predictive power. To see why this can happen, imagine adding variables to a model until there was one variable per observation. This model would perfectly predict the training data, but it would not do a particularly good job of predicting the test data. More technically, it can be shown that:

$$E[MSE_{out\,of\,sample}] = E[(y_i^{test} - \hat{y}_i)^2] = Var(\hat{y}_i) + [Bias(\hat{y}_i)]^2 + Var(\varepsilon) \tag{1.9}$$

where $Var(\hat{y}_i)$ is how much we would expect the predicted values of a model to change if a new sample was used, $[Bias(\hat{y}_i)]^2$ is how much the estimates are wrong due to the parameters of the model, and $Var(\varepsilon)$ is the unexplained random component.[11] Therefore, in order to minimize the out-of-sample prediction error we need a model that balances both precision (minimizes $[Bias(\hat{y}_i)]^2$) and robustness (minimizes $Var(\hat{y}_i)$).

I do not have two samples that would naturally lend themselves to separate training and test datasets. Instead, I use cross validation and randomly divide my data into k equal parts or folds. I reserve the first fold and train my model on the remaining k-1 folds. Then I calculate the rMSE using the reserved fold as the test data. I repeat this process on all k folds and average the resulting k rMSE. I.e.:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} rMSE_i \tag{1.10}$$

I choose k=5 and k=10, (CV(5) and CV(10)). I do this for both computational simplicity and because these values have been shown to have good statistical properties. See James et.

[11] For a less technical discussion see James et al. (2013). For a more technical discussion and proof, see Hastie et al. (2009).

al (2013) for further discussion.

Previous work has largely used a linear functional form to construct predictive models. In Figure 1.2, I plot average outcomes by high school GPA, SAT scores, and SBAC scores. In general, the relationships look largely linear, though occasionally there is some evidence of a quadratic. As such, I construct both linear and quadratic models.

1.4.2 Results

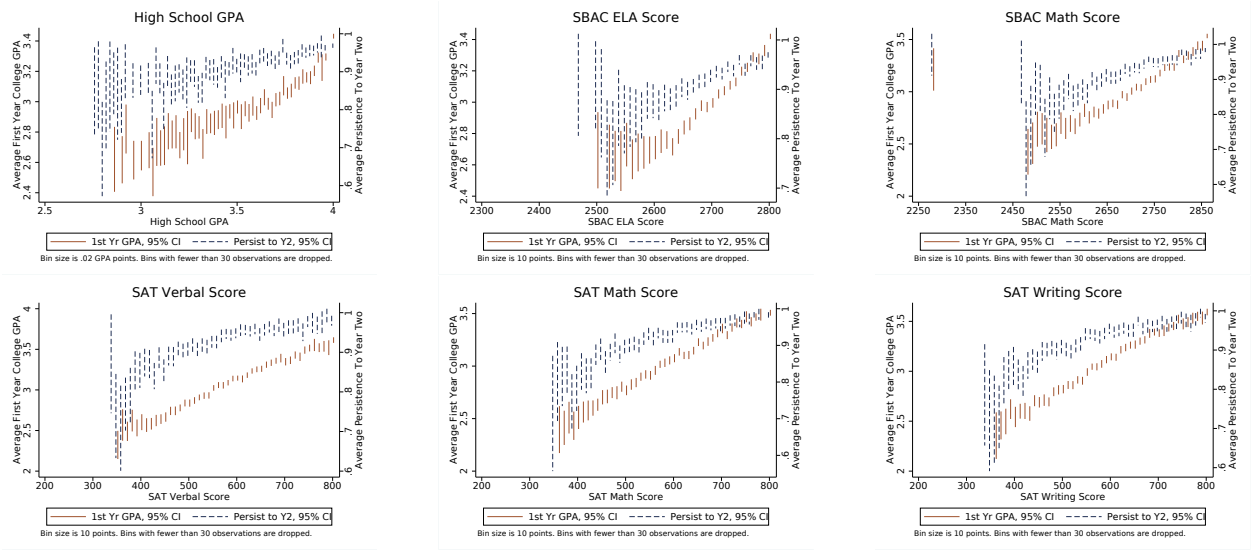
I model persistence to the second year eight different ways, using both ordinary least squares and logistic regressions, linear and quadratic functional forms, and CV(5) and CV(10). All eight methods produce virtually identical results. I only report linear models estimated with OLS and CV(5) for the sake of simplicity. The rMSE for the UCs and CSUs are reported in Table 1.6 and 7 respectively.

Table 1.6: rMSE, Persistence to Second Year (University of California)

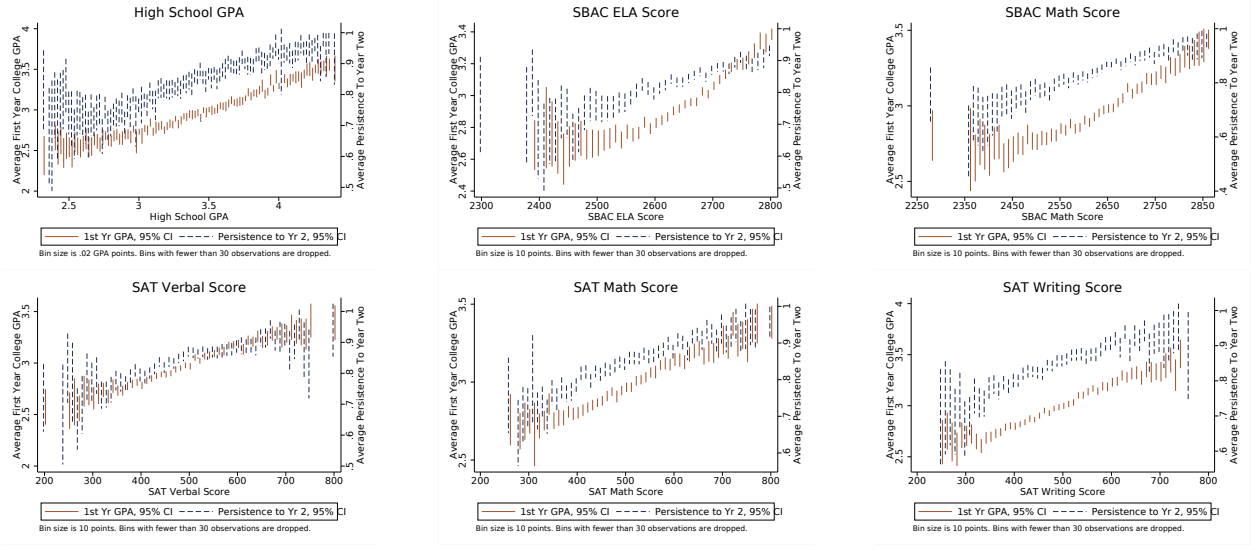
	(25)	**(26)**	(27)	(28)	**(29)**	**(30)**	(31)
Included: campus fixed effects.							
rMSE	.2456	**.2451**	.2441	.2441	**.2438**	**.2439**	.2437
Included: campus fixed effects, gender, SED, and race.							
rMSE	.2447	**.2444**	.2438	.2438	**.2436**	**.2436**	.2435
Included Covariates:							
HSGPA		**X**			**X**	**X**	X
SAT			X		**X**		X
SBAC				X		**X**	X

Note: rMSE calculated using ordinary least squares and five-fold cross validation.

Each cell in these tables contains a rMSE calculated by comparing a model's predicted persistence of a student not used to construct the model to the student's actual persistence. In Table 1.6, the top row represents models that contain campus fixed effects only. Therefore, the top cell of column (25) reports the results for a model that contains campus fixed effects and nothing else. We could think of this as predicting student outcomes using that sample average for each campus. Here the rMSE equals .2456. This is the standard deviation of the prediction error. It also means that on average using this model, our predictions are about 25 percentage points from a student's actual outcome of either persisting (100%) or failing to persist (0%) to the second year. When I add high school GPA to the model,



(a) UC.



(b) CSU

Figure 1.2: Average First-Year Outcome by Qualifying Measure.

as reported in the top cell in column (26), the rMSE becomes .2451. In other words, this improves the prediction of the model 0.05 percentage points over using the campus adjusted sample average. If instead I substitute either SAT scores or SBAC scores, the rMSE becomes .2441, an improvement of 0.15 percentage points over the sample average. Adding a test score to models that use high school grades to predict persistence also improves the models, though not by much. High school grades and SAT scores result in an rMSE of .2438 or a 0.18 percentage point improvement over the sample average and a 0.13 percentage point improvement over grades alone. Adding SBAC scores to grades instead of SAT scores produces a virtual identical result with an rMSE of .2439. It is standard to report models that also include demographics and I do so in the second row. Adding demographics to these models do not produce substantive decreases in the rMSEs.

There are three patterns to notice in Table 1.6. The first is that test scores are the best single predictor of persistence to the second year at a UC. Further, adding test scores to a model that already contains high school GPA improves the model's predictions and SAT scores add more additional information than SBAC scores. However, no model improves much on the campus-weighted sample average. The greatest reduction in rMSE is 0.21 percentage points. These differences are quite small. Another way to quantify the differences between models is to calculate the error rate for each model. To calculate the error rate, I run each model as described above. Once I have obtained a predicted probability of persistence for each student, I set the model to predicting the student as persisting to the second year if the predicted probability of persistence is greater than 50%. If a model predicts that a student has a 50% chance of persistence or less, then I set the model to predicting that the student will not persist. Then I count the number of times a student's actual outcome is different from the model's predicted outcome and divide by the number of students. More formally:

$$Error\ Rate = \frac{\sum_{i=1}^{n}(y_i \neq \hat{y}_i)}{n} \tag{1.11}$$

All 14 models have error rates of .0729. In other words, all models predict persistence incorrectly 7.29% of the time. No included assessment or set of demographics improves the prediction of a model over the campus-weighted average.

Table 1.7: rMSE, Persistence to Second Year (California State University)

	(32)	**(33)**	(34)	(35)	**(36)**	**(37)**	(38)
Included: campus fixed effects.							
rMSE	.3652	**.3614**	.3637	.3628	**.3608**	**.3605**	.3605
Included: campus fixed effects, gender, SED, and race.							
rMSE	.3639	**.3607**	.3627	.3620	**.3602**	**.3599**	.3599
Included Covariates:							
HSGPA		**X**			**X**	**X**	X
SAT			X		**X**		X
SBAC				X		**X**	X

Note: rMSE calculated using ordinary least squares and five-fold cross validation.

Table 1.7 reports results of similar models estimated on the sample of California State University Students. The top cell of column (33) reports the results from a model that predicts second year persistence using campus fixed effects and high school GPA. Here the rMSE is .3614 or a 0.38 percentage point improvement over the sample average. Adding SAT scores to this model improves the rMSE to .3608, a further improvement of 0.06 percentage points. If I substitute SBAC scores for SAT scores, the rMSE instead decreases to .3605, a 0.09 percentage point improvement. For CSU students, high school grades are the best single predictor of first-year persistence. Test scores improve the predictive power of the models, but by a lesser amount. Adding SAT scores to grades decreases the rMSE by 0.06 percentage points to .3608. If instead I add SBAC scores to high school grades, the rMSE decreases by 0.09 percentage points to .3605. Here, as opposed to in the UC sample, adding SBAC scores improves the model slightly more than adding SAT scores. Again, however, no model improves much over the sample average. Again, the error rates for all models is the same, 16.2%

I model first-year GPA using ordinary least squares four separate ways, with both linear and quadratic functional forms and CV(5) and CV(10). Again, all four methods produce virtually identical results. I only report linear models estimated with CV(5) for the sake of simplicity.[12] The rMSEs for UC and CSU models are reported in Table 1.8 and 9 respectively.

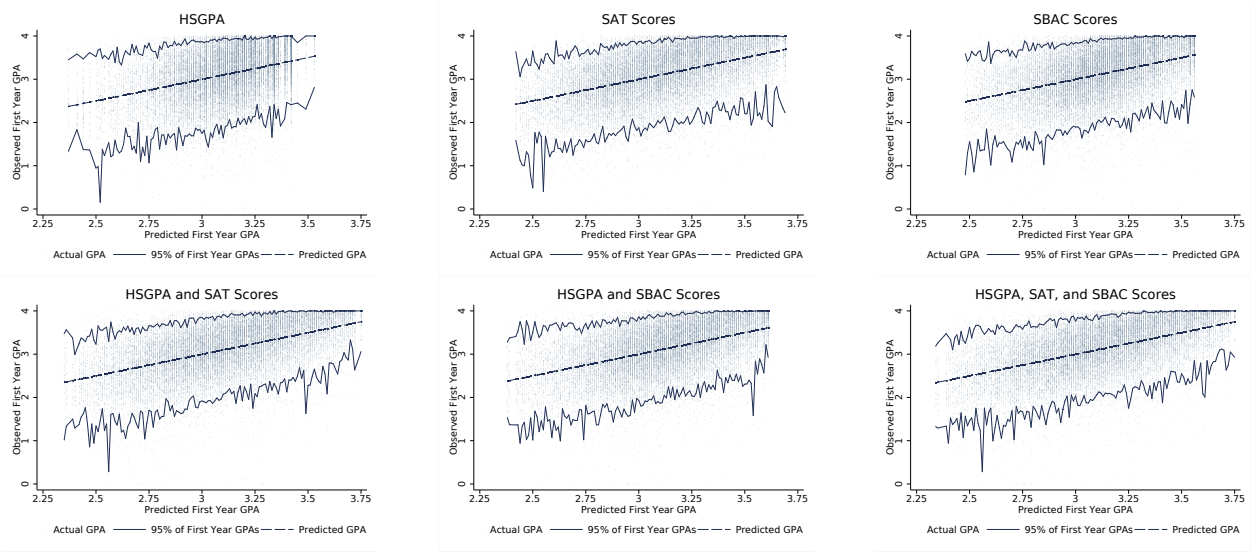
[12]All other results are available from the author by request.

Table 1.8: rMSE, First-Year GPA (University of California)

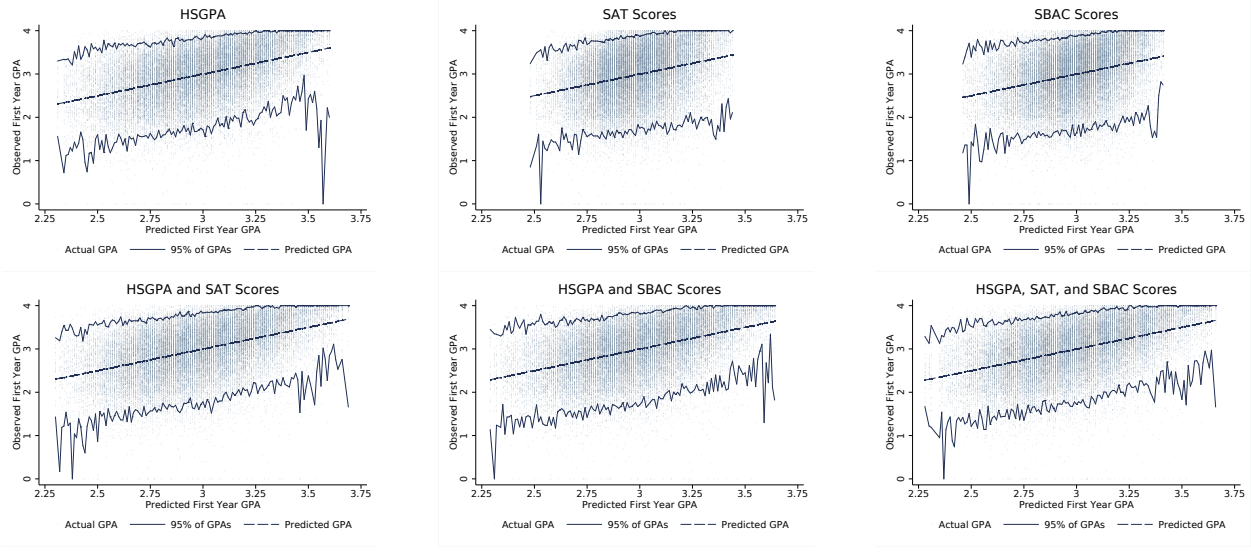
	(39)	**(40)**	(41)	(42)	**(43)**	**(44)**	(45)
Included: campus fixed effects.							
rMSE	.5703	**.5458**	.5103	.5300	**.4939**	**.5152**	.4931
Included: campus fixed effects, gender, SED, and race.							
rMSE	.5386	**.5207**	.5060	.5164	**.4916**	**.5042**	.4906
Included Covariates:							
HSGPA		**X**			**X**	**X**	X
SAT			X		**X**		X
SBAC				X		**X**	X

Note: rMSE calculated using ordinary least squares and five-fold cross validation.

Table 1.8, column (39), top row, reports the rMSE, .5703, for models that only use campus fixed effects to predict first-year GPA. If we think of the rMSE as the average distance between our predicted and actual values, then we would expect any prediction made using use campus fixed effects to be off by .57 GPA points (on a scale of zero to four) in either direction. Adding high school grades to our models reduces the expected error to .5458 or by .0245 of a grade point. Test scores reduce the expected error further by .0519 to .4939 grade points when SAT scores are added to grades and by .0306 to .5152 if SBAC scores to a model instead. As above, at the UCs, SAT scores are the best single predictor of first-year grades and SAT scores bring more predictive power to a model when included with high school grades than SBAC scores do. The differences seem small, there is only a .0213 grade point difference between the model with high school grades and SAT scores and the model with high school grades and SBAC scores. As first-year GPA is a continuous variable, I cannot construct an error rate as I did above in order to summarize the differences in the models. Instead, in Figure 1.3 I plot observed first- year GPA by the predicted first-year GPA of each model. The top six figures are models estimated using UC data, the bottom six are models estimated using the CSU data. On each graph, each dot represents one student. Each student's predicted first-year GPA is plotted on the x-axis while their actual first-year GPA is plotted on the y-axis. The dotted line represents what the graph would look like if the model predicted the GPA of all students perfectly. The darker solid lines are constructed by identifying the top and bottom 2.5% of students in each bin and connecting these points. This puts 95% of the data between these the two dark



(a) UC.



(b) CSU

Figure 1.3: Actual vs. Predicted First-Year GPAs

lines. The top left graph represents predictive models estimated using campus fixed effects and high school grades. We can see that for almost all bins, 95% of the data spans at least two GPA points. In other words, if I were to predict a student's first-year GPA to be a 3.0, its likely range would be between 2.0 and a 4.0. That is a very wide range. Top center and right graphs represent models that use SAT scores or SBAC scores, respectively, as their predictive assessment. Again, for most bins the data spans at least two GPA points.

Table 1.9: rMSE, First-Year GPA
(California State University)

	(46)	**(47)**	(48)	(49)	**(50)**	**(51)**	(52)
Included: campus fixed effects.							
rMSE	.6076	**.5550**	.5865	.5832	**.5447**	**.5463**	.5434
Included: campus fixed effects, gender, SED, and race.							
rMSE	.5912	**.5454**	.5739	.5716	**.5380**	**.5389**	.5369
Included Covariates:							
HSGPA		**X**			**X**	**X**	X
SAT			X		**X**		X
SBAC				X		**X**	X

Note: rMSE calculated using ordinary least squares and five-fold cross validation.

In the second row, I add test scores to high school GPA in my predictive models. The graph on the left displays the results when high school grades and SAT scores are both included. Again, for most bins, the data spans at least two GPA points. This result also holds for models that include high school GPA and SBAC, as shown in the middle graph.

Table 1.9 contains similar results to Table 1.8, but here the models have been estimated on the sample of California State University students. The rMSE for models with only campus fixed effects included is .6076. Adding high school grades decreases the rMSE by .0526 GPA points to .5550. Adding test scores reduces it further, by .0103 to .5447 in the case of SAT scores and by .0087 GPA points to .5463 in the case of SBAC scores. As with persistence, high school GPA is the best single predictor of first-year GPA for CSU students, though now SAT scores perform marginally better than SBAC scores when added to a model that already contains high school GPA. Figure 4b presents a visual representation of these differences. Again, the differences are small, and for almost all bins in all six models, the data spans two GPA points.

There are various ways one could investigate subgroup differences. Here I estimate models using all enrolled students at each system and then calculate the rMSEs for each subgroup separately. For brevity I only report models with campus fixed effects and no demographic controls.[13] The results are reported in Tables 10 through 14.

The patterns that could be seen in Tables 6 though 9 can be seen here as well. At the University of California, the strongest single predictor of first-year outcomes is SAT scores. Further, models that include both high school GPA and SAT scores tend to perform better than models with high school GPA and SBAC scores. At the CSU's, the strongest predictor of first-year outcomes is high

Table 1.10: rMSE by Subgroup, Persistence to Second Year (University of California)

	(53)	**(54)**	(55)	(56)	**(57)**	**(58)**	(59)
All Students							
All	.2456	**.2451**	.2441	.2441	**.2438**	**.2439**	.2437
Socioeconomic Disadvantage							
No	.2121	**.2118**	.2113	.2113	**.2112**	**.2112**	.2111
Yes	.2827	**.2821**	.2804	.2805	**.2801**	**.2802**	.2799
Race / Ethnicity							
Asian/PI	.1887	**.1887**	.1871	.1872	**.1870**	**.1872**	.1869
Black	.2883	**.2883**	.2861	.2858	**.2865**	**.2862**	.2860
Hispanic	.2981	**.2973**	.2960	.2962	**.2956**	**.2958**	.2954
White	.2402	**.2399**	.2397	.2395	**.2396**	**.2394**	.2394
Gender							
Female	.2405	**.2403**	.2391	.2392	**.2390**	**.2392**	.2389
Male	.2521	**.2514**	.2504	.2504	**.2500**	**.2501**	.2498
Included Covariates:							
HSGPA		**X**			**X**	**X**	X
SAT			X		**X**		X
SBSA				X		**X**	X

Note: rMSE calculated using ordinary least squares and five-fold cross validation. All models include campus fixed effects.

Table 1.11: rMSE by Subgroup, Persistence to Second Year (California State University)

	(60)	**(61)**	(62)	(63)	**(64)**	**(65)**	(66)
All Students							
All	.6076	**.5550**	.5865	.5832	**.5447**	**.5463**	.5434
Socioeconomic Disadvantage							
No	.6013	**.5378**	.5780	.5727	**.5285**	**.5293**	.5269
Yes	.6132	**.5700**	.5941	.5926	**.5590**	**.5611**	.5579
Race / Ethnicity							
Asian/PI	.6024	**.5457**	.5762	.5673	**.5306**	**.5304**	.5282
Black	.6287	**.5679**	.6099	.5997	**.5617**	**.5597**	.5591
Hispanic	.6107	**.5697**	.5931	.5939	**.5596**	**.5629**	.5591
White	.5988	**.5274**	.5747	.5695	**.5199**	**.5202**	.5182
Gender							
Female	.5967	**.5432**	.5711	.5688	**.5318**	**.5337**	.5305
Male	.6231	**.5718**	.6084	.6037	**.5631**	**.5642**	.5617
Included Covariates:							
HSGPA		**X**			**X**	**X**	X
SAT			X		**X**		X
SBSA				X		**X**	X

Note: rMSE calculated using ordinary least squares and five-fold cross validation. All models include campus fixed effects.

[13] Additional results are available from the author on request.

school grades. Here, models with both high school grades and SAT scores perform tend to perform better than models with both high school grades and SBAC. For both the UCs and the CSUs, the models do a better job of predicting first-year outcomes for non-SED, Asian, White, and female students than SED, Black, Hispanic, or male students.

Table 12 also contains an example of the importance of using out-of-sample methods to estimate these models. We can see that for the Hispanic subgroup, the model that includes high school GPA and SAT scores to predict first-year GPA has a smaller rMSE (.5324) than does the model that also has SBAC scores (.5327). While admittedly, this is a very small difference that has little substantive meaning in terms of policy, it

Table 1.12: rMSE by Subgroup, First-Year GPA (University of California)

	(67)	**(68)**	(69)	(70)	**(71)**	**(72)**	(72)
All Students							
All	.5703	**.5458**	.5103	.5300	**.4939**	**.5152**	.4931
Socioeconomic Disadvantage							
No	.5328	**.5072**	.4779	.4966	**.4604**	**.4813**	.4596
Yes	.6160	**.5926**	.5500	.5707	**.5348**	**.5566**	.5339
Race / Ethnicity							
Asian/PI	.5372	**.5208**	.4874	.4942	**.4699**	**.4832**	.4678
Black	.6623	**.6271**	.5837	.5888	**.5699**	**.5784**	.5672
Hispanic	.6202	**.5950**	.5462	.5757	**.5324**	**.5612**	.5327
White	.5401	**.5043**	.4860	.5126	**.4661**	**.4905**	.4658
Gender							
Female	.5570	**.5322**	.4913	.5124	**.4757**	**.4985**	.4750
Male	.5880	**.5638**	.5357	.5532	**.5181**	**.5373**	.5171
Included Covariates:							
HSGPA		**X**			**X**	**X**	X
SAT			X		**X**		X
SBSA				X		**X**	X

Note: rMSE calculated using ordinary least squares and five-fold cross validation. All models include campus fixed effects.

Table 1.13: rMSE by Subgroup, First-Year GPA (California State University)

	(73)	**(74)**	(75)	(76)	**(77)**	**(78)**	(79)
All Students							
All	.6076	**.5550**	.5865	.5832	**.5447**	**.5463**	.5434
Socioeconomic Disadvantage							
No	.6013	**.5378**	.5780	.5727	**.5285**	**.5293**	.5269
Yes	.6132	**.5700**	.5941	.5926	**.5590**	**.5611**	.5579
Race / Ethnicity							
Asian/PI	.6024	**.5457**	.5762	.5673	**.5306**	**.5304**	.5282
Black	.6287	**.5679**	.6099	.5997	**.5617**	**.5597**	.5591
Hispanic	.6107	**.5697**	.5931	.5939	**.5596**	**.5629**	.5591
White	.5988	**.5274**	.5747	.5695	**.5199**	**.5202**	.5182
Gender							
Female	.5967	**.5432**	.5711	.5688	**.5318**	**.5337**	.5305
Male	.6231	**.5718**	.6084	.6037	**.5631**	**.5642**	.5617
Included Covariates:							
HSGPA		**X**			**X**	**X**	X
SAT			X		**X**		X
SBSA				X		**X**	X

Note: rMSE calculated using ordinary least squares and five-fold cross validation. All models include campus fixed effects.

does illustrate that using an incorrect methodology can lead to misleading results.

1.5 Who Gets Accepted Under Different Acceptance Regimes

1.5.1 Methodology

Including or excluding a measure of performance in a model does more than affect how well college outcomes are predicted. It also affects the ranking of applicants. If two students have the same high school GPA but one does better on the SAT than the other, a model that only includes high school GPA will rank them equally while a model that includes high school GPA and SAT score will rank the student with the higher SAT scores above the student with the lower SAT score. Similarly, if one model includes high school GPA and SAT scores while another model includes high school GPA and SBAC scores, the first model will advantage students who do well on the SAT, all else equal, while the second model will advantage students with high SBAC scores.

To gauge the potential policy implications of including and excluding different measures of performance on the demographic make-up of applicants accepted to college, I estimate each model using ordinary least squares and the sample of enrollees. I then use the estimated parameters from each model to predict the college outcomes for California public high school applicants. In models where the outcome of interest is persistence to the second year, this gives the predicted probability of each applicant's persistence to the second year. In models where the outcome of interest is first-year college GPA, this gives the expected first-year GPA for each applicant. I then rank each student twice, once based on their predicted probability of persistence and once by their GPA, and then examine the make-up of the top 10% of the applicants, as ranked by each model. While it is unlikely that these back-of-the-envelope methods fully capture the effects of including or excluding an assessment in the admissions process, the results can give policy makers a reference point for policy conversations.

1.5.2 Results

Tables 14 and 15 report the demographic makeup of applicants in the top 10% ranked first by the probability of persisting to the second year and then by the applicant's predicted first-year GPA. Students at the top of the ranking would be predicted to have either a higher probability of persistence or a higher first-year GPA, respectively. The results for each outcome are quite similar so I will focus on the top half of each table which report the demographic makeup of the top 10% of applicants as ranked by predicted persistence to the second year.

Table 1.14: Predicted Top 10 % of Applicant Pool (University of California)

	(80)	**(81)**	(82)	(83)	**(84)**	**(85)**	(86)
Outcome: Persist to 2nd Year							
SED	.4585	**.2839**	.0854	.1503	**.0976**	**.1537**	.1089
Asian/PI	.3642	**.3656**	.6432	.5537	**.5822**	**.5274**	.5780
Black	.0325	**.0217**	.0049	.0102	**.0062**	**.0093**	.0069
Hispanic	.3523	**.2183**	.0411	.0862	**.0573**	**.0978**	.0615
White	.2205	**.3547**	.2706	.3077	**.3111**	**.3203**	.3093
Female	.5814	**.6046**	.4401	.4350	**.4851**	**.4879**	.4835
Outcome: Cumulative GPA, End of 1st Year							
SED	.1295	**.2835**	.0786	.1504	**.0957**	**.1638**	.0988
Asian/PI	.8472	**.3673**	.6121	.5531	**.5413**	**.4986**	.5411
Black	.0072	**.0214**	.0062	.0103	**.0064**	**.0099**	.0065
Hispanic	.0129	**.2154**	.0454	.0869	**.0633**	**.1075**	.0636
White	.1159	**.3542**	.2931	.3077	**.3436**	**.3389**	.3425
Female	.6730	**.6034**	.4833	.4381	**.5268**	**.5125**	.5274
Included Covariates:							
HSGPA		**X**			**X**	**X**	X
SAT			X		**X**		X
SBSA				X		**X**	X

Note: All models include campus fixed effects.

Table 14 contains estimates for the University of California. Column (81) reports results from models that only include high school GPA as a measure of student performance. In this case, the predicted top decile of applicants at the University of California contains 28.39% SED students. If instead both high school GPA and SAT scores are used to rank students, the top decile is predicted to contain 9.76% SED students. A model that uses high school GPA and SBAC scores predicts that the top decile would contain 15.37% SED students. Adding SAT or SBAC scores to high school GPA leads to a 18.63 or 13.03 percentage point reduction in SED students respectively.

In Table 15, we can see that this pattern also occurs at the California State Universities,

though the overall level of SED students is higher. When only using high school GPA to rank applicants, 37.85% of the top decile are SED students. When high school GPA and SAT scores are used in combination, that number falls by 11.04 percentage points to 26.81%. Adding SBAC scores instead of SAT scores reduces the number of SED applicants by 7.85 percentage points to 30.00%.

Table 1.15: Predicted Top 10 % of Applicant Pool (California State University)

	(87)	**(88)**	(89)	(90)	**(91)**	**(92)**	(93)
Outcome: Persist to 2nd Year							
SED	.5344	**.3785**	.1825	.2577	**.2681**	**.3000**	.2873
Asian/PI	.1818	**.2979**	.4690	.4179	**.3789**	**.3628**	.3767
Black	.0525	**.0209**	.0102	.0130	**.0142**	**.0150**	.0145
Hispanic	.4986	**.3063**	.1101	.1718	**.1937**	**.2238**	.2084
White	.2356	**.3381**	.3664	.3515	**.3711**	**.3568**	.3589
Female	.5866	**.6310**	.4088	.4527	**.5472**	**.5689**	.5565
Outcome: Cumulative GPA, End of 1st Year							
SED	.1949	**.3797**	.1596	.2572	**.2553**	**.3155**	.2690
Asian/PI	.5713	**.2985**	.4160	.3996	**.3485**	**.3380**	.3470
Black	.0095	**.0207**	.0132	.0141	**.0160**	**.0169**	.0160
Hispanic	.1230	**.3077**	.1146	.1821	**.1994**	**.2456**	.2095
White	.2752	**.3365**	.4081	.3573	**.3930**	**.3587**	.3836
Female	.7108	**.6302**	.4522	.4895	**.5775**	**.6029**	.5853
Included Covariates:							
HSGPA		**X**			**X**	**X**	X
SAT			X		**X**		X
SBSA				X		**X**	X

Note: All models include campus fixed effects.

This pattern holds for Black and Latinx students as well. Contrastingly, Asian and Pacific Islander students increase their numbers in the top decile when standardized tests are used to rank applicants, with SAT scores providing a larger advantage than SBAC scores. White students are consistently about one third of the top decile no matter which model is used to rank applicants.

1.6 Policy Implications

These results indicate that, first, no performance measure or combination of measures included here substantially improves on the sample average for predicting persistence to the second year. Second, when modeling first-year GPA, including a standardized test in models with high school grades always improves the model's predictions. Yet, this improvement

is never greater than .06 of a GPA point. Third, for UC students and for the majority of CSU students, high school grades and SAT scores are a better predictor of first-year GPA than high school grades and SBAC scores. Still, this difference is never larger than .03 GPA points at the UCs and is always less than .01 GPA points at the CSUs. Fourth, adding SBAC scores to models that only include high school grades reduces the predicted percentage of SED students rated in the top 10% of applicants by at least 12 percentage points at the UCs and 6 percentage points at the CSU. Fifth, adding SAT scores to models only containing high school grades instead of SBAC scores reduces the predicted percentage of SED students even further. In this case, the percentage of SED students ranked in the top 10% of applicants fall by 18 and 12 percentage points for the UCs and CSUs respectively.

If all that is wanted is to maximize the predictive accuracy of whatever model is used to predict college outcomes, then basing admissions decisions on high school grades and SAT scores, instead of high school grades and SBAC scores or high school grades alone, would do that. But in no case would the models improve by a substantial amount. As the same time, including standardized tests in the admissions process, particularly SAT scores, greatly affects who is ranked in the top 10% of the applicant pool. Particularly, SED students are dramatically disadvantaged when standardized test scores are added.

Considering this trade off, it seems reasonable to argue that at a minimum, policy makers should avoid strengthening the relationship between SAT scores and the college admissions process. In California, SBAC scores could be seen as a reasonable alternative to SAT scores, allowing the California public universities a second assessment that improves predictive outcomes almost as much as SAT scores while disadvantaging SED applicants to a lesser degree. However, K-12 academic standards change over time, and the tests that are used to measure those standards change with them. In the last two decades, California has used four different assessments to measure student performance, the Stanford Achievement Test, the California Achievement Test, the California Standards Test, and the SBACs. Using the SBACs as a basis for admission to college would put pressure on K-12 policy makers not to

change the way that students are assessed even when standards need to be updated.

As such, it may be time to jettison the use of standardized test scores in the admissions process and instead rely solely on high school grades. The main argument against this course of action is that doing so could render grades meaningless as schools might inflate grades without the check of some standardized score (see for example Hurwitz & Lee, 2018). Certainly, doing away with using standardized test scores during the admissions process at all colleges and universities would have a different effect than one more school becoming SAT/ACT optional. But there are checks against grade inflation already at a college or university's disposal. The high school a student attended is part of that student's application. Students can be evaluated in relation to their peers. More research needs to be done before this course of action should be endorsed wholeheartedly, but some is already underway (Cohen, forthcoming).

1.7 Conclusion

I find that standardized tests scores improve the predictive power of models of first-year college GPA over models using high school grades alone. But these improvements are small. At the same time, including standardized test scores in application process dramatically decreases the number of SED students who are ranked in the top 10% of the applicant pool at both the UCs and the CSUs. This trade-off should give pause to anyone looking to strengthen the relationship between SAT scores and the admissions process. There are a few caveats. The students in this study were the last cohort to take the SAT before the redesign, and they were the first to take the SBACs. Once more recent data becomes available, more research should be done on students who take the current format of the SATs and who were not the first cohort to take the SBACs.

1.8 References

Anderson, Nick, "A Shake-Up In Elite Admissions: U-Chicago Drops SAT/ACT Testing Requirement," Washington Post, June 14, 2018.

Black, S. E., Cortes, K. E., & Lincove, J. A. 2016. Efficacy versus equity: What happens when states tinker with college admissions in a race-blind era? Educational Evaluation and Policy Analysis. 38:2, 336-363.

California State Legislature. Assembly. Pupil assessments: Pathways to College Act. AB-1951. California Legislature 2017-2018 Regular Session., Introduced January 29, 2018.

Cohen, Jacob, 1988. Statistical Power Analysis For The Behavioral Sciences, 2nd ed. Hillsdale, NJ: Lawrence Earlbaum Associates.

Dynarski, Susan, "ACT/SAT for all: A cheap, effective way to narrow income gaps in college," Brookings, Febuary 8, 2018.

Gewertz, Catherine, Which States Require Students to Take the SAT or ACT? An Interactive Breakdown of States' 2016-17 Testing Plans, Education Week, June 15, 2018, <https://www.edweek.org/ew/section/multimedia/states-require-students-take-sat-or-act.html>.

Gulliksen, H. (1950). Theory of Mental Tests. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, 2009, The Elements of Statistical Learning, New York, NY: Springer.

Hurwitz, Michael and Jason Lee, 2018, Grade Inflation And The Role Of Standardized Testing, Jack Buckly, Lynn Letukas, and Ben Wildavsky (ed.) Measuring Success. XX – YY.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, An Introduction To Statistical Learning, New York, NY: Springer.

Kobrin, Jennifer L., Brian F. Patterson, Emily J. Shaw, Krista D. Mattern, and Sandra M. Barbuti. 2008. Validity of the SAT for Predicting First-Year College Grade Point Average. College Board Research in Review 2008-5. New York: The College Board.

Kurlaender, Michal and Kramer Cohen. 2019. Predicting College Success: How Do Different High School Assessments Measure Up? Policy Analysis for California Education. Stanford: PACE

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. Royal Society of Edinburgh, Proceedings, 62(1), 28–30.

Lewis, C. (2006). Selected topics in classical test theory. In C.R. Rao, & S. Sinharay (Eds.), Handbook of Statistics 26 Psychometrics (pp.29-42). Amsterdam, NL: North Holland Publishing.

Lord, F.M., Novick, M.R., & Birnbaum, A. (1968). Statistical Theories of Mental Test Scores. Oxford, England: Addison-Wesley.

Mattern, Krista D., Jennifer L. Kobrin, Brian F. Patterson, Emily J. Shaw, and Wayne J Camara. 2009. Validity is in the Eye of the Beholder. Robert W. Lissitz (ed.) The Concept of Validity. 213-240.

Mattern, Krista D., Brian F. Patterson, Emily J. Shaw, Jennifer L. Kobrin, and Sandra M. Barbuti. 2008. Differential Validity and Prediction of the SAT. College Board Research in Review 2008-4. New York: The College Board.

Pearson, K. (1903). Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. Philosophical Transactions of the Royal Society of London, 200, 1–66.

Rothstein, J. M. 2002. Admissions Bias: A New Approach to Validity Estimation in Selected Samples. CSHE.3.02 1-16.

Rothstein, J. M. 2004. College performance predictions and the SAT. Journal of Econometrics, 121:1-2, 297-317.

Shaw, E. J., Marini, J. P., Beard, J., Shmueli, D., Young, L., & Ng, H. 2016. The Redesigned SAT Pilot Predictive Validity Study: A First Look. College Board Research Report 2016-1. New York: The College Board.

1.9 Match methodology

. To match a student's college records to their high school achievement test scores, I used the student's first and last name, date of birth, gender, and high school in 68 different combinations. All match attempts include date of birth and some name information, the first 17 attempts also include gender and high school. Attempts 18-34 include gender but not high school, attempts 35-51 include high school but not gender, and then attempts 52-68 include neither high school not gender. I will describe the fourth set of attempts in detail, since it is the basis for all the other sets. All match attempts include date of birth.

In the first attempt, I attempt to match students using their first and last name "as is" in the SBAC and college records, with the exception that I remove all capitalization as

capitalization conventions vary between datasets. In the second attempt, all punctuation in a student's first and last name is removed, with the exception of dashes which are turned into spaces. In the third attempt all spaces are also removed.

Then I divide both the student's first name and last name into their component "sub names". For example, if a student's first name is "Mary Jane", then for the first name the first subname is "Mary", the second is "Jane", and the third is empty. If a student's last name is "Ostrom Hopper-Johnson", then for the last name the subnames are "Ostrom", "Hopper", and "Johnson". For the next two steps I used the first five first and last subnames to match followed by the first four. At this point I limit all match attempts to the first three subnames for the first and last names, and permutate through the nine combinations. I start by using the first three subnames for the last name and first three three subnames for the first name, followed by first two subnames for the first name, followed by the first subname for the first name. I repeat these steps using the first two steps subnames and then only the first subname. For the last three steps, a student's first name is limited to the first three letters of their first name. I then use three, then two, then one subnames with the first three letters of the first name as attempted matches.

1.10 Match Percentage

85.2% of the first time freshman in the UC files and 82.6% of the first time freshman from California public high schools in the CSUs files match to SBAC records. If all college bound 11th grade public high school students took the SBAC, graduated on time, applied for and enrolled in college in the Fall of 2016, could by uniquely identified by their name, date of birth, gender, and high school, and filed out the informational section of their school paperwork consistently, we would expect match rates of 100%. To better understand why this is not the case, I examine the students who do not match for evidence as to why they do not match.

To begin with, I identify all students who are not uniquely identified by name, DOB, gender and high school in the samples. In the UC sample, this information uniquely identifies all students. In the CSU sample, 24,234 of the 31,282 non matches are from entries where name, DOB, gender, and high school do not uniquely identify a student. This could be because not all students applying to the CSUs can be uniquely identified by name, DOB, gender, and high school. However, since all students in the UC data are uniquely identified by these characteristics and virtually all students in the SBAC data are also uniquely identified by these characteristics, it seems likely that some students are in the CSU data multiple times with different id numbers. Applicants in this cohort applied to each UC campus using a single application while applicants to each CSU campus used separate applications. This could account for the difference. If we assume that a student can be uniquely identified using name, DOB, gender and high school, these 24,234 entries represent 7,897 unique individuals.

This leaves 15,663 unique unmatched individuals from the UC sample and 7,048 unique unmatched individuals from the CSU sample. I compare the average age of these unmatched students to the average age of the matched students. At the UCs, the matched students are 59 days younger than unmatched students, on average. Further, given that students had to be five by December 2, 2003 in order to enroll in kindergarten in California that year, we would expect that the vast majority of students in the sample to be born between December 2, 1997 and December 2, 1998. In the matched sample 88% of the students are in fact born during this time period whereas only 74% of the unmatched students are born during that year.

In the CSU sample, matched students are 313 days younger than unique unmatched students. Further, 88% of matched students were born between December 2, 1997 and December 2, 1998 while only 49% of unique unmatched students were born in this time period. Given that in actuality not all students actually take the SBAC , and we would expect some paperwork mistakes, the CSU match seems to be the vast majority of students who could be matched between the SBAC and CSU datasets.

The number of unexplained matches between the SBAC and UC data was more puzzling. I followed up with the UC and found out that first time freshman who had not attended California private high schools had accidentally been included in the file. About 7.5% of California K-12 students attend private high schools. This percentage is probably higher in the later grades. Having private high school students in the file of first time freshmen would not affect my results in any way, but would depress the match rate. As such, 85.2% seems like a reasonable match.