Research & Occasional Paper Series: CSHE.3.02

CSHE Center for Studies in Higher Education

UNIVERSITY OF CALIFORNIA, BERKELEY
http://ishi.lib.berkeley.edu/cshe/

ADMISSIONS BIAS: A NEW APPROACH TO VALIDITY ESTIMATION IN SELECTED SAMPLES*

April 2002

Jesse M. Rothstein*
Department of Economics
549 Evans Hall #3880
University of California, Berkeley, CA 94720
jrothst@econ.berkeley.edu

ABSTRACT

Validity researchers typically work with nonrandom samples, membership in which depends in part on the exam score being investigated. In a study of the SAT's validity for freshman GPA at a particular college, for example, FGPA is not observed for the entire pool of potential applicants, but only for those who are admitted and enroll. This sample selection biases validity estimates. Corrections for restriction of range remedy the problem only when the exam score is the sole determinant of selection, and even then do not permit consistent estimation of the exam's incremental validity. Regression omitted variables results motivate a proposed validity estimator that is consistent whenever the determinants of selection are observed. An algorithm is suggested for calculation of these "robust" validity coefficients, and used to estimate the SAT's validity at the University of California. The usual validity estimates are shown to be substantially biased, in the hypothesized directions, by admissions-induced sample selection.

1. INTRODUCTION

Researchers frequently use observational samples to estimate the predictive validity of an exam score. The exam score in question commonly plays a direct role in the process by which test-takers enter the sample. Many SAT validity studies, for example, use students attending a single college; we are of course most interested in the SAT's validity at selective colleges that consider SAT scores in admissions. In situations like this, the within-sample validity coefficient may not be an appropriate estimate of the population validity.

* I thank the Center for Studies in Higher Education at the University of California, Berkeley, and a National Science Foundation Graduate Student Fellowship for research support. I am grateful to David Card, David Lee, Justin McCrary, Jack Porter, and seminar participants at Berkeley and Princeton for helpful suggestions. This paper extends results that were first introduced in Rothstein (2002).

In an effort to fix this problem, researchers frequently correct validity coefficients "for restriction of range" (Camara and Echternacht, 2000; Willingham et al., 1990; Breland, 1979). However, validity estimates with the usual Pearson-Lawley correction (Gulliksen, 1950; Bridgeman et al., 2000) are free of selection bias only under very specific, infrequently satisfied assumptions about the selection process. Namely, the sample must be selected purely on the basis of the exam score itself; other variables that might have independent predictive power for the outcome in question must not enter into the selection process. The situation is even worse when the parameter of interest is the exam's incremental validity. Here, range-corrected validities are inappropriate whenever the exam score is even a partial determinant of selection.[1]

An analogy is perhaps helpful. Consider the validity of height for the prediction of performance on the basketball court. If we could measure performance for a representative sample of the population, it would surely be strongly correlated with players' height. But suppose that we were instead to estimate the correlation on a sample of National Basketball Association players. This sample would certainly have a much smaller range of heights than the population as a whole, so one might think that the range-corrected within-sample estimate would be appropriate. But a consideration of the sample selection reveals that the problems do not end with restricted range of the predictor.

Individuals can only enter the NBA sample if they have high "basketball performance." In particular, short players only make it to the NBA if their athleticism and native ability are so exceptional as to make up for the disadvantage conferred by their height. Tall players are not so strongly selected for native ability, as they do not have to overcome a height disadvantage. Thus, it seems likely that mean native ability is higher among short NBA players than among their tall teammates. This would lower the correlation between height and performance in the sample. Indeed, it might be that height has zero predictive power within the selected NBA sample, but a great deal of power in the population. The problem is fundamentally one of omitted variables: Estimates of height's validity or incremental validity that do not take into account the interaction of height with native ability in sample selection will not be informative about the population relationship.

I propose here an alternative approach to the estimation of validity coefficients in samples selected partly on the basis of the score to be evaluated. The new approach embeds validity estimation in a linear regression framework, and uses omitted variables results to obtain validity coefficients that are robust to more general selection assumptions than are required for traditional validity and incremental validity estimates. My approach can be more demanding in terms of data availability than is the traditional approach, but even when not all of the required data are available it suggests a useful diagnostic for the presence of serious selection problems in traditional validity estimates.

My critique of traditional validity estimates and my alternative approach are applicable in many psychometric contexts. For expository purposes, however, I focus on an illustrative example, SAT-based predictions of freshman grade point average (FGPA) at the University of California.[2]

[1] By "inappropriate," I mean biased in the limit, or inconsistent: the estimate will not converge to the population value even as the sample size goes to infinity. Mangling ordinary usage, I sometimes call a consistent estimator "unbiased." The finite-sample properties—to which "unbiased" ordinarily refers—of correlation coefficients are quite complex and beyond the scope of this paper (van der Vaart, 1998, pg. 30-31).

[2] Here and throughout, I use "SAT" to refer interchangeably to the pre-1995 SAT exam and to the post-1995 SAT I exam.

I consider the estimation of two parameters that are of frequent interest: the SAT's univariate validity and its incremental validity when baseline predictions consider only the high school grade point average (HSGPA).

The University of California (UC) has several attractive properties for my purposes. In particular, the UC system's exact eligibility formula is published annually and, for the 1993 entering class considered here, is based almost exclusively on a student's HSGPA and SAT. This means that—subject to the maintained assumption that sample selection comes purely through the admissions decision and not through students' application and enrollment decisions—I can observe the distribution of all determinants of sample selection in both the sample and the population, reducing what is ordinarily intractable "selection on unobservables" to more manageable "selection on observables." This is precisely what my estimation strategy requires.

I find that traditional estimation techniques applied to an appropriate sample understate the SAT's uncontrolled validity but overstate its incremental validity, by ten and seven percent respectively, for University of California FGPAs. I see no reason to believe that the example used here is unique; it seems likely that traditional validity estimates from other colleges and contexts are similarly erroneous.

The paper proceeds as follows. In Section 2, I introduce the University of California data and discuss the UC admissions process. I also present SAT validity estimates for this sample using the traditional methodology. In Section 3, I provide evidence that admissions rules bias univariate validity coefficients downward and incremental validity coefficients upward. Neither form of bias is repaired by the usual restriction of range corrections. Section 4 outlines an algorithm, derived from regression omitted variables results, for calculating validity coefficients that are free of admissions-induced selection bias. Section 5 applies the new approach to the UC data, confirming the presence of serious selection bias in estimates derived from the usual methodology. Section 6 concludes.

2. ADMISSIONS AT THE UNIVERSITY OF CALIFORNIA

The data used in the present analysis are extracted from a University of California administrative database and describe the class of in-state freshman matriculants in the 1993-1994 academic year.[3] I have individual level observations on all 22,526 students in this cohort; variables include high school grades and SAT scores as well as longitudinal information about students' collegiate academic progress. I compute a freshman GPA for each student based on the first year for which that student registered in any courses and discard observations with missing data, leaving a sample of 18,587 students.[4]

Table 1 displays validity coefficients for this sample, estimated using the traditional methodology

[3] I am grateful to Saul Geiser and Roger Studley of the UC Office of the President for providing me access to the data, which are not publicly available. One shortcoming of the current data relative to those used in other validity studies is that only the composite SAT score is reported, not the separate math and verbal components. Thus, I depart from common practice and analyze only the composite SAT's validity.

[4] One notable subset of observations which must be discarded corresponds to students at the Santa Cruz campus, where letter grades are optional and infrequently assigned. If Santa Cruz students differ systematically from the typical UC student, this may introduce selection bias even in my proposed algorithm. Rothstein (2002) estimates campus-specific validity coefficients that are not subject to this bias, and finds no evidence that it is a major problem.

and corrected for restriction of range in the usual way. This range correction requires information about the distribution of SAT scores in the population of interest. I take the relevant population to be all California SAT-takers, and calculate the distribution from a College Board data set of all California SAT-takers from the 1994-1998 cohorts.[5]

Panel A pools all UC students into a single sample, imposing the assumption that grading standards are identical across campuses and majors within the system. In Panel B, I display validity estimates for an adjusted FGPA that permits grading standards to differ across UC campuses and majors.[6] Other researchers have used preferable (but more data intensive) adjustment mechanisms (Keller et al., 1994; Young, 1990; Braun and Szatrowski, 1984a,b), but the one used here seems adequate for the example at hand, in which the campuses are all part of the same system and may, therefore, have similar grading standards. The table shows that validities rise when major and campus are controlled, suggesting that low-SAT and low-HSGPA students are disproportionately assigned to campuses and majors with easier-than- average grade scales. Unless otherwise noted, all further validity estimates use the adjusted FGPA as the variable to be predicted.

Table 1: Validity estimates using usual methodology, entire UC sample

	Within Sample	Corrected for Restriction of Range
Panel A: Treating all UC FGPAs as directly comparable		
Single variable		
SAT	0.358	0.438
HSGPA	0.430	0.562
Bivariate: SAT+HSGPA	0.470	0.600
SAT incremental	0.040	0.037
Panel B: FGPAs adjusted for campus and freshman major		
Single variable		
SAT	0.413	0.498
HSGPA	0.482	0.617
Bivariate: SAT+HSGPA	0.531	0.661
SAT incremental	0.050	0.044

Note: Sample includes 18,568 observations from seven UC campuses. See text (note 6) for description of "adjusted" FGPA.

[5] I am grateful to David Card, Alan Krueger, the Mellon Foundation, and the College Board for permitting me access to these proprietary data. A minor problem is introduced by the recentering of SAT scores in 1995. I use the College Board's crosswalk tables (College Board, 2002) to obtain a pre-recentering score that approximately corresponds to each student's recentered score. Since these data are used only to obtain the joint population distribution of SAT and HSGPA, imprecision in the crosswalk tables should not affect my results. All validity estimates in this paper should be understood to apply to the pre-1995 SAT scoring system.

[6] This is implemented by including fixed effects for each campus and each major group in a regression model along with both SAT and HSGPA, then subtracting the fixed effect coefficients from the raw FGPA to compute an adjusted FGPA. In effect, this denies predictive “credit” to SAT and HSGPA for their ability to predict campus and major.

The SAT validity coefficients reported in Table 1 are somewhat smaller than those seen in other studies, but not outside the usual range (Bridgeman et al., 2000). However, because the sample consists only of students attending the UC, even range-corrected validity coefficients are potentially subject to selection bias introduced by the UC's admissions criteria. This would be true for a sample drawn from any selective college, but the UC's admissions criteria are unusually simple and transparent (at least at the systemwide level that concerns me here). As this feature is the key to my estimation strategy, it is helpful to take a moment to describe the UC admissions process.

University of California admissions proceed in two-stages. The first is a systemwide eligibility determination, and is based on a few easily observed variables. Each year, the UC application packet includes a grid detailing the high school GPA and SAT score requirements for eligibility. Table 2 reproduces the 1993 grid. Students who complete a normal academic courseload in high school, who submit scores on all required exams, and who meet the published SAT-HSGPA thresholds are eligible for admission to the UC system; students who do not meet these requirements ordinarily are not (Student Academic Services, multiple years).[7]

Table 2: Minimum high school GPA and SAT scores for UC eligibility

If HSGPA is:	Minimum SAT score:	If HSGPA is:	Minimum SAT score:
>=3.30	elig. with any score	3.05	1050
3.29	490	3.04	1080
3.28	520	3.03	1100
3.27	540	3.02	1120
3.26	560	3.01	1150
3.25	590	3	1170
3.24	610	2.99	1190
3.23	630	2.98	1220
3.22	660	2.97	1240
3.21	680	2.96	1260
3.2	700	2.95	1290
3.19	730	2.94	1310
3.18	750	2.93	1330
3.17	770	2.92	1360
3.16	800	2.91	1380
3.15	820	2.9	1400
3.14	840	2.89	1430
3.13	870	2.88	1450
3.12	890	2.87	1470
3.11	910	2.86	1500
3.1	940	2.85	1520
3.09	960	2.84	1540
3.08	980	2.83	1570
3.07	1010	2.82	1590
3.06	1030	<2.82	not elig.

Source: University of California Office of the President: *Introducing the University*, 1993.

[7] This is a slightly simplified version of the process governing the 1993 admissions cycle that generated the data used here. Additional complexities in the actual rules—for example, that students with extraordinarily high SAT and Achievement (SAT II) scores are eligible even with low HSGPA—are ignored here and do not invalidate my conclusions. It is worth noting, however, that UC admissions rules have changed substantially since 1993, and the estimation strategy used here would need slight modification to incorporate these changes.

After the eligibility determination is made, the eight UC campuses make separate admissions decisions from among their eligible applicants. The campus admissions offices consider many more variables than are used in eligibility determination: student essays, recommendations, and personal background characteristics all play a role. However, all eligible applicants are guaranteed admission to at least one of the UC campuses; an eligible student who is denied admission by all of the campuses to which she applied is offered admission at a less competitive campus.[8] Ineligible students get no such guarantee and, in fact, can be admitted only "by exception"; such exceptions may not account for more than a few percent of any campus' admission offers.

3. SELECTION BIAS IN VALIDITY COEFFICIENTS

An accurate model of the campus admission processes would require unusually rich data on student characteristics that are ordinarily unobserved by the researcher. However, the eligibility determination that governs access to the system is more easily modeled. The most important eligibility criterion is the grade and test threshold. It seems reasonable to expect that most students who cross this threshold also meet the other eligibility requirements.[9] 17,346 students in my sample have sufficiently high grades and SAT scores to satisfy the test score criterion. Assuming that these students met the coursework and test-taking criteria, they were guaranteed admission to at least one of the UC campuses.[10] The remaining 1,241 students did not satisfy the test score criterion, and must therefore have been admitted "by exception," probably because they had other characteristics—athletic talent or an exceptional essay, perhaps—that called them to admissions officers' attention.

Table 3: Range-corrected validity estimates, sample split by eligibility status

	UC Eligible	UC Ineligible
N	17,346	1,222
Single variable		
SAT	0.479	0.235
HSGPA	0.626	0.205
Bivariate: SAT+HSGPA	0.674	0.416
SAT incremental	0.048	0.211

Note : Validities are for prediction of adjusted FGPA; see text (note 6) for definition.

3.1. Selection on Observables

We might expect students in the group of eligible applicants to be roughly typical of all (similarly-

[8] This system is made possible because some campuses—usually Riverside and Santa Cruz—admit essentially all eligible applicants.

[9] It is worth noting, however, that in the College Board SAT-taker data, many students whose SAT scores are above the eligibility threshold have not taken the Achievement—SAT II—exams that constitute another eligibility criterion (the Achievement scores themselves were not considered at the eligibility stage in 1993-1998; the student simply had to submit some scores).

[10] Unobserved variables *are* relevant to campus assignment within the system, even within this group. In this paper, I assume this complication away by treating the eight campuses as a single college, with grade scales that differ only by a location parameter across campuses. The approach taken here is extended to permit campus-specific validity coefficients in Rothstein (2002).

eligible) students with the same SAT scores and HSGPAs, but it seems unlikely that students in the ineligible group are similarly representative of the low-scoring population.[11]

If the unobserved characteristics that lead admissions officers to make exceptions to the eligibility rules are correlated with academic potential, the unrepresentativeness of the latter group might bias SAT validity coefficients. On average, the lower an ineligible admittee's SAT score, the better must have been the unobserved characteristics that recommended him for admission. Thus, we should expect that the SAT-FGPA relationship is weaker in the pool of ineligible students than among those who meet the eligibility criterion. Table 3, which displays SAT validity coefficients estimated separately for the two groups, confirms this. Only range-corrected coefficients are shown; uncorrected coefficients are not comparable across groups because the range of SAT scores among eligibles is much different than that among ineligibles. As expected, FGPA predictions are substantially less accurate in the ineligible pool than in the eligible pool.

This does not indicate that FGPAs are fundamentally less predictable for low-SAT, low-HSGPA students than for students with higher scores and grades. It occurs because the subset of low-scoring students who are admitted to the UC is less representative of the population of low-scoring students than the admitted subset of high-scoring students is of the population of high-scoring students. The inclusion of low-scoring students in the sample on which the validity coefficients of Table 1 were estimated biases validities, mostly downward. Hereafter, I exclude the ineligible students, restricting analysis to the subsample of UC-eligible students.

3.2. Selection on Unobservables

Above, I argued that the inclusion of students admitted on the basis of unobserved characteristics can bias validity coefficients. Less obviously, even admission on observed characteristics biases within-sample coefficients, in predictable directions. The argument is made more rigorously in Section 4, but the basic idea is best explained without mathematics: although students in the UC-eligible subsample may be presumed representative of the population of students with the same SAT scores and HSGPAs, they are not representative of the wider group of students with the same SAT scores: conditional on SAT score, high-HSGPA students are overrepresented and low-HSGPA students are underrepresented. Consider the population of students with SAT scores of exactly 900. The College Board data—describing the universe of California SAT takers in 1994 through 1998—include nearly 8,000 students with this score; their HSGPAs have mean 3.25 and standard deviation 0.55. The subsample of eligible UC students contains 205 with SAT scores of 900; their HSGPAs have mean 3.69 and a standard deviation of only 0.36. The reason for the disparity in HSGPA distributions is seen in Table 2: Students with 900 SATs need HSGPAs of at least 3.12 to be UC-eligible, so the eligibility rules omit 900-SAT students with low HSGPAs.

Because HSGPA has independent predictive power for FGPA, we should expect the UC-eligible students with SAT scores of 900 to achieve higher average FGPAs than would have been seen if the UC had admitted all applicants with 900 SAT scores. If the UC sample were equally unrepresentative of the population at all levels of the SAT distribution, the sample selection

[11] The former claim relies on an assumption that students' decisions to apply to and enroll at the UC are uninformative about their academic ability. If this assumption—maintained throughout but unverifiable in the data used here—is violated, validity coefficients could be biased either up or down, depending on the direction and magnitude of the induced selection pressure at different points in the SAT distribution. Of course, traditional validity estimates would be subject to the same bias.

described here would not necessarily bias the SAT validity coefficient. However, the UC eligibility rules—like all admissions processes—allow high SAT scores to offset low HSGPAs, making the HSGPA selection more stringent at lower SAT scores. Students with 1500 SAT scores are UC eligible with HSGPAs of 2.86 or higher; 99 percent of students at or above this SAT score are UC-eligible. However, students with SAT scores below 490 must have HSGPAs of at least 3.3 to be eligible, a requirement that only 15 percent can meet. In between, 62 percent of students with SAT scores between 900 and 950 are eligible. Thus, UC students are less representative of the population at the bottom of the SAT distribution than at the top.

Figure 1A: Conditional Expectation of HSGPA Given SAT

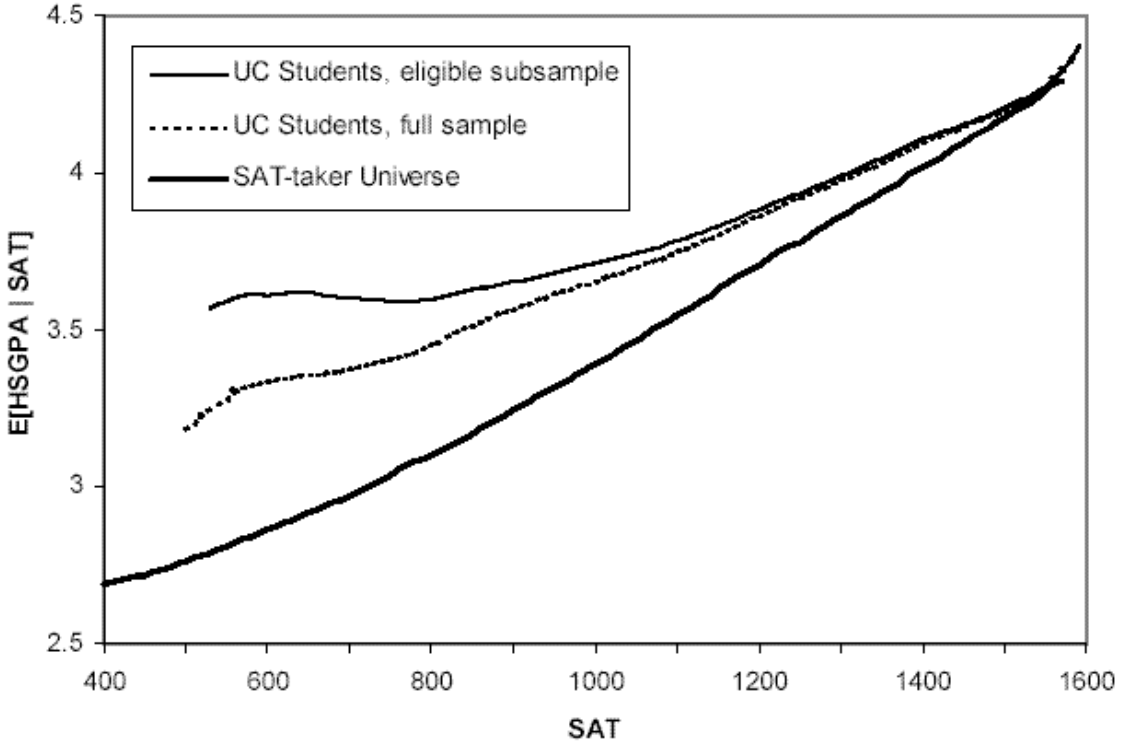

Figure 1A displays this argument in graphical form. It shows the relationship between SAT scores and conditional mean HSGPA in three groups: The population of SAT-takers in California, the original sample of all UC students, and the subsample of UC students who appear to be UC-eligible. The figure shows that low-SAT students in the UC samples have substantially higher HSGPAs, on average, than do typical low-SAT students, and that the disparity is largest at the lowest SAT scores. This tends to flatten the relationship between FGPA and SAT in the UC sample, depressing the measured SAT validity.

It is worth emphasizing that this is a different problem than the bias introduced by the inclusion of ineligible students discussed in Section 3.1: That bias came from the consideration of unobserved variables in admissions decisions, while this problem arises because an observed variable that was considered in eligibility and admissions is not accounted for in the partial correlation of SAT scores with FGPAs. The bias indicated by Figure 1A would exist at any selective college that considered characteristics other than the SAT score itself in admissions.

Figure 1B: Conditional Expectation of SAT Given HSGPA

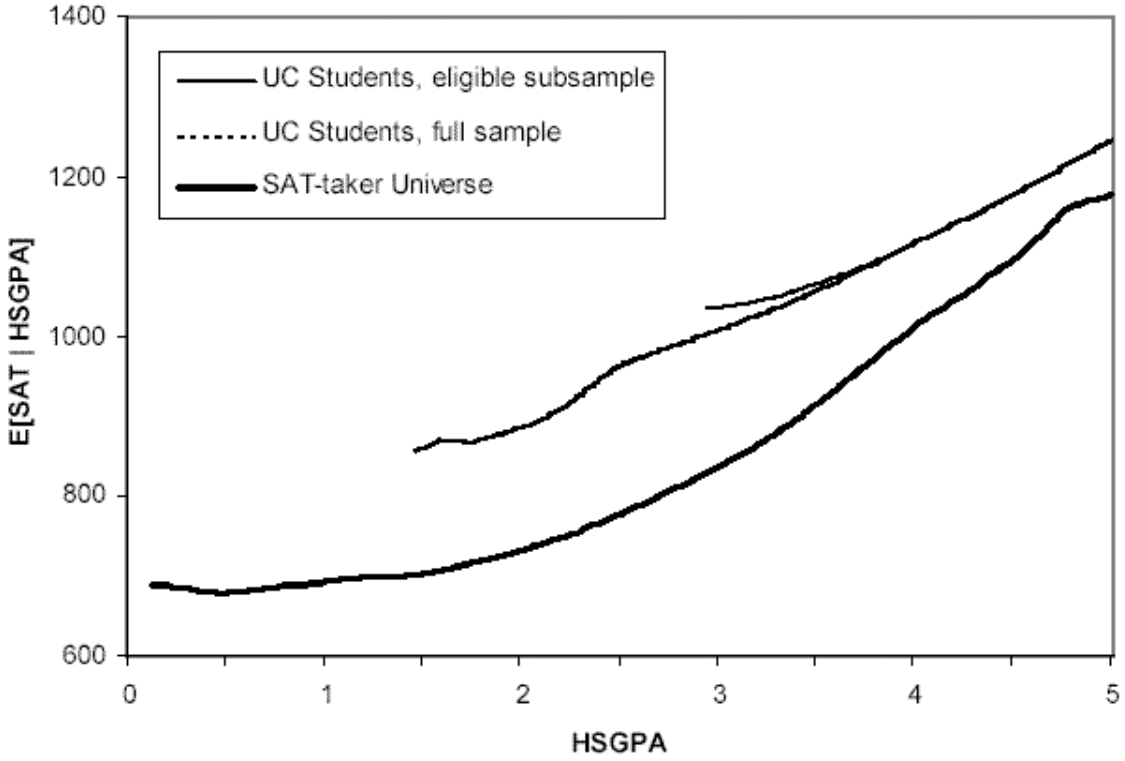

Figure 1B provides a different view of the data, displaying the relationship between HSGPAs and conditional mean SAT scores. UC students at low HSGPA levels are unrepresentative of the population of low HSGPA SAT-takers: their SAT scores are much higher than is typical in the population. This is simply the converse of the phenomenon discussed earlier; the UC eligibility rules allow high SAT scores to make up for low HSGPA. However, it means that the HSGPA validity measured in the UC sample is also biased downward. This is not important if the SAT's validity is the only parameter of interest, but it does bias estimates of the SAT's incremental validity. Incremental validity coefficients are estimated by subtracting the HSGPA validity coefficient from the joint validity coefficient of HSGPA and SAT, taken together. Unless bias in the latter exactly offsets bias in the former, the SAT's incremental validity is biased.

In fact, the joint validity coefficient is not biased when estimated on the UC-eligible subsample and corrected for restriction of range. To see this, consider students with 1000 SAT scores and 3.50 HSGPAs. All students of this sort are UC eligible. If the UC were to admit the 63 students with these qualifications in the College Board data—as, in fact, it probably did, although I cannot match College Board observations to UC outcomes—there is no reason to think that their mean FGPA would be any different from the 3.48 seen among the ten students in the UC data with the same qualifications. This is selection on observables: As long as HSGPA and SAT are the only variables determining selection into the analysis sample, a regression of FGPA on both HSGPA and SAT yields unbiased coefficient estimates (Wooldridge, 2000, pg. 299).

The selection-on-observables result guarantees only that the multiple regression coefficients are unbiased. The restricted range of HSGPA and SAT in the UC sample still biases uncorrected validity estimates downward. However, this is exactly the situation for which the usual "restriction of range corrections" are designed: given unbiased regression coefficients, they

produce an unbiased corrected validity coefficient (Gulliksen, 1950). Thus, the UC-eligible subsample permits unbiased estimation of the *joint* validity.

It is important to note that range corrections do not correct the bias in the HSGPA or SAT univariate validity coefficients. The argument is spelled out in greater detail in Rothstein (2002), but for the present purposes it suffices to point out that a regression of FGPA on either HSGPA or SAT separately yields coefficients downward-biased by selection-on-unobservables (the omitted variable is observed in the data, but because it is not controlled for in the regression the bias is not avoided). Restriction-of-range corrections only magnify bias in regression coefficients, so even range-corrected univariate validity estimates are downward-biased. It follows from this that SAT's incremental validity is upward-biased.

4. A MORE ROBUST APPROACH

The above discussion suggests an estimator of the SAT's univariate validity that is free of the selection bias shown above. The problem, essentially, is to estimate how much of an FGPA gap one should expect to see between a randomly-selected student with an 1100 SAT score and another with a SAT score 100 points lower. The estimation approach implied by traditional validity research is to regress FGPA on SAT in the sample, estimating the mean gap at one hundred times the SAT's coefficient. The FGPA gap implied by the UC-eligible SAT validity estimate in Column A of Table 3 is 0.14 GPA points.[12]

However, HSGPA's role in sample selection, coupled with its independent predictive power for FGPA, means that this is not a reasonable estimate of the population FGPA gap between students 100 SAT-points apart. Conceptually, the gap can be decomposed into two parts: one arising from the different HSGPAs we expect the two students to have and one reflecting the SAT's incremental predictive power. As Figure 1A shows, the population HSGPA gap between the two students is larger than that seen in the sample, so the first part of the decomposition is understated by the traditional approach. Using the methods introduced below, I find that the expected FGPA gap between two students with SAT scores 100 points apart is 0.18 GPA points, nearly one-quarter larger than that implied by the within-sample estimate.

To express the above argument more formally, it is useful to turn to a regression model of the relationship between predictors and outcome. Let X_i and Z_i (SAT and HSGPA) be two predictors of an outcome y_i (FGPA) for individual i. Assume that in the population the conditional expectation of y_i given X_i and Z_i satisfies

$$\mathrm{E}\left[y_i \middle| X_i, Z_i\right] = \alpha_0 + X_i\beta_0 + Z_i\gamma_0 \tag{1}$$

This implies that when we condition only on i X, as when we are estimating X's univariate validity,

$$\mathrm{E}\left[y_i \middle| X_i\right] = \alpha_0 + X_i\beta_0 + \mathrm{E}\left[Z_i \middle| X_i\right]\gamma_0 = \alpha_1 + X_i\beta_1 \tag{2}$$

where

$$\alpha_1 = \alpha_0 + \left(\mathrm{E}[Z] - \mathrm{E}[X]\,\rho_{XZ}\sqrt{\frac{\mathrm{var}(Z)}{\mathrm{var}(X)}}\right),$$

[12] The validity coefficient is a transformation of the regression slope; the exact relationship is discussed below.

$$\beta_1 = \beta_0 + \rho_{XZ}\sqrt{\frac{\text{var}(Z)}{\text{var}(X)}}\gamma_0 \quad , \tag{3}$$

and ρ_{XZ} is the correlation between X and Z (Wooldridge, 2000, pg. 89). The crucial parameter is β_1.[13] The traditional approach implicitly uses the within-sample correlation between X and Z (0.38 in the UC-eligible data) and within-sample variances in (3). The essence of my approach is to use the population correlation (0.52 in the College Board data) and variances in place of their sample analogues. Thus, the proposed algorithm, incorporating traditional range corrections and new omitted variables corrections, is as follows:

4.1. An algorithm for validity estimates without selection bias

1. Regress y on X and Z in the analysis sample. Extract from this regression estimates of α_0, β_0, γ_0, and $\sigma^2 \equiv \text{var}(y \mid X, Z)$, the residual variance. Because the sample is assumed to be selected on the basis of X and Z alone, these estimates are consistent for their population values (Wooldridge, 2000, pg. 299). The analysis sample can now be discarded; all other calculations are carried out using population data.

2. For the restricted regression models

$$\text{E}\left[y_i \mid X_i\right] = \alpha_1 + X_i\beta_1 \text{ and} \tag{4}$$

$$\text{E}\left[y_i \mid Z_i\right] = \alpha_2 + Z_i\gamma_2 , \tag{5}$$

calculate the coefficients and residual variance implied by the parameters estimated in step 1 and the population moments of X and Z. $\hat{\beta}_1$ is as above in (3); the other omitted variables estimates are:[14]

$$\begin{aligned} \hat{\gamma}_2 &= \gamma_0 + \rho_{XZ}\sqrt{\frac{\text{var}(X)}{\text{var}(Z)}}\beta_0 , \\ \hat{\text{var}}(y \mid X) &= \sigma^2 + \gamma_0^2\left(1-\rho_{XZ}^2\right)\text{var}(Z) \text{, and} \\ \hat{\text{var}}(y \mid Z) &= \sigma^2 + \beta_0^2\left(1-\rho_{XZ}^2\right)\text{var}(X) . \end{aligned} \tag{6}$$

---

[13] X's validity is $R_X \equiv \rho_{Xy} = \beta_1\sqrt{\frac{\text{var}(X)}{\text{var}(y)}}$. The terms in the latter fraction are those repaired by corrections for restriction of range, which can permit a consistent estimate even when the sample variance of X is small but which requires a clean estimate of (1).

[14] These formulae are needed when only the variance-covariance matrix of (X,Z) is available for the population. With individual-level observations, a simple way to calculate the omitted variables coefficients and residual variances is to generate predicted values from the regression estimated in step 1, then regress these predicted values on X and Z —separately—in population data (in my example, the College Board data). The coefficients from these regressions are the omitted variables coefficients described in (3) and (6); the residual variances are the final terms in the expressions for $\hat{\text{var}}(y \mid X)$ and $\hat{\text{var}}(y \mid Z)$ in (6).

(I omit the estimators for α_1 and α_2; these are nuisance parameters that, because they are constants, do not contribute to explained variation.)

3. If individual-level observations for the population are available, compute predicted values for each observation in the population based on each of the three regression models, using the omitted-variables coefficient estimates from step 2:

$$\begin{aligned} \hat{y}_{0i} &= \hat{\alpha}_0 + X_i\hat{\beta}_0 + Z_i\hat{\gamma}_0 \\ \hat{y}_{1i} &= \hat{\alpha}_1 + X_i\hat{\beta}_1 \\ \hat{y}_{2i} &= \hat{\alpha}_2 + Z_i\hat{\gamma}_2 \end{aligned} \tag{7}$$

Under the selection assumptions, these—not fitted values from least-squares estimation of (4) and (5)—are the best linear predictors for the population. This step is useful primarily as an aid to intuition; the next step suggests how it can be skipped when only aggregate population variances and correlations are available.

4. If the hypothesized model (1) holds, and if the actual y_i could be observed for every individual in the population, we would see that $\text{var}(y) = \text{var}(\hat{y}_0) + \sigma^2 = \text{var}(\hat{y}_1) + \text{var}(y|X) = = \text{var}(\hat{y}_2) + \text{var}(y|Z)$. Thus, using the interpretation of validity coefficients as the square root of the explained share of variance, estimate them as:

$$\begin{aligned} \hat{R}_{XZ} &\equiv \sqrt{\frac{\text{var}(\hat{y}_0)}{\text{var}(y)}} \\ &= \sqrt{\frac{\hat{\beta}_0^2\,\text{var}(X) + 2\hat{\beta}_0\hat{\gamma}_0\,\text{cov}(X,Z) + \hat{\gamma}_0^2\,\text{var}(Z)}{\hat{\beta}_0^2\,\text{var}(X) + 2\hat{\beta}_0\hat{\gamma}_0\,\text{cov}(X,Z) + \hat{\gamma}_0^2\,\text{var}(Z) + \hat{\sigma}^2}}; \\ \hat{R}_X &= \sqrt{\frac{\text{var}(\hat{y}_1)}{\text{var}(\hat{y}_1) + \hat{\text{var}}(y \mid X)}} = \sqrt{\frac{\hat{\beta}_1^2\,\text{var}(X)}{\hat{\beta}_1^2\,\text{var}(X) + \hat{\text{var}}(y \mid X)}}; \text{ and} \\ \hat{R}_Z &= \sqrt{\frac{\text{var}(\hat{y}_2)}{\text{var}(\hat{y}_2) + \hat{\text{var}}(y \mid Z)}} = \sqrt{\frac{\hat{\gamma}_2^2\,\text{var}(Z)}{\hat{\gamma}_2^2\,\text{var}(Z) + \hat{\text{var}}(y \mid Z)}} \end{aligned} \tag{8}$$

Given the regression coefficient estimates from step 2, the final expression of each validity requires only population variances and covariances to calculate, and can be used by researchers without access to individual-level population data. The above provide consistent estimates of the population validities as long as the sample is selected on the basis of X and Z alone. X's incremental validity is simply $\hat{R}_{XZ} - \hat{R}_Z$.

Several of the above steps can be collapsed into algebraically simpler (but conceptually less clear) calculations. Indeed, the usual validity estimator is identical to that produced by this algorithm, with one important exception: its estimates replace the population variances and correlations in (3) and (6) with their sample analogues. As demonstrated in Section 3, when both *X* and *Z* enter into the selection rule, the within-sample correlation between *X* and *Z* can be a poor estimate of its population value; as a result, neither β_1 nor γ_2 is consistently estimated by within-sample regression.[15]

The above algorithm is easily generalized to cases where more than two variables enter the selection process, or where we are interested in the incremental validity of the test score with

[15] When only one of the variables affects selection, one of the regression coefficients and one of the univariate validities is consistent; the other is not.

respect to a variable other than those determining selection. The key point is that the original regression model, (1), must include all variables determining selection as well as all variables used in prediction. Omitted variables formulae based on population covariances are then used to remove variables from this original model without introducing selection bias. Thus, the approach can only be fully implemented if the researcher can observe all determinants of sample selection. However, even when the data do not permit this, a limited version of the approach can serve as a diagnostic for selection bias: If validity estimates produced using the traditional methods differ greatly from those produced by the above algorithm when all available variables that might predict *y* are included in (1), the traditional estimates are likely biased.

It is worth repeating that the proposed approach is only sensible if population validity coefficients are of interest, in which case it is consistent under far more general sample selection assumptions than is the usual methodology. In the event that the researcher is interested in the within-sample validity, however, there can by definition be no selection bias in the traditional methods.

5. RESULTS

Table 4 presents regression parameters and validity coefficients for the UC-eligible subsample. The first column lists the parameters implied by the traditional approach to validity estimation, while the second implements the omitted variables approach proposed in Section 4. As argued there, the traditional approach underestimates β_1 and γ_2, leading to downward-biased univariate validity coefficients. The difference in estimates from the two approaches is substantial: The usual estimator for the SAT's univariate validity, when applied to the UC-eligible subsample, is 10 percent lower than the selection-corrected estimate, while the SAT's incremental validity is 7 percent too high.

Table 4: Traditional and new estimates of validity parameters

	Traditional Estimates	New Approach	Bias in Traditional
Population validities			
SAT	0.479	0.533	-10.2%
HSGPA	0.626	0.630	-0.5%
Bivariate: SAT+HSGPA	0.674	0.674	---
SAT incremental	0.048	0.045	7.1%
Regression parameters			
corr(*SAT, HSGPA*)	0.378	0.520	-27.3%
β_1	0.0014	0.0018	-19.3%
var(*y* | *SAT*)	0.346	0.398	-13.2%
γ_1	0.726	0.744	-2.4%
var(*y* | *HSGPA*)	0.326	0.336	-3.1%

Notes : Validites are for prediction of adjusted FGPA; see text (note 6) for definition. All estimates based on subsample of 17,346 UC-eligible observations.

Table 5 provides several alternative specifications. The first row repeats estimates from Table 4. In the second row, I eliminate the campus and major adjustment that has been used throughout, treating all UC grades as directly comparable. As noted above, unadjusted FGPAs are not as

accurately predictable as adjusted FGPAs, suggesting that campuses and majors with low-SAT and low-HSGPA students have less rigorous grading standards. The selection bias in validity coefficients is even stronger in this specification, however.

The second row considers a different measure of student outcomes: The 5-year cumulative GPA, adjusted by the same mechanism as that described for FGPA in footnote 6. This variable is slightly less predictable than is FGPA, but estimates of the SAT's validity for this outcome are seen to be subject to the same selection bias as those for FGPA. Finally, the remaining rows present estimates of SAT validity for FGPA computed separately for each race-gender combination. Again, the general results are similar to those seen in earlier estimates that pooled all students into a single group.

Table 5: SAT validity in alternative specifications

		SAT Univariate Validity			SAT Incremental Validity		
		Trad.	New	Bias	Trad.	New	Bias
Base Model		0.479	0.533	-10.2%	0.048	0.045	7.1%
No Campus/Major Controls		0.416	0.476	-12.6%	0.041	0.037	12.9%
Pred. Of 5-year Cumul. GPA		0.452	0.509	-11.2%	0.046	0.042	9.5%
Within-Gender/Race							
Male	White	0.493	0.535	-7.8%	0.050	0.042	20.6%
	Black	0.321	0.399	-19.5%	0.165	0.086	92.5%
	Hispanic	0.386	0.459	-16.0%	0.070	0.057	23.2%
	Asian	0.472	0.533	-11.4%	0.047	0.036	30.9%
Female	White	0.521	0.552	-5.6%	0.047	0.045	3.1%
	Black	0.314	0.364	-13.7%	0.091	0.082	11.4%
	Hispanic	0.383	0.449	-14.6%	0.048	0.051	-5.2%
	Asian	0.502	0.536	-6.4%	0.053	0.053	-0.4%

Notes: All estimates based on subsample of 17,346 UC-eligible observations (broken into 8 groups for the within-gender/race cells). Validities are for unadjusted FGPA (row 2), adjusted five-year cumulative GPA (row 3), and adjusted FGPA (all other rows). See text (note 6) for description of adjustment procedure.

6. CONCLUSION

The validity literature has paid little attention to the substantial statistical difficulties introduced by the use of selected samples for estimation. The primary exception to this generalization, the use of restriction of range corrections, has frequently been treated as a magical technique that repairs all sample selection bias (Camara and Echternacht, 2000; Hezlett et al., 2001). In fact, range corrections are appropriate only under very specific selection assumptions and are never appropriate for more than one statistic in a given sample: If the sample is assumed to be selected solely on the basis of SAT scores, SAT's univariate validity may appropriately be range-corrected; if selection is on SAT and HSGPA, but nothing else, their joint validity—but not SAT's univariate or incremental validity—may be corrected.

There are no statistical assumptions that can justify the standard practice of estimating both univariate and incremental validity parameters that do not take account of sample selection except through simple restriction-of-range corrections. I have proposed an approach to sample selection bias that is consistent under plausible assumptions—namely, that only variables observed by the researcher in both the sample and the population influence selection. These assumptions are difficult to satisfy, but they are always more general than those required for the consistency of unadjusted or range-corrected validity estimates. Moreover, even when the necessary assumptions are not met, the new approach suggests a diagnostic for some forms of selection bias in validity estimates.

Under the sorts of conditions that generate most estimation samples, uncorrected and range-corrected validity coefficients are substantially biased, in predictable directions. Estimates of univariate validity understate the true parameters, while incremental validity estimates are biased upwards. Since incremental validity coefficients are often the true measure of an exam's utility, the selection bias discussed here may lead to the overuse of uninformative exams.

REFERENCES

Braun, Henry I. and Ted H. Szatrowski (1984a). "The Scale-Linkage Algorithm: A Universal Criterion Scale for Families of Institutions," *Journal of Educational Statistics 9* (4), Winter, p. 311-330.

--- and --- (1984b). "Validity Studies Based on a Universal Criterion Scale," *Journal of Educational Statistics 9* (4), Winter, p. 331-344.

Breland, Hunter M. (1979). "Population Validity and College Entrance Measures," Research Monograph 8, The College Board, New York.

Bridgeman, Brent, Laura McCamley-Jenkins, and Nancy Ervin (2000). "Predictions of Freshman Grade-Point Average From the Revised and Recentered SAT I: Reasoning Test," Research Report 2000-1, The College Board, New York.

Camara, Wayne J. and Gary Echternacht (2000). "The SAT I and High School Grades: Utility in Predicting Success in College," Research Notes RN-10, The College Board, Office of Research and Development, July.

College Board (2002). "SAT V+M Composites: Original to Recentered Scale," Web page http://www.collegeboard.com/sat/cbsenior/equiv/rt027027.html.

Gulliksen, H. (1950). *Theory of Mental Tests*, Hillsdale, NJ: John Wiley & Sons.

Hezlett, Sarah A., Nathan R. Kuncel, Meredith A. Vey, Allison M. Ahart, Deniz S. Ones, John P. Campbell, and Wayne Camara (2001). "The Predictive Validity of the SAT: A Meta-Analysis," in Deniz S. Ones and Sarah A. Hezlett, eds., *Predicting Performance: The Interface of I-O Psychology and Educational Research*, Sixteenth Annual Conference of the Society for Industrial and Organizational Psychology 2001.

Keller, Dana, James Crouse, and Dale Trusheim (1994). "The Effects of College Grade Adjustments on the Predictive Validity and Utility of SAT Scores," *Research in Higher Education 35* (2), p. 195- 208.

Rothstein, Jesse M. (2002), “College Performance Predictions and the SAT," UC Berkeley Center for Labor Economics Working Paper #45, October.

Student Academic Services (multiple years). “Introducing the University of California: Information for Prospective Students," Application handbook, University of California Office of the President.

van der Vaart, A. W. (1998). *Asymptotic Statistics*, Cambridge University Press.

Willingham, Warren W., Charles Lewis, Rick Morgan, and Leonard Ramist, eds (1990). *Predicting College Grades: An Analysis of Institutional Trends Over Two Decades*, Educational Testing Service.

Wooldridge, Jeffrey M. (2000). *Introductory Econometrics: A Modern Approach*, South-Western College Publishing.

Young, John W. (1990). “Are Validity Coefficients Understated Due to Correctable Defects in the GPA?," *Research in Higher Education 31* (4), p. 319-325.