SAT Scores, High Schools, and Collegiate Performance Predictions

Jesse Rothstein[*]

Princeton University

Abstract

Family background is correlated with collegiate performance, as is the SAT. Moreover, SAT scores are themselves highly correlated with family background. Validity research has rarely taken account of these connections, though the extent to which the SAT's predictive validity derives from its correlation with background is crucial to the interpretation of prediction models. Some of the most powerful proxies for family background are measures of the demographic characteristics and test performance of the high school attended, perhaps because parents who are involved with choosing good schools for their children tend also to be involved with their children's education in other ways. Using data from the University of California, I examine the role of high-school-level characteristics in SAT validity models. I find that the school average SAT score is a substantially more powerful predictor of collegiate performance than is the individual deviation from that average. A portion, but by no means all, of the school average SAT effect can be attributed to its correlation with demographic characteristics of the school, particularly the racial composition. Validity models that use the unadjusted SAT score without demographic controls thus overstate the direct contribution of the individual SAT to prediction, attributing to it substantial variation that is better attributed to readily-observed school characteristics.

[*] Industrial Relations Section, Firestone Library, Princeton, NJ 08544. (609) 258-4045; jrothst@princeton.edu. I thank Saul Geiser and Roger Studley at the University of California Office of the President for access to data, and Daniel Koretz for helpful comments.

I. Introduction

Perhaps the strongest argument for using scores from the SAT exam in college admissions is that these scores help admissions offices to predict students' eventual performance in college. It is in nobody's interest to admit an applicant who is unprepared to succeed in college, so any tool that helps to predict success has clear value in the admissions process (Barro, 2001; McWhorter, 2001; Camara, 2001).

A long literature has established the *predictive validity* of the SAT score for collegiate performance, usually measured as the grade point average (GPA) in the freshman year. Examples include Bridgeman et al., 2000; Camara and Echternacht, 2000; Stricker, 1991; and Willingham et al., 1990. The most common methodological concerns in this literature are sample selection, deriving from the fact that most validity studies use samples of students who attend a single college and are therefore not random samples from the population (Breland, 1979; Camara and Echternacht, 2000), and the construction of a standardized measure of collegiate GPA that has similar meaning for students who enroll in very different courses (Elliot and Strenta, 1988; Goldman and Widawski, 1976; Young, 1993).

I argue in this paper that the most important methodological limitation of traditional validity studies is neither of these, but rather the paucity of predictor variables included on the right-hand-side of regression models.[1] I take as my starting point a presumption that the goal of validity studies is to estimate the amount of information provided by an applicant's SAT score, over and above the other information available in the application, about how well the applicant will do if admitted. Crucially, this understanding of validity abstracts from issues of causal inference: The causal path underlying any correlation between a predictor variable and student outcomes is irrelevant, so long

[1] Willingham (1985) and Willingham and Breland (1982) explore a few non-standard predictor variables.

as the former variable helps to predict the latter. I focus on information about the student's high school. The school that a student attends conveys a great deal of information about his or her future performance, but school-level variables are rarely included in validity models.[2]

I demonstrate that simple, widely-available school-level variables are important predictors of both individual SAT scores and collegiate performance, and that the exclusion of these variables from validity models leads to substantial overstatement of the independent importance of the individual SAT. In particular, students from schools with high average test scores earn substantially higher freshman GPAs (hereafter, FGPAs), on average, than do students with similar scores from schools with lower averages. To some extent, this reflects a correlation of the school average SAT score with other characteristics of the school: Schools with relatively low shares of black and Hispanic students, with high shares of Asian students, with parents with high levels of education, and with high scores on state accountability exams typically have higher average SAT scores and higher average FGPAs than do more disadvantaged schools. However, even when all of these characteristics are controlled, the school average SAT retains substantial predictive power for students' college performance.

Admissions offices generally do not use mechanical rules, and frequently admit students with lower predicted performance while rejecting students who are predicted to earn higher GPAs on average. This arises for the simple reason that predicted performance, while often important, is not the only criterion in admissions decisions. Thus, for example, it is well established that black students typically earn lower FGPAs than do white students with similar SATs and HSGPAs

[2] Admissions officers are well aware of the value of school characteristics for prediction. One commonly hears, for example, that selective colleges give preferences to students from schools that have historically sent many ultimately-successful students to the college in question. Although I do not have access to measures of the performance of students from prior cohorts, the results presented here suggest that this is probably a useful strategy for identifying students likely to succeed.

(Young, 2001),[3] but it is nevertheless common practice to admit blacks over whites with similar credentials. My results imply a similar situation for so-called "strivers," students who earn higher SAT scores than their high school classmates. It seems clear that colleges would like to give preferences to these students (Marcus, 1999). My results suggest that the performance-maximizing rule would be the opposite, with students who are below average at high-scoring high schools given preference over competitors with similar scores from schools with much lower averages.

This has no particular implication for admissions policy, as colleges may well have good reasons to admit "strivers" despite their low predicted FGPAs. However, the flip side of this result is that within-school differences in SAT scores have much less predictive power than would be implied by traditional validity models that fail to distinguish across- from within-school variation. Only the within-school component can plausibly be called the individual SAT score's predictive contribution. The results thus suggest that the SAT plays a substantially less important role in identifying qualified students than would be indicated by traditional validity studies.

II. A Note on Causation

As demonstrated below, high schools' demographic characteristics are strong predictors of their students' eventual performance in college. It is worth emphasizing that this says little about causality. Although it is certainly possible that there is a direct causal path leading from school composition to future student performance—for example, students may learn better study habits when surrounded by students with highly educated parents—there are a variety of other hypotheses that could explain the predictive relationship. To take but one example, demographically advantaged

[3] Sander (2005) implicitly attributes this gap to the effects of affirmative action, as a result of which blacks have the option of attending more selective schools than do observationally-equivalent whites. But the gap appears even within the same institution, belying the affirmative action explanation.

families may seek out well-run high schools, leading demographic characteristics to serve as a proxy for difficult-to-measure school quality in prediction models.

It is crucial to keep in mind, however, that *predictive validity* has little to do with causation. Evidence for the predictive validity of SAT scores, for example, cannot be interpreted as evidence that SAT scores have a causal effect on student performance.[4] Rather, predictive validity studies are often interpreted as evidence for statements such as "The SAT has proven to be an important predictor of success in college. . . . SAT scores add significantly to the prediction" (Camara and Echternacht, 2000). The question at issue in this study is whether statements like this are in fact true, or whether SAT scores merely duplicate information about future performance that is otherwise available from demographic variables. If so, admissions offices may be able to obtain equally accurate predictions even without access to the SAT.

To the extent that the SAT's predictive power is found to reflect its association with characteristics—like demographics—that are not seen as appropriate admissions variables, predictive validity evidence alone cannot support arguments for the SAT's importance to the admissions process. Even the addition of judgments of content validity is of limited relevance: It would certainly seem possible to design an exam—perhaps a math exam set in the context of calculating golf handicaps—that appeared to have content validity but which in practice acted primarily as a proxy for student background. In the most extreme case, if the SAT adds zero predictive power to what is otherwise available from demographic variables, it would be inappropriate to allow the SAT to "launder" the demographic variation by excluding these variables from our prediction model, as is commonly done. Even in less extreme cases, if a substantial portion of the information provided by

[4] Indeed, it is not clear what it would mean for a test score to have a causal impact of this sort. Clearly, interventions that raise SAT scores might also improve future performance, but test coaching, cheating, re-taking the exam (Clotfelter and Vigdor, 2003), and studying the underlying material, all of which might raise SAT scores, would be expected to have quite different impacts on future performance.

SAT scores about future performance is also available from other variables—even if these variables lack a simple causal relationship with either the SAT or the performance measure—then only the portion of the SAT's predictive power that is not otherwise available should count in its favor.

III. Omitted Variables in the Basic Validity Model

The basic validity model relates the individual SAT score to the FGPA that the student ultimately earns:

(1) $FGPA_i = \alpha_1 + SAT_i \beta_1 + \varepsilon_i$.

The SAT is judged to have predictive validity if it is an important predictor of FGPA. One might use any of several measures of the SAT's importance. The coefficient β_1, for example, measures the amount by which a student's FGPA would be expected to exceed, on average, that of another student whose SAT score was one point lower. Validity studies typically focus instead on an alternative measure,

(2) $R = corr(FGPA, SAT) = \beta_1 * [Var(SAT) / Var(FGPA)]^{1/2}$

The bivariate correlation between SATs and FGPAs provides an upper limit to the SAT's predictive contribution. It is widely recognized that other variables—most commonly, the high school GPA or class rank—are available for prediction, and that the SAT's contribution is only the extent to which predictions are improved by its addition to the model. This implies a regression of the form

(3) $FGPA_i = \alpha_2 + SAT_i \beta_2 + HSGPA_i \gamma_2 + \varepsilon_i$.

The SAT's contribution might be measured by β_2 or by the increment to R over a model that excludes the SAT. These are related by

(4) $\Delta R = \beta_2 * [(1 - \rho^2) * Var(SAT) / Var(FGPA)]^{1/2}$,

where $\rho = corr(SAT, HSGPA)$.

A standard omitted variables formula gives the relationship between β_2 and β_1:

(5) $\beta_2 = \beta_1 - \theta * \gamma_2$,

where $\theta = \varrho * [\text{Var(HSGPA)} / \text{Var(SAT)}]^{1/2}$. Assuming that both ϱ and γ_2 are positive, β_2 will be smaller than β_1. Moreover, because $(1 - \varrho^2)$ is less than 1, ΔR must be less than R by an even greater degree: $(\Delta R / R) \leq (\beta_2 / \beta_1)$. Thus, by either measure, the SAT's estimated contribution must (weakly) decline with the addition of HSGPA to the prediction model.

There is no principled reason to stop here. Ideally, a number of other variables—the quality of the essay, the strength of the recommendation letters, a measure of the difficulty of the high school curriculum, etc.—should be included in (3) along with the SAT and HSGPA. Typically, these are excluded simply because they are much more difficult to measure, and because the data sets used for validity estimation do not provide the necessary variables. Once again, if these variables could be added to (3), the SAT's ΔR necessarily declines. For likely parameter values, β will decline as well, although it is in principle possible for it to increase.[5] As it makes no sense to credit the SAT with predictive power that is available in any case from other aspects of the application, the conceptually correct measure of the SAT's contribution is the β or ΔR from a model that controls for all other variables considered in the admissions decision.

It is slightly less clear that variables not considered in admissions but available at the time of the student's application should be included as predictors in (3). Consider, for example, the racial composition of the high school, at a university that practices race-blind admissions.[6] It is entirely conceivable that the racial composition of the school is correlated both with SAT scores and with FGPAs. Suppose, hypothetically, that β (and therefore ΔR) declined to zero when the school racial

[5] β would decline if the partial correlation of the added variables with SAT and the partial correlation with FGPA are of opposite signs. An example might be if students with strong essays had lower SAT scores but higher FGPAs than students with weaker essays.

[6] I assume for this hypothetical that "race-blindness" precludes direct consideration not just of individual race but also of the school racial composition, though this need not be true.

composition was included in (3). This would indicate that the SAT's predictive power for FGPAs derived entirely from its association with the school racial composition, and that an admissions office that had access to the racial composition variables would derive no additional benefit from considering individual SAT scores. At our hypothetical university, the former variables are not available for admissions, so predictions are improved by using the SAT score. But this hardly seems consistent with the goal of race-blindness: The SAT functions in the admissions process merely as a proxy for the disallowed race variables. A university that is serious about race-blindness should not be willing to use the SAT to "launder" the disallowed variation, so should eschew the SAT in admissions. Similarly, a researcher should consider that the SAT score provides no incremental information about students' eventual performance, a conclusion that is unaffected by the university's decision not to take advantage of all of the available information.

In the real world, of course, things are not as simple as in this hypothetical example. As I demonstrate below, high school variables are indeed correlated with both SAT scores and FGPAs. Their inclusion in the validity model substantially reduces the estimated β, but does not drive it to zero. Thus, the SAT does provide some information about FGPAs beyond what is available from the school-level variables. Nevertheless, because the school-level variables that I consider are available in the admissions process, the portion of the SAT's contribution to FGPA prediction in models that exclude these variables overstates the amount of independent information provided by the individual SAT score. A more appropriate measure of the information provided by the individual SAT score is β or ΔR from the richer model that includes the school-level variables.

A. Selection, Bias, and Restriction of Range

The discussion thus far presumed that the validity model, (3), was estimated on a random sample of applicants to a college. This is infeasible, as FGPAs are only observed for those students

who are admitted and who enroll. The distribution of each variable, and the correlations among variables, are likely to be different in the sample of matriculants than in the population of applicants. This introduces two important complications. First, sample selection may bias estimates of the regression coefficients in models like (3). Second, even with unbiased estimates of β_2 and γ_2, ΔR may be biased by the "restriction of range" of SAT scores and HSGPAs in the sample of admittees, as the admissions process almost certainly causes the variance of these variables (as well as Var(FGPA) and ϱ) to be lower in the sample than in the population.

The validity literature has long noted the second problem, and estimates of ΔR are typically "corrected for restriction of range," computed using estimates of Var(SAT) from population rather than sample data, with a similar correction for Var(FGPA).[7] But the first problem has generally been ignored. Rothstein (2004) points out that usual practice cannot be justified by any sample selection assumptions, and that restriction-of-range corrected estimates of the SAT's validity are consistent only if sample selection is random (i.e. uncorrelated with ε or SAT) conditional on HSGPA. This is implausible at any college that considers the SAT in admissions. With endogenous sample selection, β_1 and likely β_2 and γ_2 are biased. Rothstein (2004) proposes an alternative procedure for restriction of range correction that produces consistent estimates of ΔR so long as the coefficients of model (3) can be estimated without bias.

In the current paper, I present only the coefficients from models like (3), and do not extend them to estimate ΔR. This avoids the complication of correction for restriction of range. It is still necessary to obtain unbiased regression coefficients. As noted by Rothstein (2004), sufficient conditions for unbiasedness are that all of the variables considered in the admissions decision are included as predictor variables in model (3), and that decisions to enroll once admitted are not

[7] Note that $Var(FGPA)=Var(\alpha_2 + SAT*\beta_2 + HSGPA*\gamma_2) + Var(\varepsilon)$. In range-corrected estimates of ΔR, the first term of this is computed from population data (using sample estimates of the regression coefficients). $Var(\varepsilon)$ is assumed consistently estimated from the sample residuals.

correlated with the regression error ε. These conditions are implausible in most contexts, as the researcher rarely has access to all variables used in admissions. I argue below, however, that it is somewhat plausible in the data used for the current analysis.

B. Reliability

No exam is perfectly reliable. It is well known that measurement error in tests—an imperfect correlation between an underlying "true" score and the actual score obtained on a test of finite length—attenuates coefficients in prediction models. Indeed, under conventional assumptions, standard formulae for errors-in-variables regression models give the large-sample attenuation bias (Deaton, 1997, p. 99). Let r be the reliability of the SAT exam. Then the coefficient of the univariate regression of FGPAs on SATs is attenuated by a factor r:

$$(6) \qquad \text{plim}\ \hat{\beta}_1 = r\beta_1 < \beta_1 .$$

The attenuation factor for a multivariate regression of FGPAs on SATs and other variables (HSGPA, etc.) is slightly more complex:

$$(7) \qquad \text{plim}\ \hat{\beta}_2 = \frac{r - Q^2}{1 - Q^2}\beta_2 < \beta_2 ,$$

where Q^2 is the explained share of variance from a regression of SAT on the other included variables. This is a declining function of Q^2, so the more variables that we include in our model, the greater the attenuation of the SAT coefficient.

This effect cannot be neglected in the current study, as declines in the SAT coefficient that derive purely from the addition of variables that "soak up" the signal in the SAT do not offer evidence for the differential predictive power of demographic and non-demographic components of the SAT score. I present models below that include as a control variable in the validity model the school average SAT score. Because this average is likely more reliable than is the individual SAT, its

inclusion will tend to reduce the individual SAT coefficient even if across- and within-school variation in the characteristics measured by the SAT are equally predictive of future performance.

Equation (7), however, helps us to understand the extent to which declines in the SAT coefficient with the addition of more variables can be written off as purely statistical. The reliability of the SAT is quite high, above 0.9.[8] I demonstrate below that, while SAT scores vary substantially across high schools—more than do HSGPAs or FGPAs—three-quarters of the variation in SAT scores is still within schools. Thus, the addition of school effects to prediction models should be expected to reduce the attenuation factor in (7) from 0.9, with only the SAT included as a predictor, to 0.86 (=(0.9-0.25)/(1-0.25)), with the school effects.[9] The small difference between these suggests that any large change in the estimated SAT coefficient cannot be attributed to reliability issues, and must indicate that SATs convey different information across and within high schools.

IV. Data and Methods

I use an unusually large and rich data set extracted from University of California (UC) administrative records. The data contain observations on all California residents who applied to any of the UC campuses for admission as regular freshmen for the 1993-1994 academic year. Variables include self-reported SAT scores[10] and high school GPAs, as well as identifiers for the high school attended. For those students who enrolled at one of the UC campuses, additional variables describe

[8] The College Board (2003) reports that the reliability of the math and verbal scores are each 0.91-0.93. The reliability of the composite score must be at least that high if the correlation between the two true subscores is non-negative.
[9] This calculation neglects the HSGPA, which is included in the validity models that I estimate but has little impact on this back-of-the envelope estimate. Rothstein (2004) presents coefficient estimates that are corrected for the imperfect reliability of the SAT.
[10] Scores are reported on the pre-1994 scale. I do not have separate math and verbal scores, but only the composite. When I combine the UC data with auxiliary data sets that report post-1994, "recentered" scores, I convert the latter to the earlier scale.

the campus attended, the major during the freshman year, and the grade point average during the freshman year.[11]

These data have two important features that act to minimize selection problems which otherwise plague validity estimates. First, the campuses of the UC system are quite disparate in their selectivity. Thus, while validity studies that use data from a single campus often rely on a very thin slice of the college-going population, the UC sample represents a wide range of SAT scores and HSGPAs. By itself, this reduces sample selection biases, which are most severe for observations close to the admissions threshold where exceptional unobservable characteristics would have been required to offset marginal SATs.

The second advantageous feature of the UC data is that eligibility to the UC system is a known, deterministic function of SAT scores and HSGPAs. Eligible students are guaranteed admission to at least one campus, while ineligible students can be admitted only "by exception," with these exceptions limited to no more than 6% of each campus's admissions offers. I am able to compute an approximate measure of eligibility;[12] the admissions rules guarantee that any eligible student—regardless of his or her unobserved characteristics—had the option to attend the University of California. It is at least plausible that the matriculation decisions of admitted students are uncorrelated with ε. If this is true, estimates of equation (3) from the subsample of eligible students are unbiased by sample selection.

I include full sets of campus and major effects in all specifications to absorb differences in grading standards. Of course, admission to individual campuses takes into account both observables

[11] One of the eight campuses, Santa Cruz, allows students the option of taking courses without grades, and many students from that campus do not have valid GPAs. All observations from that campus are dropped from all analyses here.

[12] The full eligibility rules impose requirements on the high school courseload, and use a slightly different construct of HSGPA than my measure. I ignore these complications, and assume that every student whose SAT scores and observed HSGPAs meet the eligibility threshold is in fact eligible.

and unobservables, so the campus effects may be endogenous. Rothstein (2004) presents evidence that any endogeneity of campus choice does not substantially bias the SAT and HSGPA coefficients.

I begin with simple decompositions of SAT, HSGPA, and FGPA into across- and within-school components. I then estimate a variety of models of the form

$$(8) \qquad FGPA_{ij} = \alpha + SAT_{ij}\,\beta + HSGPA_{ij}\,\gamma + AvgSAT_j\,\theta + AvgGPA_j\,\omega + \varepsilon_{ij},$$

where $FGPA_{ij}$ is the FGPA of student i from school j; SAT_{ij} and $HSGPA_{ij}$ are his or her SAT score and high school GPA; and $AvgSAT_j$ and $AvgGPA_j$ are the average SAT and average HSGPA at the school. I explore different computations of these averages—over students in the sample, over all UC applicants, and over all SAT-takers—with little effect on the results. Note that when the averages are computed over the sample used for the estimation of the regression, β and γ are identified only from within-school variation in the predictor variables, while θ and ω reflect the predictive power of across-school variation over and above that of within-school variation.[13] I also explore specifications that add to (8) measures of the socioeconomic composition of the school's student body and of the school's scores on state accountability exams, to explore the impact of this on the θ coefficient. Finally, I discuss the relationship between my results and the literature on "overprediction" (Ramist, Lewis, and McCamley-Jenkins, 1993).

V. Within- and Between-School Components of the SAT

Table 1 presents preliminary validity models, first on the full sample of matriculants at the University of California (Columns A-C) and second on the subsample of matriculants who were UC-eligible (Columns D-F). As argued earlier, endogenous admissions decisions may bias prediction coefficients in models that include ineligible students; the selection-on-observables of the

[13] That is, equation (6) can be re-written as $FGPA_{ij} = \alpha + (SAT_{ij} - AvgSAT_j)\beta + (HSGPA_{ij} - AvgGPA_j)\gamma + AvgSAT_j\,(\beta + \theta) + AvgGPA_j\,(\gamma + \omega) + \varepsilon_{ij}$. Estimates of β and γ are not precisely identical to those obtained from true within-school models with high school fixed effects, as within-school models implicitly control for the school-level averages of the campus and major fixed effects as well, but are quite similar in practice.

eligibility determination may, if enrollment decisions are uncorrelated with ability, permit for unbiased estimation using the sample of eligible students. Note, however, that the eligibility rule encompasses both SAT scores and HSGPAs, so only the model containing both (Column F) is even plausibly unbiased. The HSGPA coefficient is notably lower in Column F than in Column C, consistent with the HSGPA's important role (much larger than that of the SAT) in the eligibility formula. Rothstein (2004) explores alternative specifications designed to detect bias deriving from endogeneity of matriculation or of campus or major choice, with little impact on the relevant coefficients.

The key concern of this paper is the potential for differential predictive validity of the across- and within-school components of the variation in SAT scores. As a preliminary step, Table 2 presents variance decompositions of SAT scores, HSGPAs, and FGPAs into across- and within-school shares, computed from the estimation sample of eligible UC matriculants. Column D of the table presents a similar decomposition for SAT scores in the population of California SAT-takers. The across-school share of the variation in SAT scores is substantially higher than that of either HSGPAs or FGPAs.

Table 3 presents an exploration of the predictive power of across- and between-school variation in SAT scores and HSGPAs in the UC-eligible sample. Column A repeats the specification from Column F of Table 1. Column B presents the SAT and HSGPA coefficients from a specification that includes fixed effects for each high school in the sample. Not surprisingly, these fixed effects add substantially to the predictive power of the regression. Note, however, what happens to the SAT and HSGPA coefficients: The SAT coefficient falls by nearly half, while the HSGPA coefficient rises substantially. Evidently, SAT scores are much worse at predicting FGPAs within schools than they are in the sample as a whole, while HSGPAs are much better at within-

school prediction. (The latter result, of course, is consistent with the idea that HSGPA scales differ substantially across schools, perhaps in part a consequence of differential grade inflation.)

The across- and within-school distinction is made clearer in Column C, which replaces the school fixed effects with the sample means of HSGPA and SAT. If the across- and within-school variation in these variables were equally predictive of FGPAs, the school means would have coefficients of zero. The positive coefficient on the school average SAT indicates, once again, that across-school SAT variation is substantially—over three times—as predictive of college performance as is within-school variation. The individual and school mean HSGPA have opposite signs and are similar in magnitude, indicating that within-school variation in HSGPAs is quite predictive of eventual FGPAs but that across-school variation provides very little information.

The models in Columns B and C are not fair validity models, however, as they include as predictors information that is arguably unavailable at the time of the admissions decision, when it is impossible to compute average SATs among students who will eventually matriculate. The fixed effects model in Column B represents an upper limit of the extent to which admissions offices can use across-high-school variation to predict FGPAs. To the extent that the school fixed effects reflect transitory differences between schools that are unobserved by the admissions office, the performance of this model could not be approached in real-world admissions.[14] Similarly, the sample average SATs and HSGPAs in the high school that are used in Column C would be available only after admissions and enrollment decisions were completed, as only then would the sample be determined. Thus, it is possible that reliance on these measures is misleading about what an admissions office could do to use the across-school component of FGPAs in predictions.

[14] One way to evaluate this would be to include the estimated school fixed effect from a previous year as a predictor in the current year's validity model. This would approximate the practice of giving preferences to students from schools whose past graduates have been successful. I do not have access to data that would permit this.

The remaining columns of Table 3 present specifications that estimate school averages over samples that are progressively closer to what might be available at admissions time. In Column D, the averages are computed over all eligible UC enrollees (including those at UC Santa Cruz and others without valid FGPAs); in Column E over all eligible UC applicants, including those who did not ultimately enroll; and in Column F over all UC applicants, eligible or not. None of these alterations has important effects on the prediction coefficients.

All of the preceeding models have averaged SATs and HSGPAs over subsets of students who applied to the UC. Even more readily available to an admissions office is the average SAT among all SAT-takers in the school, whether or not they applied. I compute this using College Board data on all SAT-takers from the 1994 cohort (one year behind the cohort from which my UC sample is drawn). Although students are asked their HSGPA when taking the SAT, it is reported only in discrete categories, so is not directly comparable to that available in the UC data. As a result, comparable average HSGPAs among SAT-takers at the school are unavailable. Panel B of Table 3 repeats the earlier estimates without the school average HSGPA. The exclusion of this variable reduces the school average SAT coefficient and inflates the coefficient on the individual SAT score. Finally, Columns G and H measure the school SAT average over all UC-eligible SAT-takers, whether or not they applied, and over all SAT-takers at the school. This modification causes the school average SAT coefficient to decline notably, but has essentially zero effect on the other coefficients.

VI. School Demographics and Between-School SAT Variation

It would be arguably legitimate to give preferences in admissions to students who attended good high schools. On the other hand, many policymakers would, quite understandably, balk at the idea of basing admissions in part on the racial composition of the high school, or even at rewarding

school quality if the quality measure turned out to be highly correlated with the fraction of white students at the school. It is not clear a priori which of these extreme cases best characterizes the school average SAT score. Thus, a central question for the interpretation of the results in Table 3 is the extent to which the school average SAT scores provide independent information about students' preparedness, rather than simply proxying for school demographic characteristics.

As a preliminary step toward analyzing this question, Table 4 presents school-level models that take the school average SAT score—averaged over all test-takers—as the dependent variable. The sample in this table is public schools in California that sent at least one student to the UC in 1993-1994, and estimates are weighted by the number of test-takers at the school.

As schools vary widely in the fraction of students who take the exam, and as test-takers are not sampled randomly from within the school population, an important concern in relating school SAT averages with school characteristics is the potential bias introduced by the selectivity of the test-taking subpopulation. The first model in Table 4 includes only a single explanatory variable, an inverse Mill's ratio computed from the school SAT participation rate (Gronau, 1974; Heckman, 1979). The coefficient on this variable is large and negative, indicating that schools with higher participation rates also have substantially higher scores. This contradicts expectations about the sign of selectivity bias—one expects that test-takers earn higher scores than would non-takers—and suggests that there are important omitted factors that vary across schools and influence both test participation rates and scores.

Column B adds to the model controls for the racial composition of the school. Schools with high fractions of blacks and Hispanics have substantially lower SAT scores than do schools that are largely white. Inclusion of these variables shrinks the selection coefficient substantially, though it is still significantly negative. Column C further adds the average education of parents of the students

at the school, measured on a 1-5 scale.[15] This, too, is highly significant, indicating that students at a school whose parents are all college graduates earn SAT scores 130 points higher than do those from schools where parents all attended college but failed to graduate. Inclusion of parental education has dramatic effects on the racial composition coefficients, and eliminates the effect of the school participation rate.

Finally, Columns D and E add the "School Characteristics Index" (SCI) and "Academic Performance Index" (API). The API is the score used by the California Department of Education in its school accountability program, computed from the distribution of scores on standardized exams given to all students in California schools. The SCI is computed as the predicted value from a regression of the API score on the full set of demographic characteristics collected by the California Department of Education (Technical Design Group, 2000). APIs are scaled to range from 200 to 1000, while SCIs range from 1 to 2.

Both the SCI and API are important predictors of school SAT scores. When the SCI is included, in Column D, the Mill's ratio coefficient finally becomes significantly greater than zero, as expected. It becomes even larger when the API is added as well, in Column E.

Of course, there is an important difference between the API score and the other measures, as this variable—unlike parental education, for example—plausibly measures the school's quality as well as its demographic composition. It is worth noting, however, that the addition of the API score to the model does not eliminate the effect of the demographic characteristics, suggesting that at least a portion of the association between SAT scores and demographics cannot be attributed to school quality (at least as measured by APIs).

---

[15] The parental education variable was collected in 1999 as part of the California school accountability program via a take-home survey. There are thus two important sources of error: First, response rates were less than perfect, with a response rate of about 80% at the average high school. Second, a school's parental education may have changed between 1993 and 1999. (The latter would be a problem as well for the school racial composition, as this too is from 1999.) Either type of error would be expected to attenuate the coefficient.

Table 5 explores the impact of including these school-level variables in the FGPA prediction model. Columns A and B repeats the specifications from Columns A and C of Table 3. Columns C and D report the same specifications, this time estimated only on the subsample of students from California public high schools (where the demographic variables are available). There is little difference between the two pairs of columns. The remaining columns add the school demographic controls and, ultimately, the API score to the model. The inclusion of the school racial composition reduces the coefficient on the school SAT average by about one quarter.[16] Additional demographic characteristics have relatively little impact, and are generally insignificant. On the other hand, when the school API score is included (in Column H), it is has significant predictive power for FGPAs, and it causes the school mean SAT coefficient to decline by an additional 5%.

The results in Table 5 indicate that a substantial portion of the school average SAT's predictive power for FGPAs derives from its association with the school racial composition. A substantially smaller additional portion derives from a component that is also reflected in the API score, though the API score retains substantial predictive power on its own.

VII. Demographic Variation and Underprediction

A long literature (Young, 2001, is a recent example) shows that race-blind models tend to overpredict the performance of black students, who have lower average SAT scores than white students and who tend to earn even lower relative FGPAs than would be predicted based on their SAT scores. There is a close connection between this "underprediction" and the analyses conducted here. It is straightforward to show that underprediction is equivalent to the result that between-race SAT gaps have stronger associations with FGPAs than do within-race SAT differences. That is,

[16] A natural concern is that the school racial composition variables are proxying for individual race. Rothstein (2004) explores models that include both the individual race and the school composition; both have predictive power for FGPAs.

whenever group A has lower average SAT scores than group B, group A's performance will be underpredicted by a race blind model if and only if the SAT coefficient declines when a group indicator is added as a predictor variable.[17]

One implication of this result, in combination with the analyses shown earlier, is that standard approaches will tend to find evidence of overprediction of FGPAs for students from schools with below-average SAT scores or with demographic characteristics that are typically associated with below-average scores. Table 6 presents a demonstration of this. I first estimate a standard validity model, including only SAT scores and HSGPAs. I then compute residuals from this model—individual students' over- and under-performance—and regress them on school average SAT scores and on school demographic characteristics. Positive coefficients in this latter regression indicate that FGPAs are over-predicted for students with low values of the indicated variable and underpredicted for students with high values, holding all other variables constant. Table 6 thus indicates that standard models that do not take account of demographic variables systematically overpredict the FGPAs of students from schools with low average SAT scores, with high fractions black or Hispanic; with low fractions other races; and with low API scores. By contrast, the individual SAT coefficient is negative, indicating overprediction for students whose SAT scores exceed the averages at their schools.

VIII. Conclusion

The results presented here indicate that SAT scores are substantially more predictive of eventual student performance across high schools than within.[18] In other words, the average SAT

[17] A proof of this assertion is available upon request.

[18] A natural explanation for this result might be that the school average SAT score is more reliable than the individual score, so that inclusion of the average leads to substantially greater attenuation of the effect of the individual score. This effect, while certainly present, cannot account for the results: As indicated in Table 2, only one quarter of the variation

score at a student's school is substantially more informative about that student's eventual FGPA than is the student's own score. An admissions office that lacked individual SAT scores but used school mean scores (or good proxies for them) would not suffer greatly in the accuracy of its predictions. On the other hand, an admissions office that ignored school mean scores would tend to over-predict the performance of students from low-scoring schools and under-predict the performance of those from high-scoring schools.

One might conclude from this that colleges should give preferences to students from high-scoring schools. This would amount to penalizing "strivers," students whose scores are higher than one would expect given their scores, relative to other students from more advantaged backgrounds with similar test scores. Indeed, the optimal penalty is quite large: A university choosing between applicant A and B, where A's school has average scores 10 points larger than B's school, should prefer A even if A's own score is 20 points *lower* than B's.

The predicted-performance-maximizing admissions rule amounts to affirmative action for socioeconomically advantaged students. Whether it represents desirable policy depends on one's view about the appropriate role of school demographic characteristics in admissions decisions. A non-trivial portion of the SAT's across-school predictive power is due to its association with the school racial composition: Students from high-minority-share high schools earn lower FGPAs than do students from primarily white high schools, and the school average SAT score acts in part as a proxy for racial composition in models that exclude the latter. Colleges that prefer not to discriminate against applicants from diverse high schools may prefer not to take advantage of the portion of the SAT average's predictive power that derives from its association with racial composition variables, as to do so permits the SAT to "launder" the impermissible variables.

---

of SAT scores is across schools, placing a fairly high lower bound on the reliability of the individual deviation from the school mean.

Regardless of the college's goals for admissions, the individual SAT score itself is less important than is implied by existing research. The school average SAT, despite its association with school demographic characteristics, contains much of the information needed for predicting FGPAs. Schools may wish to reduce the weight that they place on individual SATs, relying primarily on HSGPAs for within-school comparisons.

References

Barro, Robert J., 2001. "Economic Viewpoint: Why Colleges Shouldn't Dump the SAT," *Business Week*, April 9, 20.

Breland, Hunter M., 1979. *Population Validity and College Entrance Measures*, College Entrance Examination Board, New York.

Camara, Wayne J., 2001. "There is No Mystery When it Comes to the SAT I," *College Board News*.

Camara, Wayne J., and Gary Echternacht, 2000. "The SAT I and High School Grades: Utility in Predicting Success in College," Research Report RN-10, College Entrance Examination Board, New York.

College Board, 2003. "Test Characteristics of the SAT: Reliability, Difficulty Levels, Completion Rates." Downloaded from http://www.collegeboard.com/prod_downloads/about/news_info/cbsenior/yr2003/pdf/table_5.pdf on June 27, 2005.

Deaton, Angus, 1997. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*, Johns Hopkins University Press / The World Bank, Baltimore, Maryland.

Elliot, Rogers, and A. Christopher Strenta, 1988. "Effects of Improving the Reliability of the GPA on Prediction Generally and on Comparative Predictions for Gender and Race Particularly," *Journal of Educational Measurement* 25, 333-47.

Goldman, Roy D. and Melvin Widawski, 1976. "A Within-Subjects Technique for Comparing College Grading Standards: Implications in the Validity of the Evaluation of College Achievement," *Educational and Psychological Measurement* 36, 381-90.

Gronau, Reuben, 1974. "Wage Comparisons--a Selectivity Bias," *The Journal of Political Economy* 82 (6), 1119-1143.

Heckman, James J., 1979. "Sample Selection Bias as a Specification Error," *Econometrica* 47, 153-161.

Marcus, Amy Dockser, 1999. "New Weights Can Alter SAT Scores—Family is Factor in Determining Who's a 'Striver,'" *Wall Street Journal*, August 31, B1.

McWhorter, John H., 2001. "Eliminating the SAT Could Derail the Progress of Minority Students," *Chronicle of Higher Education*, March 9, B11-B12.

Ramist, Leonard, Charles Lewis and Laura McCamley-Jenkins, 1993. "Student Group Differences in Predicting College Grades: Sex, Language, and Ethnic Groups." Research Report 93-1, College Entrance Examination Board, New York.

Rothstein, Jesse, 2004. "College Performance Predictions and the SAT," *Journal of Econometrics* 121, 297-317.

Sander, Richard, 2005. "A Systemic Analysis of Affirmative Action in American Law Schools," *Stanford Law Review* 57 (2) 367-484.

Technical Design Group, 2000. "Construction of California's 1999 School Characteristics Index and Similar Schools Ranks," PSAA Technical Report 00-1, April. http://www.cde.ca.gov/ta/ac/ap/documents/tdgreport0400.pdf.

Vigdor, Jacob L., and Charles T. Clotfelter, 2003. "Retaking the SAT," *Journal of Human Resources* 38 (1), January, 1-33.

Willingham, Warren W., 1985. *Success in College: The Role of Personal Qualities and Academic Ability*. College Entrance Examination Board, New York.

Willingham, Warren W., and Hunter M. Breland, 1982. *Personal Qualities and College Admissions*. College Entrance Examination Board, New York.

Young, John W., 1993. "Grade Adjustment Methods," *Review of Educational Research* 63, 151-65.

Young, John W., 2001. *Differential Validity, Differential Prediction, and College Admissions Testing*. College Board Research Report 2001-6.

Table 1. Basic validity models on full sample and eligible subsample

	Full Sample			UC-eligible subsample		
	(A)	**(B)**	**(C)**	**(D)**	**(E)**	**(F)**
SAT / 1000		1.36	0.94		1.28	0.94
		(0.03)	(0.03)		(0.03)	(0.03)
HS GPA	0.62		0.51	0.67		0.57
	(0.01)		(0.01)	(0.01)		(0.01)
N	18,711	18,711	18,711	17,504	17,504	17,504
R^2	0.19	0.14	0.24	0.17	0.13	0.22

Notes: Dependent variable is freshman GPA. All models include campus and major (as of the freshman year) fixed effects. Statistics are uncorrected for restriction of range. Sample in columns D-F excludes students who appear to be ineligible according to the UC published rules (and who therefore appear to have been admitted "by exception").

Table 2. Across- and within-school share of variance of SATs, HSGPAs, and FGPAs

	UC students, eligible subsample			All SAT-takers
	FGPA	HS GPA	SAT	SAT
	(A)	(B)	(C)	(D)
Standard deviation	0.63	0.39	169	225
S.D. of school mean (indiv. level)	0.23	0.14	88	112
S.D. of deviation from school mean	0.58	0.36	144	195
Between-school share of variance	14%	13%	27%	25%
Within-school share of variance	86%	87%	73%	75%

Table 3. Validity models on eligible subsample, distinguishing within- and across-school variation

	School averages are computed over:							
			Est. sample	Eligible applicants who enroll at UC	Eligible UC applicants	All UC applicants	Eligible SAT-takers	All SAT-takers
	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
SAT / 1000	0.94	0.50	0.49	0.49	0.49	0.53		
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)		
HS GPA	0.57	0.73	0.72	0.73	0.72	0.68		
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)		
Mean SAT at school / 1000			1.28	1.27	1.33	1.34		
			(0.06)	(0.06)	(0.06)	(0.06)		
Mean GPA at school			-0.63	-0.64	-0.72	-0.45		
			(0.03)	(0.03)	(0.04)	(0.03)		
School FEs	n	y	n	n	n	n		
N	17,504	17,504	17,504	17,504	17,504	17,504		
R^2	0.22	0.34	0.26	0.26	0.26	0.25		
SAT / 1000			0.59	0.59	0.59	0.62	0.64	0.67
			(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
HS GPA			0.62	0.62	0.62	0.62	0.63	0.62
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Mean SAT at school / 1000			1.11	1.12	1.16	1.05	0.89	0.86
			(0.06)	(0.06)	(0.06)	(0.05)	(0.04)	(0.05)
School FEs			n	n	n	n	n	n
N			17,504	17,504	17,504	17,504	17,251	17,260
R^2			0.24	0.24	0.24	0.24	0.24	0.24

Notes: All models include campus and major fixed effects. Statistics are uncorrected for restriction of range. Sample in all columns excludes students who are not UC-eligible.

Table 4. Models for school average SAT scores

	(A)	(B)	(C)	(D)	(E)
Inverse Mills ratio	-179.0	-82.0	16.6	29.4	55.4
	(12.2)	(9.0)	(8.8)	(8.9)	(9.4)
Fraction Black		-372.1	-224.7	-78.8	-36.9
		(19.1)	(17.2)	(30.7)	(30.3)
Fraction Hispanic		-266.4	-1.8	50.5	63.4
		(9.9)	(15.8)	(18.0)	(17.5)
Fraction other race (Asian, Native		-41.8	45.4	12.4	14.9
American, etc.)		(14.5)	(12.5)	(13.6)	(13.1)
Average parental education			130.1	62.5	49.6
			(6.7)	(13.6)	(13.2)
School Characteristics Index				493.8	308.4
				(86.7)	(87.9)
Academic Performance Index					0.38
					(0.05)
N	723	723	723	723	723
R^2	0.23	0.70	0.80	0.81	0.82

Notes: Dependent variable is the school average SAT score, computed over all test-takers at the school. Sample consists of public schools in California with non-missing data. Regressions are weighted by the number of SAT-writers at the school. Inverse Mill's ratio is computed from the measured school SAT participation rate.

Table 5. School demographics and school average scores as predictors of collegiate grades

	Full sample		CA Public schools					
	(A)	**(B)**	**(C)**	**(D)**	**(E)**	**(F)**	**(G)**	**(H)**
Individual SAT / 1000	0.94	0.59	0.92	0.58	0.58	0.58	0.58	0.58
	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Individual HSGPA	0.57	0.62	0.58	0.63	0.64	0.64	0.64	0.64
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
HS: Mean SAT in sample /		1.11		1.17	0.86	0.88	0.87	0.83
1000		(0.06)		(0.06)	(0.08)	(0.09)	(0.09)	(0.09)
HS: Fraction Black					-0.20	-0.21	-0.11	-0.10
					(0.05)	(0.05)	(0.08)	(0.08)
HS: Fraction Hispanic					-0.11	-0.13	-0.10	-0.09
					(0.03)	(0.04)	(0.05)	(0.05)
HS: Fraction other race (Asian,					0.17	0.16	0.13	0.12
Native American, etc.)					(0.03)	(0.03)	(0.03)	(0.03)
HS: Average parental						-1.24	-6.18	-7.39
education / 100						(1.64)	(3.57)	(3.59)
HS: School Characteristics							0.34	0.00
Index							(0.22)	(0.24)
HS: Academic Performance								0.53
Index / 1000								(0.13)
N	17,504	17,504	14,141	14,141	14,141	14,141	14,141	14,141
R^2	0.22	0.24	0.22	0.24	0.25	0.25	0.25	0.25

Notes: All models include campus and major fixed effects. Statistics are uncorrected for restriction of range. Sample in all columns excludes students who are not UC-eligible; columns C-H additionally exclude students from private or non-California high schools.

Table 6. School-level variables and over-/under-prediction

	Full sample		CA Public schools					
	(A)	**(B)**	**(C)**	**(D)**	**(E)**	**(F)**	**(G)**	**(H)**
Individual SAT / 1000	0.00	-0.32	-0.02	-0.33	-0.33	-0.33	-0.33	-0.33
	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Individual HSGPA	-0.06	-0.02	-0.06	-0.02	-0.01	-0.01	-0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
HS: Mean SAT in sample /		1.01		1.05	0.74	0.78	0.77	0.73
1000		(0.05)		(0.06)	(0.08)	(0.09)	(0.09)	(0.09)
HS: Fraction Black					-0.19	-0.21	-0.11	-0.10
					(0.05)	(0.05)	(0.08)	(0.08)
HS: Fraction Hispanic					-0.11	-0.15	-0.11	-0.11
					(0.03)	(0.04)	(0.05)	(0.05)
HS: Fraction other race (Asian,					0.15	0.14	0.11	0.10
Native American, etc.)					(0.03)	(0.03)	(0.03)	(0.03)
HS: Average parental						-1.79	-6.71	-7.75
education / 100						(1.63)	(3.55)	(3.57)
HS: School Characteristics							0.34	0.03
Index							(0.22)	(0.24)
HS: Academic Performance								0.46
Index / 1000								(0.13)
N	18,690	18,690	15,044	15,044	15,044	15,044	15,044	15,044
R^2	0.00	0.02	0.00	0.02	0.03	0.03	0.03	0.03

Notes: Dependent variable in every column is the residual FGPA from a prediction model that is estimated on UC-eligible students and includes HSGPA, SAT, and campus and major fixed effects. Samples for reported regressions are not restricted to UC-eligible students. Samples in columns C-H, however, exclude students from private or non-California high schools.