DOCUMENT RESUME

ED 383 736 TM 023 187

AUTHOR Leonard, David K.; Jiang, Jiming
TITLE Gender Bias ir the College Predictions of the SAT.
PUB DATE 21 Apr 95
NOTE 50p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Students; Evaluation Methods; *Females; *Grade Point Average; Grades (Scholastic); Higher Education; High Schools; *High School Students; Males; *Prediction; *Sex Bias; Sex Differences; Test Bias; Test Construction
IDENTIFIERS College Entrance Examination Board; *Scholastic Aptitude Test; University of California Berkeley

ABSTRACT

This paper demonstrates that the various College Board examinations, most importantly the Scholastic Aptitude Tests (SATs), make predictions of grade point averages at the University of California at Berkeley that are biased against women. This finding persists even when one has made corrections for differences in fields that women and men study and for selection bias. Because women in fact are better students than the SATs predict, they are underrepresented both proportionately and relative to their merit in Berkeley's freshman classes (as they would also be in scholarship competitions and at other selective colleges). The differences in predicted grades are small, but if unbiased tests were used, Berkeley would have at least 5% more women in its freshmen classes (200 to 300 per year). Various solutions to the bias in the SATs are explored. Simple mathematical "fixes" by admissions officers will not work well. The best solution would be for the Educational Testing Service to correct the bias by altering the mix of questions in its tests. Failing this, college admissions staff should significantly expand the numbers of women's applications that are evaluated qualitatively. Five tables and six figures present analysis results. (Contains 18 references.) (Author/SLD)

ED 383 736

Abstract of

GENDER BIAS IN THE COLLEGE PREDICTIONS OF THE SAT

by
David K. Leonard (Department of Political Science),
and
Jiming Jiang (Department of Statistics),
The University of California at Berkeley

A paper presented to the annual conference of the American Educational Research Association, San Francisco, California, April 21, 1995.

The paper demonstrates that the various College Board examinations (most importantly the Scholastic Aptitude Tests) make predictions of grade point averages at the University of California at Berkeley that are biased against women. This finding persists even when one has made corrections for differences in the fields that women and men study and for selection bias. Because women in fact are better students than the SATs predict, they are under-represented both proportionately and relative to their merit in Berkeley's freshman classes (as they would be also in scholarship competitions and at other selective colleges). The differences in predicted grades are small, but if unbiased tests were used Berkeley would have at least 5 per cent more women in its freshman classes (200-300 a year).

Various solutions to the bias in the SATs are explored. Simple mathematical "fixes" by admissions officers will not work well. The best solution would be for the Educational Testing Service to correct the bias by altering the mix of question items in its tests. As long as the ETS refuses to take corrective action and falsely minimizes the extent of the problem (of which it has been aware for 50 years), college admissions staff should counter the bias by significantly expanding the numbers of files of women's applications that are evaluated qualitatively. For large, highly-competitive public universities, such as Berkeley, this solution is expensive and painful to implement in these resource-constrained times.

April 17, 1995

GENDER BIAS IN THE COLLEGE PREDICTIONS OF THE SAT

by
David K. Leonard (Department of Political Science),
and
Jiming Jiang (Department of Statistics),
The University of California at Berkeley

The College Board has been administering the Scholastic Aptitude Test since 1926 (Jencks & Crouse, 1982: 35). It was designed to predict the performance of secondary school students in college, and in the nearly 70 years that have ensued it has remained dedicated to that task. Last year the SAT was relabeled the Scholastic Assessment Tests and its format changed. But the purpose of predicting the grades that applicants will earn when they get to college remains and the SAT is always justified in terms of this purpose (College Entrance Examination Board, 1991). The Educational Testing Service, which develops and administers the SATs on behalf of the College Board, has meticulously evaluated the tests' effectiveness as predictors (e.g., Willingham, et al., 1990: esp. p. 90).

Thus to those of us in the academic community who rely on the SAT to guide us in our admissions tasks, it is a shock to learn that not only are the tests flawed in their predictions but that this problem has been known to insiders for over half a century. Women earn higher grades in college than men with identical SAT scores. The problem here is not that on

 © Leonard

average women score less well than men do, which is an entirely different issue. Our concern instead is that the SAT quite simply under-predicts the college performance of women (Wagner & Strabel, 1935; Thorndike, 1963; Lavin, 1965; Linn, 1973; Stricker, Rock & Burton, 1991).

The Educational Testing Service formally acknowledges the existence of this problem. The dominant response, however, has been to state that a large part of the under-prediction derives from a difference between men and women in course taking patterns and that when these differences are factored out the remaining under-prediction is only a small fraction of the GPA (Elliott & Strenta, 1988; Rigol, 1989; College Board, 1991: 2). This response is technically correct. It is then asserted or implied, however, that these difference are too small to have much practical effect and that they would be handled by respecting the Board's statement of the error margins on the tests (L.W. Hecht in Striker, Rock & Burton, 1991, pp. vii-viii.) This latter assertion is false.

In this paper we use data from the University of California at Berkeley to demonstrate that this "small" difference in fact negatively impacts a large number of women, reducing the number of women in Berkeley's freshman class by over 5 per cent. In the process we will show that a full half of the under-prediction of women's college grades remains after one has taken out the effects of choice of

program of study and that this under-prediction exists across the range of scores in which highly competitive colleges and universities actually make their cut-off decisions. We also will demonstrate that some of the methods that have been suggested for correcting the bias in the SAT are inadequate, unstable or inappropriate.

At the outset we would like to point out that none of those involved in writing this paper have ever had professional relationships with the Educational Testing Service or the College Board.* We also all are men and never have written on the subject of gender discrimination before. Indeed the study out of which these findings arise was originally undertaken with a quite different set of concerns in mind. We are as close to an unbiased set of observers as this problem is likely to get.

THE DATA

This study is based on the academic performance at

---

*In addition to the authors, those involved in this study were Professor Terry Speed (Department of Statistics). Dr. Tom Cesa (Office of Student Research) and Walter Wong (Office of Undergraduate Admissions). Assistance also was provided by Professor Marcia Linn and Kathy Kessel (School of Education). The authors are very grateful to all of them for their help, while taking full responsibility for all interpretations presented in this paper. At the time this study was begun David Leonard was chair of the Berkeley Academic Senate's Committee on Admissions and Enrollment.

Berkeley of all students admitted as freshmen between 1986 and 1988 (N= ~10,000). In 1986 California high school students were permitted for the first time to apply to an unlimited number of campuses of the University of California. As a consequence, applications to attend Berkeley escalated by 60 per cent and the competition changed the nature of the admissions process. Our initial purpose in undertaking this study was to examine the determinates of successful completion of an undergraduate degree at Berkeley. At the time we began this study 1988 was the last year for which a minimum of four years of data on academic performance were available.

The dependent variable in our analysis is the student's cumulative Grade Point Average (GPA) after graduation, transfer, withdrawal or termination from the university, whichever came first. Since the National Collegiate Athletic Association now evaluates institutions by the percentage of its entering freshmen who have graduated six years later and since very few Berkeley undergraduates complete their degrees more than six years after their initial enrollment, we decided to examine up to six years of data on the academic performance of each student.

Our study thus differs from a number of others on this subject in two respects. First, we examine students' performance over their entire undergraduate careers. Many

others concentrate on only the freshman year. Colleges and universities admit students for their full undergraduate contribution to the campus, however, not just for the freshman year. We believe that many other studies concentrate on the freshman year largely in order to deal with the effect that choice of classes has on the differential grades of men and women (Stricker, Rock & Burton, 1991; Willingham, et al., 1990). We deal with this problem in another way.

Second, we include the performance of those who have left the Berkeley campus without graduating. By and large other studies either have examined only four year graduates (e.g., Elliott & Strenta, 1988) or have treated as failures all those who do not complete their degree at their original institution in a specified period of time (the NCAA approach). Neither of these approaches is consistent with contemporary student behavior, especially at large public universities. A great many take more than four years to graduate (most often because they are simultaneously working); many others transfer in good standing to other institutions because the first did not meet their expectations; and still others failed out or left because their status was marginal. Berkeley does not consider the first two sets of students to be failures and would not wish to discriminate against them in its admissions practices. It

does wish to distinguish between those who left the campus in good or poor standing; thus our study includes an examination of the grades of such students.

We focus on the prediction of undergraduate cumulative GPA because Berkeley has a powerful (if not primary) interest in identifying and admitting those prospective undergraduates who will excel academically as students on the campus. There are other considerations -- such as artistic achievement, athletic competitiveness and ethnic diversity -- that contribute to excellence and that sometimes cause the campus to supplement expected academic achievement with them as criteria of admission. But expected academic performance is never ignored, and it is the primary (if not sole) determinant of the admissions decision in the vast majority of cases. Predicting undergraduate excellence accurately and efficiently therefore is a high priority in the admissions process.

This study is based on the premises that Berkeley professors are those best qualified to define and assess the academic quality of an undergraduate and that they do this through assigning students grades in their courses. Operationally undergraduate academic achievement is nearly synonymous with a high GPA. Obviously professors make mistakes in assigning grades, but Berkeley faculty are strongly convinced that no one else is better prepared than

they are to define the meaning of academic excellence in their respective fields of study. The SAT was designed to predict college GPA, not some other undefined concept of excellence. If the SAT and the college GPA differ in their evaluations, validity must be presumed to reside in the latter.

THE BERKELEY ADMISSIONS PROCESS

The most important determinant of admissions throughout the University of California system is a composite measure called the Academic Index Score (AIS). The AIS is made up of two equally weighted components: high school grade point average (HS GPA) and standardized tests. Five test scores are used: the verbal and mathematical portions of the Scholastic Aptitude Test and three Achievement Tests (English, Mathematics, and one other of the applicant's choice). Each test is scored between 200 and 800, so the range of possible values for this component of the AIS is 1000 to 4000.

The high school grade point average is comprised of those courses taken in the 10th and 11th grades that fulfill University of California requirements (known as "a-f" courses). The standard 4 point grade scale is used, but applicants are awarded an extra grade point for each University-recognized honors course they have taken. Thus it

is possible to have an adjusted HS GPA above 4.0 and at Berkeley a majority of those admitted have done so.** To calculate the AIS the HS GPA is multiplied by 1000 and added to the test scores. An applicant with a 4.0 high school GPA and 800 on all five College Board exams would receive 8000 AIS points.

The first 50 per cent of Berkeley's freshman class is admitted exclusively by reference to this AIS score.*** At the moment the cut-off for these "Tier 1" admits is 7120 to 7150 AIS points. Applicants down to about 7050 will be in the "Special Promise Read Pool" (which means that even if they have a HS GPA of 4.0 they would need to average 610 on the five College Board examinations). Admissions decisions within this group are made on the basis of the applicants' essays and their various achievements, but of course the AIS has determined whether or not their file gets read in this pool. Applicants who meet the minimum requirements for admission to the University of California systme will be offered the option of consideration for deferred admission in the spring semester; in practice those admitted under the "defer to spring" option have an AIS of about 6300 at the present time.

**In the period for which this study was undertaken the University of California "capped" HS GPA at 4.0. Berkeley has now removed this "cap" in its internal admissions decisions, but this issue is peripheral to the findings of this study.

***Before 1989 it was 40 per cent.

(So if an applicant's HS GPA were 3.4 she would need an average of 580 on the tests.) Some other students (e.g., athletes and under-represented minorities) will be admitted outside these limits, but it is clear that the AIS plays the determinative role in Berkeley's admission process. What in fact is the predictive "power" of the AIS?

THE GENDER BIAS IN BERKELEY'S CURRENT ADMISSIONS

Women have been under-represented in Berkeley's freshman classes for some time. They were 45.6% of the freshmen admitted between 1985 and 1988 and they were 45.5% of the Fall 1992 freshman class. Berkeley has never knowingly had any admissions policies that would discriminate against women, so the question arises as to where this disparity comes from.

It is <u>not</u> coming from the pool of those eligible for admission to the University of California (UC) system. Women were 50.7% of California's high school graduating classes in 1990. They were a still higher proportion of the high school graduates who had met UC's eligibility requirements. The California Post-Secondary Education Commission reports that in 1990 13.3% of the state's female high school graduates were fully eligible for UC, as opposed to 11.6% of the males. Another 7.5% of the women were potentially eligible if they took one or more tests, as against 5.6% of the men. Thus

women represented 54.1% of the fully eligible students and 55.4% of the eligible pool, if one included the potentially eligible. By these measures women should have been a majority of Berkeley's freshman classes.

One part of the problem lies in the application process. For example, for Fall 1992 women were 46.8% of the applicants, 47.2% of those admitted and 45.5% of those enrolled. These figures suggest that some of the remedy lies in recruitment -- in persuading women that they are as good as men with the same high school grades. Females' diminished sense of self-worth may be partly caused by the SATs, however.

The larger part of the problem lies in the biased estimate produced by the indicators Berkeley uses to predict excellence. The AIS under-predicts the undergraduate grades of women and over-predicts those of men. On average, women will have a tenth of a point higher GPA at Berkeley than men with identical AIS scores--a figure that appears with regularity in studies such as these (Stricker, Rock & Burton, 1991). Figure 1 uses the "smoothed line" method**** to plot the average cumulative GPAs of men and women admitted to the

---

****A smoothed line is formed by joining successive points fitted locally by a robust weighted least squares method. This smoothed line, although locally linear, can take any shape in its larger configuration and is an asset in estimating the best functional form for a relationship (Cleveland, 1979).

College of Letters and Science (L&S) at Berkeley in 1986 against their AIS scores. It will be seen that the female GPA line is consistently higher. The same point is illustrated in Table 1 for L&S 1988 freshmen. There students are rank ordered according to their AIS scores and then grouped in rising ten per cent intervals (deciles). The average cumulative GPA percentile of each AIS decile is then presented for men and women. Again, the pattern is one of consistently better performance by the women.

Fig. 1: Average College GPA Per Academic Index Score Interval for the College of Letters and Science for the Class Entering as Freshmen in 1986

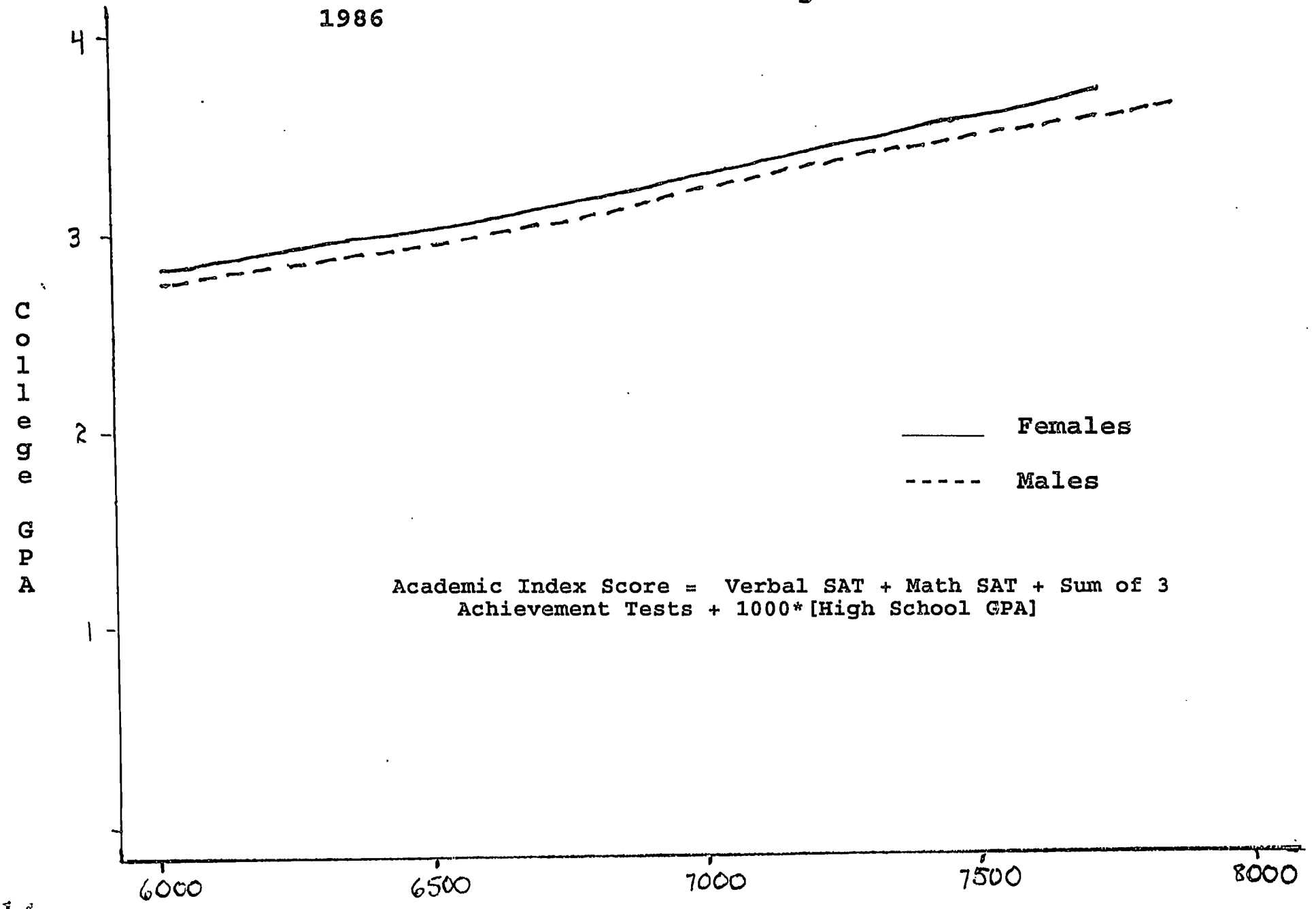


Table 1

The Average Cumulative Grade Percentile of Men and Women Admitted to the College of Letters and Science as Freshman in 1988 By AIS Decile

	AIS PERCENTILE:									
	0-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
Male		30	23	28	33	41	53	60	63	69	75
Female	34	25	33	41	48	59	71	72	76	79	
All	32	25	31	37	44	56	66	67	72	76	
Diff.	-4	-2	-5	-8	-7	-6	-11	-9	-7	-4	

Total difference between male and female (male - female): -63

The most frequent hypothesis offered to explain the under-prediction of women's college grades is gender influences on the selection of courses (Elliott & Strenta, 1988; Rigol, 1989; Willingham, et al., 1990: 74). It might be that women do better than men at Berkeley because they choose to study subjects that are graded more leniently than those taken by men. It is true that the mathematically-based disciplines are graded somewhat more harshly than other fields at Berkeley and that the density of men is greater in these fields. Thus, for example, in 1992 the average GPA of seniors in the College of Chemistry, where women are 35% of the total, was 2.93 versus 3.16 for the College of Letters and Science, where 51% are women. On the other hand, women do outperform men in both Chemistry and Business Administration, although not in Engineering (for seniors,

2.95 v. 2.92; 3.32 v. 3.29; and 3.02 v. 3.11, respectively). Furthermore, AIS scores under-predict female GPA within Letters and Science, not just with respect to these professional schools (3.20 v. 3.12).

A better test of the hypothesis that choice of classes is raising the grades of women is an examination of grades in the mathematically-based colleges of Chemistry and Engineering. (See Figure 2 and Table 2.) The fact is that AIS scores under-predict women's performance in these colleges as well. The bias in the prediction is about half as large as it is for the campus as a whole, however. These figures suggest that choice of major by women is a part but only a part of the explanation. This finding is consistent with the other research done on this subject -- there remains an irreducible gender bias in the college GPA predictions of the SAT (Elliott & Strenta, 1988; Stricker, Rock & Burton, 1991; Wainer & Steinberg, 1991; Willingham, et al., 1990: 74).

Fig. 2: Average College GPA Per Academic Index Score Interval for the Colleges of Chemistry and Engineering for Classes Entering as Freshmen Between 1986 and 1988

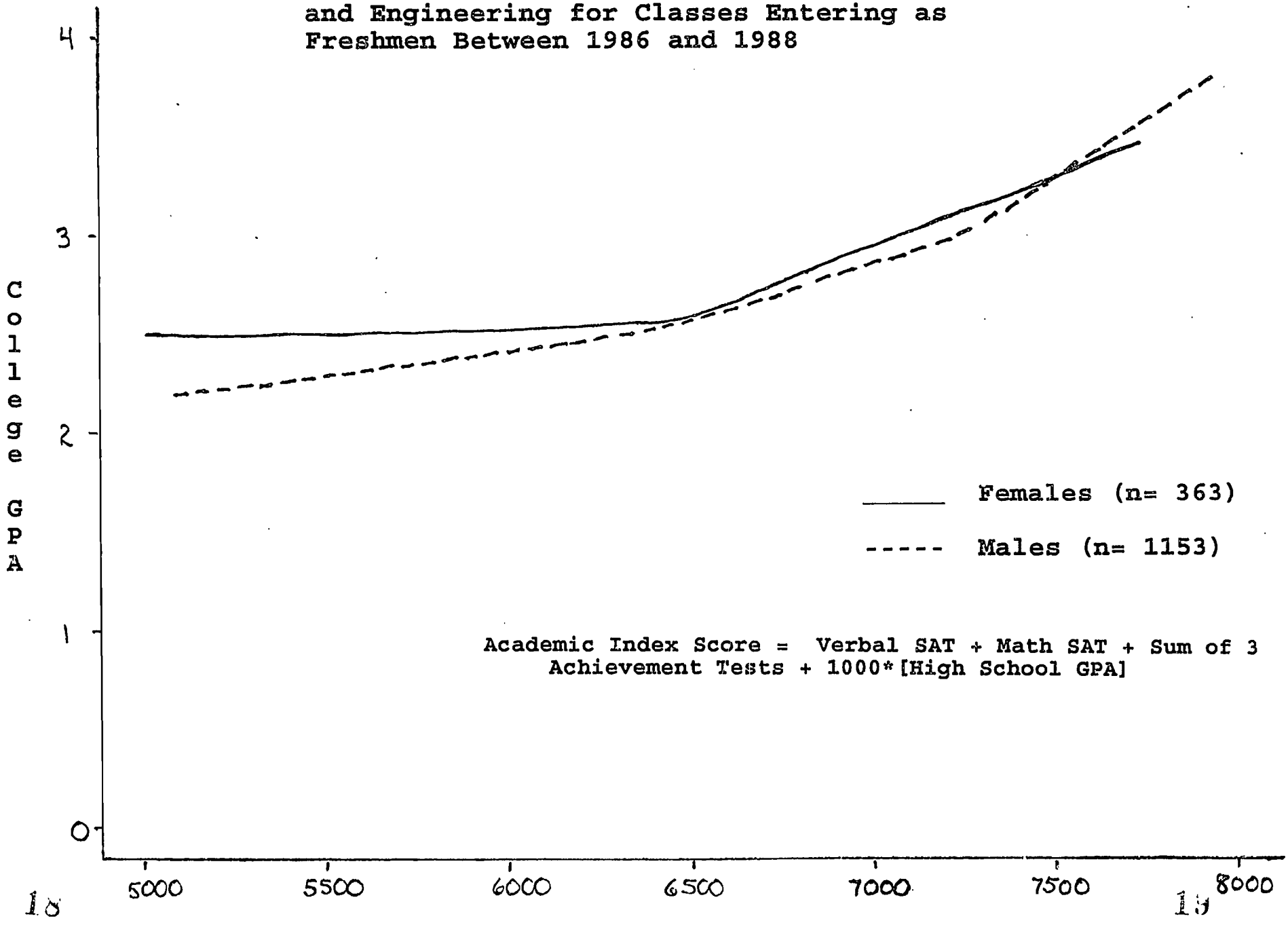


Table 2

The Average Cumulative Grade Percentile of Men and Women Admitted to the Colleges of Chemistry and Engineering as Freshman Between 1986 and 1988 By AIS Decile

	AIS PERCENTILE:									
	0-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
Male		29	32	39	43	47	50	58	62	65 75
Female	27	31	45	49	54	55	61	63	67	76
All	28	32	41	45	49	51	58	62	66	75
Diff.	+2	+1	-6	-6	-7	-5	-3	-1	-2	-1

Total difference between male and female (male - female): -28

Even the preceding test is flawed, however. In order to make a still stronger test of the hypothesis that choice of courses accounts for the AIS' under-prediction of women's GPAs, we have extended the control for subject matter and made two other ones as well. First, we have changed our focus from the college into which the student was admitted and instead look at the field in which they graduated. It could be that women drop out of some fields more than men do and move to other, "easier" areas of study. Second, we have grouped students not only by the college in which they graduated but by the division of their major within it, so that we now are looking at groupings such as -- Biological Sciences, Chemistry, Engineering, Humanities, Social Sciences, Physical Sciences, etc.

Third, we have made a correction that is not possible in

most studies of this kind -- we control for the effect of exogenous selection criteria. Generally the measurement of the effect of any given selection criterion for college admission is compromised statistically by the fact that those who have a low score on that dimension are not a random sample of the population; applicants who score low on one criterion have been admitted precisely because of the fact tnat they score high on other criteria. Note that the left hand tail of the female smoothed line in Figure 2 is essentially flat. These women have been admitted to these colleges in spite of, not because of, their AIS scores. That judgement was based on the presence of other indicators of probable success, ones that turned out to be largely correct.

Berkeley's admission process offers a rare opportunity to correct for this statistical problem. As indicated above, half of the freshman class is admitted on the basis of their AIS score alone -- a purely mechanical combination of their HS GPA and the College Board test scores with no element of human judgement intruding. Although this cut-off number varies somewhat between fields of study and across years, it is constant and precisely known within any field at any one time. If we examine the performance of only those students admitted by their AIS scores alone, we can see its predictive power uncontaminated by other, compensating elements of

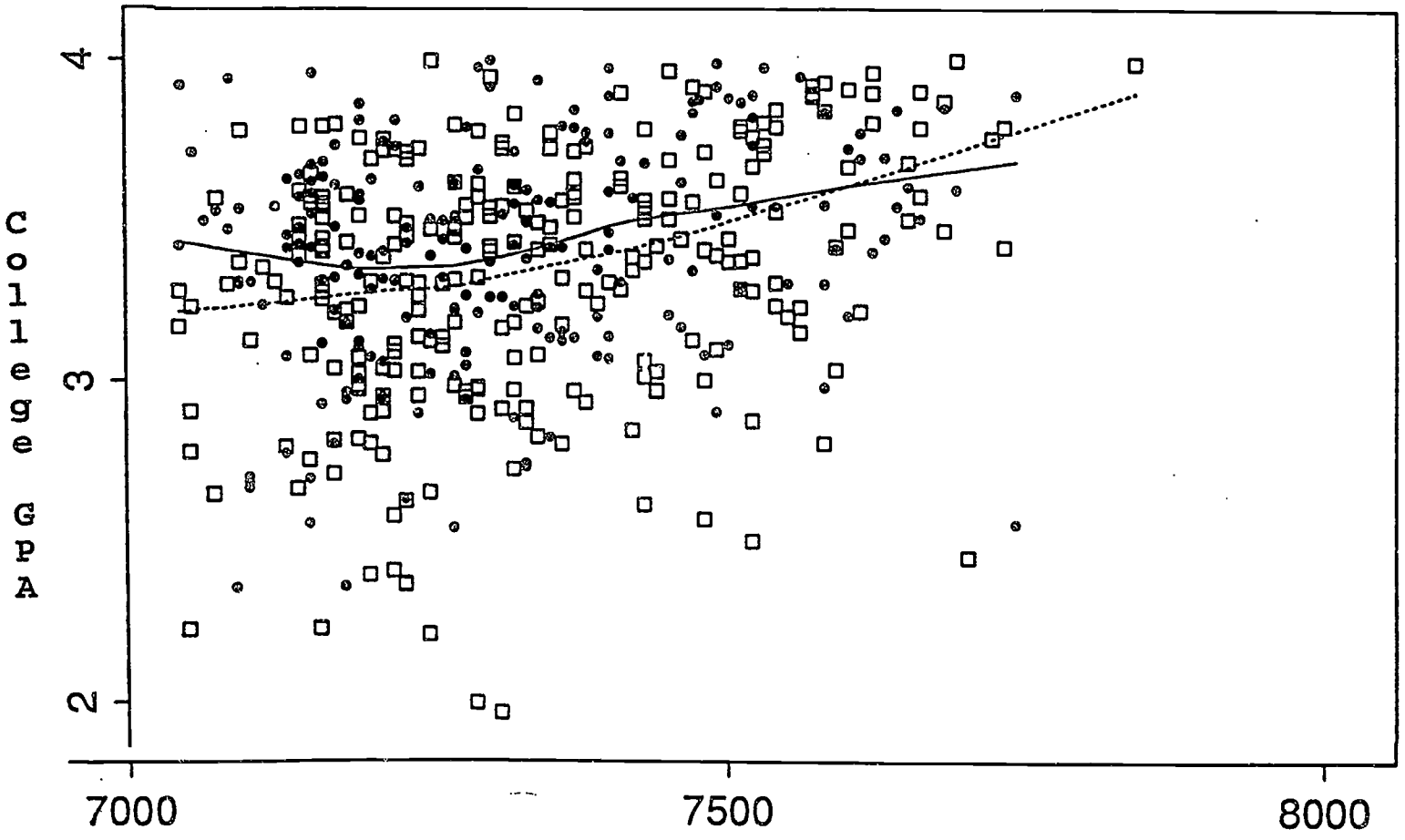
judgement.*****

When we control for field of study and for the effect of compensating indicators of excellence, we find that the under-prediction of women's college grades persists. This is illustrated clearly by Figure 3, which gives a scattergram and smoothed line representation of the relationship between AIS and GPA for men and women in the Division of Biological Sciences in the College of Letters and Science (L&S) who were admitted as freshmen between 1986 and 1988.

A more general representation of the under-prediction of women's grades and an indication of its persistence across fields is provided in Table 3. An inspection of the within-field data indicate a good deal of variation in the patterns of relationship between AIS and GPA, but the functional form that comes closest is a linear one. Thus using linear regression we derive the slope and intercept of the line that predicts GPA from AIS. Taking the slope and intercept for the men (columns 1 &2) we can calculate the expected GPA (col. 4) of a male who had the exact minimum Academic Index Score that was needed to gain admission in most of the 1986-88 period for his field of study (col. 3). Using the slope and intercept calculated by linear regression on the data for

*****Note that within the College of Letters and Science, which is where most undergraduates at Berkeley are enrolled, admission is made without reference to an applicant's prospective field of study.

Fig. 3: College GPAs By Academic Index Score for Men and Women Majors in the Division of Biological Sciences Who Entered as Freshmen Between 1986 and 1988



_____ Females

----- Males

N = 452

Dots: Females

Boxes: Males

Academic Index Score = Verbal SAT + Math SAT + Sum of 3 Achievement Tests + 1000*[High School GPA]

Table 3: Gender Calculations on AIS

Field of Major	(1) Male Intercept	(2) Male Slope	(3) Admiss. Cut-off	(4) Expected M-GPA @Cut-off	(5) Female Intercept	(6) Female Slope	(7) Equivalent Female Cut-off	(8) Male minus Female Difference	(9) Number of Cases
Engin	-3.219	0.00087	7500	3.306	-3.039	0.00085	7465	35	222
Env. Des.	1.374	0.00026	6900	3.168	1.252	0.00029	6607	293	132
Chem	-4.424	0.00103	7000	2.786	-5.238	0.00114	7039	-39	107
Natural	0.473	0.00037	6800	2.989	-0.253	0.0005	6484	137	132
L&S-Interdis	1.306	0.00029	7100	3.365	-0.304	0.00052	7056	44	225
L&S-Bio	-2.391	0.00078	7100	3.147	-0.423	0.00052	6865	235	452
L&S-Phy	-1.504	0.00065	7100	3.111	2.853	7E-05	3686	3414	240
L&S-Hum	-0.565	0.00054	7100	3.269	0.091	0.00048	6808	292	320
L&S-Soc	0.011	0.00046	7100	3.277	-2.568	0.00083	7042	58	605
									2435

Weighted difference =	462
Weighted difference w/o Physical Sci.=	139
Weighted difference of math.-based subjects=	1447
Weighted difference of non-math. subjects=	161

females (cols. 5 & 6), we can calculate the AIS score a woman applicant would need in order to have the same expected GPA as this "last admitted" man (col 7). The difference between this figure and the cut-off is given in column 8 and is an estimate of the magnitude of the bias against women in the AIS in that field of study. (A negative figure indicates an estimated bias against men.) Note that although there is stochastic variation in the size of the estimated bias from field to field, the bias is against women in every field but one. Chemistry, the one exception, has the smallest number of cases and the deviation toward men is modest. It is reasonable to assume that this one field represents a chance anomaly and the stable underlying pattern is one of bias against women. At the base of the table we give a weighted average of the bias in the AIS against women -- 462 points. We do not wish to put too much emphasis on this particular magnitude of bias, for the estimates show a good deal of variation among them. (The AIS is such a poor predictor of the grades of women who do finally study in the Physical Sciences, which includes Mathematics, that the estimated bias for that field is unrealistically high. If we exclude the Physical Sciences from our calculations, we get an estimated bias of 139 and if we exclude all the other mathematically based fields as well -- Engineering, Biological Sciences, Chemistry, and Natural Resources -- the size of the bias is

161.)

Another way of expressing the gender gap is to indicate how much better the grades of a woman with a given AIS score would be than a comparable man. At the AIS cut-offs indicated on Table 3, the difference is .09 of a grade point (.08 if one drops the Physical Sciences from the calculations). Thus it is true that once one has corrected for course selection the prediction gap between men and women is smaller -- about .08 or .09 of a grade point v. the .10 we indicated earlier. This gender difference is small but far from trivial. If women had been given the 139 AIS points we estimated they deserved, two to three hundred white and Asian women would have been admitted in Berkeley's Tier 1 and Special Promise categories who otherwise were denied or forced to defer their admission to the spring.****** This would have increased the numbers of women in the freshman class by about 7 per cent. This problem is now in the process of being corrected at Berkeley, but substantial numbers of women applicants are almost certainly being disadvantaged at other large, highly competitive universities that make a preliminary admissions ranking "by the numbers."

******Minority women were treated affirmatively and would have been admitted if they met the University of California's minimum entrance criteria and were judged to have a reasonable prospect of graduation. Thus it is non-minority women (i.e., whites and Asians) who were most seriously impacted by the gender bias in the AIS.

Similarly, any scholarships that are awarded solely on the basis of grades and/ or College Board test scores (such as the National Merit Scholarships) disadvantage women financially.

Note that the argument made here for gender bias in the predictions of the College Board examinations is not based on any a priori assumption that men and women should perform at equivalent levels and that any difference in admissions between the two is evidence of discrimination. Given the complex ways in which sex roles interact in our society with patterns of maturation, preparation for college and choice of subjects for study, pure equality of result would be prima facie evidence that violence were being done to equality of performance. Since women do not take as many math and science courses in high school as men do, we would expect that any performance-based method of selection would produce more men than women in these fields. But similarly, since women do better than men in high school overall, college admissions processes based solely on expected performance should probably produce more women than men, once field of study is ignored.

THE ROLE OF THE SATS IN GENDER DISCRIMINATION

We have demonstrated conclusively that the use of HS GPA plus College Board examinations as the sole criteria for

admissions decisions creates gender bias at highly competitive colleges. Where does this bias come from? It does not come from high school grades. Our aggregate data indicate that high school grades are an almost perfectly gender neutral predictor of Berkeley GPA. (To the extent that there is any effect of sex evident it is tilted against women. At Berkeley cumulative GPA is positively correlated with being female at .044, but when one controls for HS GPA the partial correlation coefficient increases slightly to .049; N=13398.)

To examine the bias in the SATs, we have examined their predictive validity for each field of study. We have used the same truncated sample discussed above to reduce the contaminating effect of other admissions criteria. Thus the following analysis is restricted to those talented applicants who were admitted to Berkeley on the basis of their HS GPAs and their College Board test scores alone. Unfortunately, an examination of SAT scores alone is still a somewhat contaminated one, however, for those admitted with lower test scores have to have had very high HS GPAs.

Tables 4 and 5 repeat for the Verbal and Math SATs the analysis done for the AIS in Table 3. The cut-off point used here is a score of 650 on an individual test, the lowest score one could average and be assured of admittance to Berkeley with a HS GPA of 3.85 (unless one were an athlete, a

Table 4: Gender Comparisons on the Mathematics SAT

Field of Major	(1) Male Intercept	(2) Male Slope	(3) Sample SAT Score	(4) Expected M-GPA @Score	(5) Female Intercept	(6) Female Slope	(7) Equivalent Female SAT	(8) Difference: Male minus Female	(9) Number of Cases
Engineering	1.329	0.00256	650	2.993	2.106	0.00149	595	55	222
Environ.Design	2.539	0.00099	650	3.1825	3.497	-0.00025	1258	-608	132
Chemistry	0.183	0.00402	650	2.796	1.371	0.00237	601	49	106
Natural Res.	2.226	0.00125	650	3.0385	2.974	0.00035	184	466	137
L&S-Interdis.	3.82	-0.00054	650	3.469	3.206	0.00035	751	-101	225
L&S-Bio. Sci.	1.983	0.00188	650	3.205	2.853	0.0008	440	210	452
L&S-Phys. Sci.	1.537	0.00242	650	3.11	2.595	0.00111	464	186	240
L&S-Humanities	3.677	-0.0004	650	3.417	3.86	-0.00053	836	-186	320
L&S-Soc. Sci.	3.808	-0.00058	650	3.431	2.698	0.00113	649	1	605
									2439

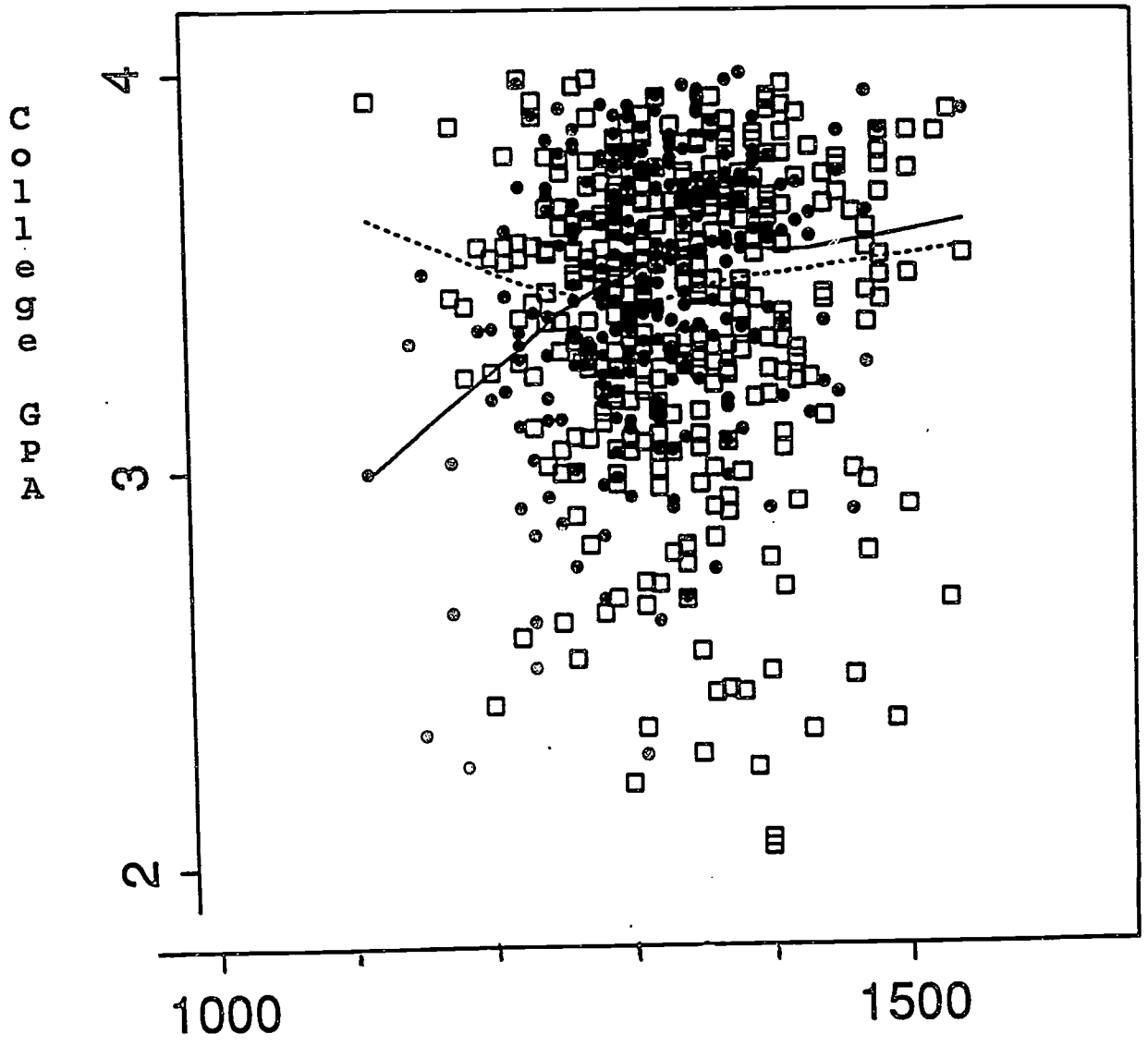
Weighted difference =	24
Weighted difference w/o Physical Sci.=	6
Weighted difference of math.-based subjects=	109
Weighted difference of non-math. subjects=	-2

Table 5: Gender Comparisons on the Verbal SAT

Field of Major	(1) Male Intercept	(2) Male Slope	(3) Sample SAT Score	(4) Expected M-GPA @Score	(5) Female Intercept	(6) Female Slope	(7) Equivalent Female SAT	(8) Difference: Male minus Female	(9) Number of Cases
Engineering	2.86	0.00056	650	3.224	3.121	8E-05	1287	-637	222
Environ.Design	3.657	-0.00071	650	3.1955	2.951	0.00064	382	268	132
Chemistry	2.528	0.00096	650	3.152	2.081	0.00163	657	-7	106
Natural Res.	2.645	0.00077	650	3.1455	2.675	0.00094	501	149	137
L&S-Interdis.	3.414	4E-05	650	3.44	3.014	0.00068	626	24	225
L&S-Bio. Sci.	2.896	0.00071	650	3.3575	3.235	0.00028	438	212	452
L&S-Phys. Sci.	3.149	0.00026	650	3.318	3.691	-0.00048	777	-127	240
L&S-Humanities	2.505	0.00137	650	3.3955	3.066	0.00067	492	158	320
L&S-Soc. Sci.	3.191	0.00033	650	3.4055	2.686	0.00123	585	65	605
									2439

Weighted difference =	31
Weighted difference w/o Physical Sci.=	48
Weighted difference of math.-based subjects=	-304
Weighted difference of non-math. subjects=	132

Fig. 4: College GPAs By SAT Scores for Men and Women Majors in the Social Sciences Division Who Entered as Freshmen Between 1986 and 1988



Females: Solid Line & Dots

Males: Dashed Line & Dashes

N = 605

musician or socio-economically disadvantaged). A score of 650 is a bit above the median for this group and reduces the problem of distortions in the estimate being introduced by compensatingly high scores on the other components of the AIS. At this score the summary table conforms best to what can be seen by inspecting all the smoothed line scattergrams individually. The modal pattern is one of predictions biased against women in the range where decisions are actually being made. Figure 4 for the Social Sciences in L&S is representative of this modal pattern. In the left tails of the smoothed lines, where the number of cases is smaller and where compensating judgments are being made, the bias seems to favor women. But in the dense range of combined verbal and math scores of 1200 to 1400, where the decisions are being made mechanically, women's performance is clearly under-predicted by the tests.

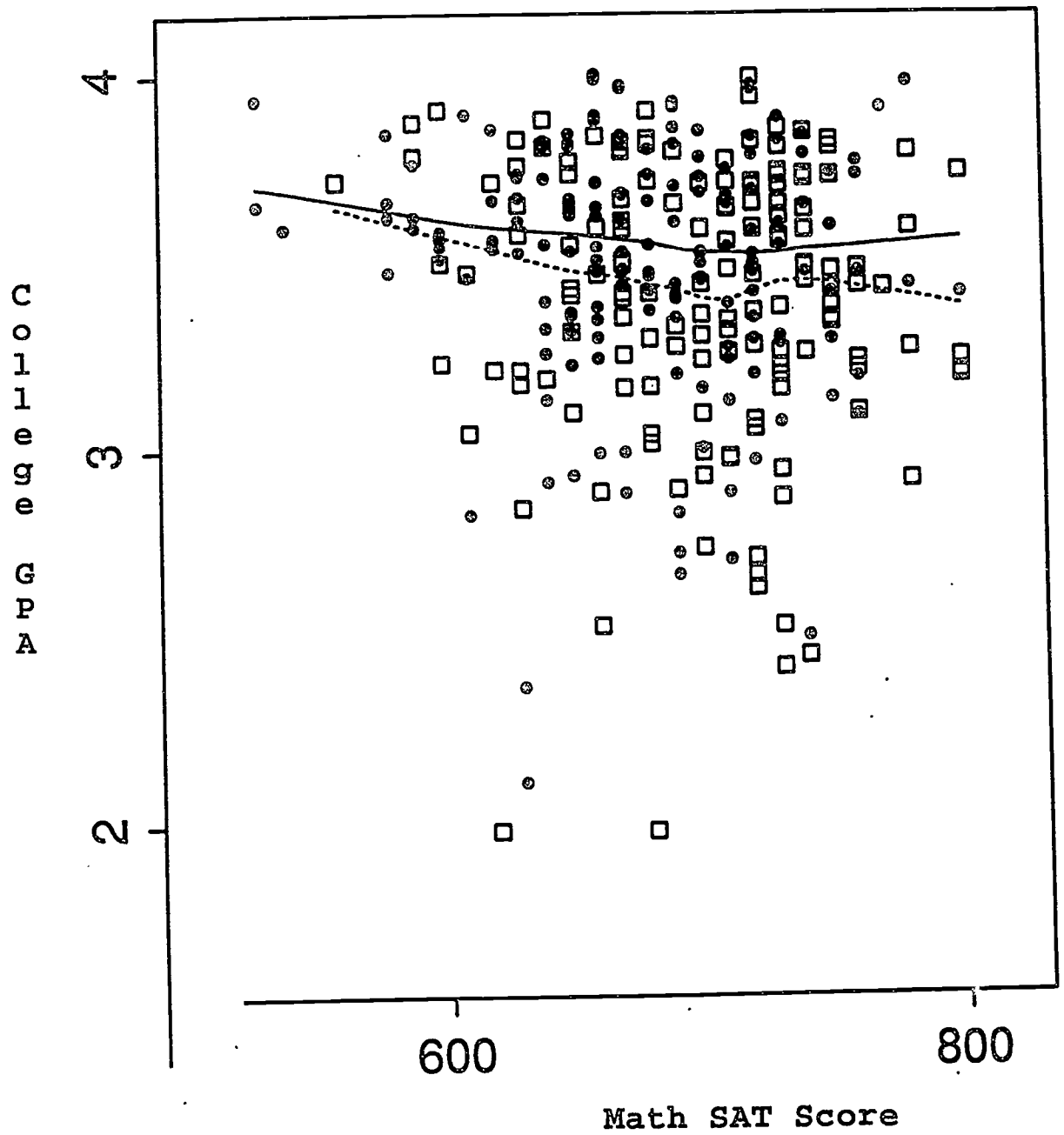
It is helpful to focus on the apparent exceptions to the general test bias against women, however, for in doing so our conclusions are strengthened still further. Note that the fields in which the tests seem to discriminate against men are Engineering and the Physical Sciences for the Verbal SAT and Environmental Design (Architecture) and Humanities for the Math SAT. These are all fields in which any thinking admissions officer would heavily discount these particular tests anyway. Furthermore, the problem for the Physical

Sciences with the Verbal SAT and for the Humanities with the Math one is that the factors confounding women's choices of these majors are so strong that these particular tests are negatively related to college GPA (Figure 5). Despite this negative relationship an inspection of the scattergrams indicates that women still are outperforming men with similar test scores in the range where the decisions are being made. (The scattergram for the Verbal SAT in the Physical Sciences demonstrates this same point, although less vividly.)

In general one would say that it is a mistake to put much if any weight on the Math SAT for applicants expecting to go into the Humanities and on the Verbal one for those going into Engineering. Where tests are relevant, gender bias in the vicinity of 50 to 100 points seems the norm. In the mathematically-based disciplines this translates into women's having a GPA .14 points better than men with the same score of 650 on the Mathematics SAT. For the non-mathematical fields, the same score of 650 on the Verbal SAT yields a female advantage of .08 GPA points.

The evidence that the SATs give biased predictions of the performance of women in Engineering and Chemistry is weak (Figure 6). At first one is tempted to argue that one could continue to use the tests in these fields without worrying about gender bias. But then how do we account for the clear under-prediction of female grades given by these tests in the

Fig. 5: College GPAs By Math SAT Scores for Men and Women Majors in the Humanities Division Who Entered as Freshmen Between 1986 and 1988

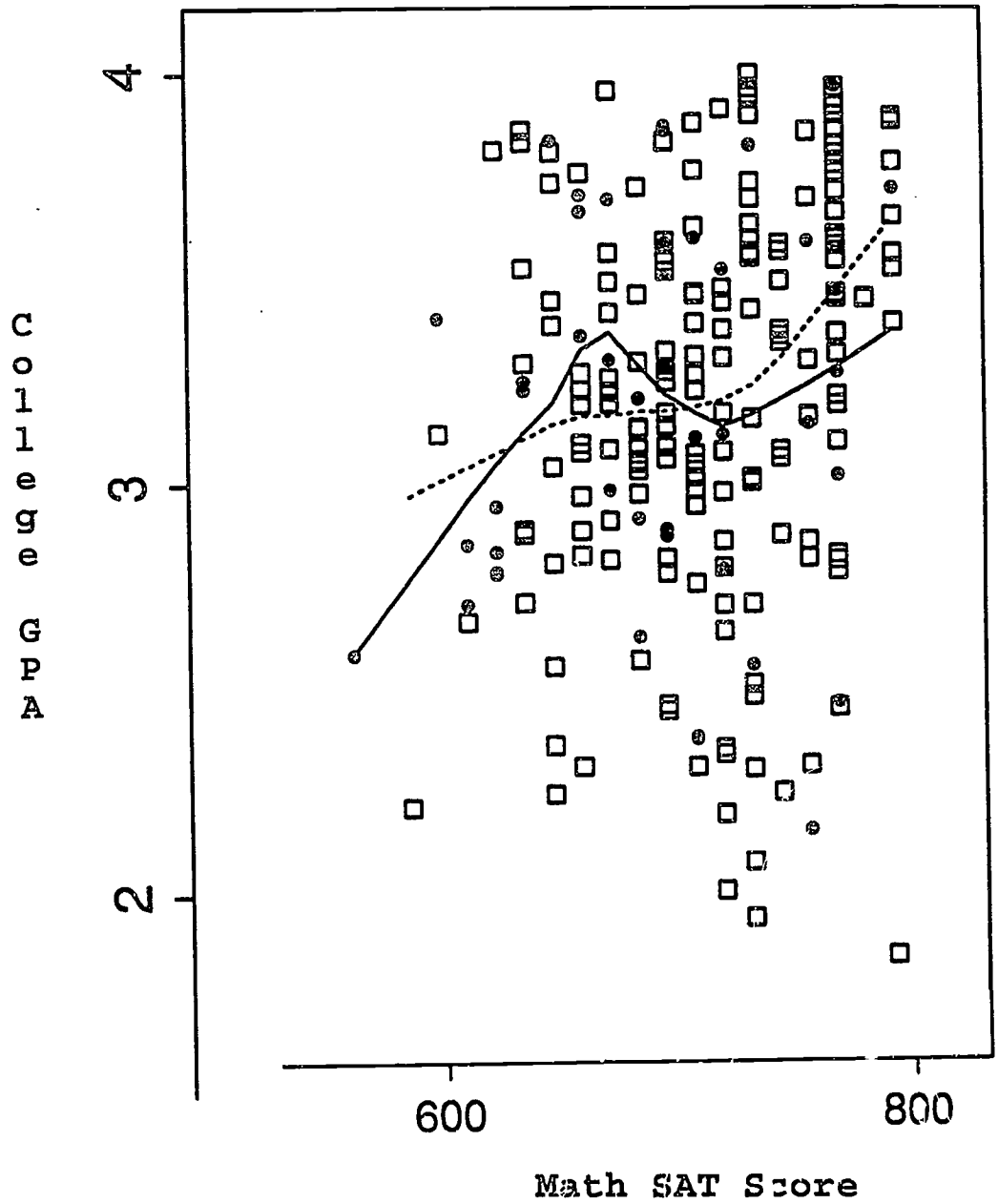


Females: Solid Line & Dots

Males: Dashed Line & Dashes

N = 320

Fig. 6: College GPAs By Math SAT Scores for Men and Women Majors in Engineering Who Entered as Freshmen Between 1986 and 1988



Females: Solid Line & Dots

Males: Dashed Line & Dashes

N = 222

Biological and Physical Sciences? With these powerful "exceptions" one can hardly argue that the Math SAT is unbiased for disciplines that are mathematically based. Is the explanation instead that Engineering is the only mathematically-based field at Berkeley that has an affirmative action program for women?

INAPPROPRIATE CORRECTIVES

Given that the AIS and the SATs are generally gender biased, what is the appropriate corrective? Berkeley's Academic Index Score is based on a set of weights between high school grades and the various College Board tests that are purely arbitrary (from the point of view of prediction). The obvious place to begin a search for correctives is to alter the relative importance attached to the various indicators used. This also has been the most common attempted remedy historically. For example, when the Federal courts ruled that the State of New York discriminated against women by using the SATs to award scholarships, the corrective was to require that HS GPA and SATs be used as co-equal criteria. The apparent reasoning was that since men do better on the SATs and women do better in high school grades, marrying the two would balance out the bias in each (Holden, 1989). This logic is faulty, however. The fact that women's secondary school GPAs show a higher average than men's does

not necessarily mean they are biased predictors of college performance. As we indicated above, our aggregate data indicate that high school grades are almost perfectly gender neutral as a predictor of Berkeley GPA. Thus it is impossible to correct the gender bias in the SATs with a compensating weight for HS GPA. At any weight less than 100 per cent (i.e., the elimination of the SAT as a criterion) the biasing effect of the College Board examinations might be diminished but would still remain.

A second attempted remedy has been to alter the weights assigned to the various College Board tests themselves. For example, a study done on all freshmen enrolling in 1988 at the nine campuses of the University of California found that it could eliminate the gender bias in the prediction of the first year GPA by using a regression formula that weighted each SAT, the three Achievement tests, and HS GPA separately. Inspection of the formula reveals that in order to predict women's college performance accurately it was necessary to give overwhelming influence to HS GPA and the English Achievement Test (Kowarski, 1993). The latter is the least biased instrument in the College Board repertoire. When the Verbal SAT was changed in the 1972 it was tilted in favor of men (Linn, 1992: 15; Petersen & Dubas, 1992: 124; Wellesley, 1992: 55).

There are three reasons why the remedy of weighting test

by test is inappropriate--two of principle and one of infeasibility. First, in order to undo gender bias one would be giving almost no weight to tests of knowledge and skills in mathematics and science. This would be inappropriate for many fields of study and it would send an extremely unfortunate message to today's high schoolers about what their study priorities should be. Second, the tremendous emphasis that this method places on English language skills would put non-native speakers at a severe disadvantage. Even if this were pedagogically justifiable (which is questionable) it would be politically and socially unacceptable.

Third, however, we doubt that this weighting method would be feasible when widely applied. The various College board test scores are highly inter-correlated (with r values at Berkeley ranging from .51 to .82). The multi-colinearity makes any regression coefficients attached to them in a prediction equation unstable. The formula we developed with the data for one entering class was not the appropriate one for use with the next. Thus the ex post fitting of a regression equation to the grades of one class gives one an illusory sense of the precision of one's ability to correct biases and predict appropriately for later classes on an ex ante basis--which is what admissions decisions are supposed to do after all.

Our conclusion is that <u>if</u> admissions decisions are to be made in a purely mechanical manner, the only statistically justifiable procedure--at least for Berkeley--is to sum the scores of all the College Board tests being used, effectively giving them equal weights. This is consistent with present practice and also has the considerable administrative advantage of being the simplest thing to do. (It is appropriate, however, to weight differentially HS GPA and the sum of College Board test scores. Comparision of multiple regression equations for each of the admission years suggest that the appropriate weighting for the highly competitive process at Berkeley would best approximated by [HS GPA times 1000] plus [2 times the sum of five tests].)

A third remedy to the gender bias in the SATs would be to cease using them altogether. Some of those concerned with equity advocate this practice and it has been implemented in a number of college admissions processes (Petersen & Dubas, 1992: 125). We do not find this remedy optimal for a highly competitive institution such as Berkeley. The SATs may be a biased instrument but they do improve a college's ability to predict the performance of applicants. High school GPA alone predicts 21 per cent of the variance in Berkeley GPA; the predicted variance rises to 24 per cent with an appropriately weighted use of the sum of the College Board scores (which as we suggested above is double the HS GPA). Three percentage

points of variance may not seem like very much, but at an institution such as Berkeley where large numbers of highly qualified applicants compete for admission and HS GPAs are tightly packed, it probably makes a difference in at least 1000 admissions decisions.

A fourth remedy to the gender bias in the College Board examinations is to correct for it mathematically in the admissions formula(e) used to rank order applicants. MIT uses separate formulae for each sex in order to predict MIT GPA and then uses that GPA to rank order candidates for admissions. The College Board appears to endorse this method (although not very aggressively; two colleagues of ours in admissions who serve on national advisory boards for the College Board/ ETS were unaware of it before we brought it to their attention). However this particular methodology requires a fairly technical approach to admissions and the resources for campus specific validity studies (College Board, 1991: 2; Petersen & Dubas, 1992: 125-26). This is beyond the means and capabilities of most institutions. Furthermore, even if the technical resources were available our own attempt at such a validity study at Berkeley is not encouraging. We found that although gender bias was consistently present, the patterns and magnitudes it took varied from year to year and from discipline to discipline. We were unable to find a mathematical corrective in which a

responsible and careful statistician could have any confidence.

When such a purely mathematical corrective to the gender bias in the SATs was mooted at Berkeley, it also provoked an interesting debate on the faculty committee which oversees the admissions process. Although such a correction would have been designed to restore a "level playing field" to women, it was felt that it would seem to the public that something "special" were being done to "help" women overcome a deficiency inherent to their sex. Thus the unintended consequence could be to denigrate the capabilities of women and to attract public criticism to the measure. The Office of the General Counsel of the University of California [system] advised the committee that the peculiarities of the case law and legal environment on this issue made a suit likely if such a simple "quick fix" on the AIS were attempted. The only case law on this subject is the New York Regents Scholarship one, where the court accepted the equal use of HS GPA and the SATs as a remedy. Something being done "for" women would probably attract a suit by aggrieved male applicants. They probably would lose but the suit would be expensive and politically inconvenient.

APPROPRIATE CORRECTIVES

Public perceptions and the intractability of estimating

the gender bias in the tests consistently and accurately has led us to recommend a qualitative solution. Berkeley bases only 50% of its admissions decisions of the AIS alone. The other half of the admits follow a full reading of the applicant's file, from which qualitative judgments are made. The problem was that for the College of Letters and Science (where most of the students are admitted) the only white and Asian files read (save for athletes and the socio-economically disadvantaged) were those 120 AIS below the automatic-admit cut-off. Since the bias against women in the AIS is about 150 for L&S, it could not possibly be corrected in a "read pool" of this size--even if not a single male were admitted in this part of the process.

Prior to our study a "read pool" 120 AIS points wide had seemed acceptable, given the resource constraints created by the extremely large numbers of applicants at the highly competitive public universities. The College Board had advised that its tests each had a standard error of measurement of about 34 points and had urged that differences in scores less than this amount not be used to make admissions decisions (1984, p.34). When Berkeley decided to work with a "read pool" 120 points wide, it was adding 44 points to the 76 that accumulating a <u>random</u> error of 34 across 5 tests would dictate--apparently a safe margin. Our study shows, however, that the College Board's confidence

interval contains not only random measurement errors but a systematic gender bias as well. This latter element (unlike the former) accumulates across tests and exceeds the Board's statement of the tests' error margin when several scores are added together. Thus for a school such as Berkeley that uses five College Board tests the "read pool" for males should never be less than 76 points wide but the only certifiably safe size for the female "read pool" would be 170 points.******* A college might choose to work with larger margins than these but the female margin would always need to be 100 points wider than the male one.

Once admissions staff have selected such a "read pool" they must proceed to examine each individual file for the variety of other factors that might indicate that the applicant will be a good student and/ or make a special contribution to campus life. There is some evidence that full attention to these other indicators of excellence corrects the bias of the SATs (Stricker, Rock & Burton, 1991).

Such a qualitative solution to the SAT gender bias problem seems to us to be the best that any admissions office

---

*******If the standard errors are independent of one another (as they seem to be for men), the standard error for their combination would be the square root of the sum of the squares of the individual standard errors, the square root of 5 times 34 squared, or 76. But any systematic errors (such as gender bias) would be additive, thus requiring 5 times 34, or 170.

acting on its own can do. Nonetheless it is cumbersome and imprecise -- and it is expensive. The best estimate is that at Berkeley it will cost at least $150,000 in additional resources. For beleaguered public institutions this is not a trivial sum.

There is a still better solution to this problem -- that the Educational Testing Service correct the gender bias in its tests itself. The sources of the discrimination against women in these tests is a subject of considerable debate in the educational psychology literature (e.g., Linn, 1992; Wellesley, 1992) and is beyond our professional competence. Correction certainly is possible for at least some of the exams, through altering the mix of test items (Burton, 1995). It was done in favor of men by changing the mix of questions when the Verbal SAT was revised in 1972.

Correction is most urgent for the SATs, which are justified only by their ability to predict college grades, than it is for the Achievement tests, which can be defended as measures of student knowledge that may or may not predict college performance. It seems possible that the new format for the SATs has lessened the gender bias on the Verbal SAT but it remains for the Math one (Burton, 1995). Since correction of gender bias was not among the objectives for the change it is very unlikely to have been eliminated entirely. Given the history of the College Board's

obfuscation and avoidance on this issue for the last 50 years, users would be unwise to assume that a good faith effort will be made to correct the problem without external prodding. The Education Testing Service should be asked to undertake a study such as ours to prove the truth of any assertion that the problem no longer exists.

The financial welfare, academic opportunities, and sense of self-worth of female students depend on further changes being made in the SATs. It is urgent that, after a half century of delay, the College Board finally mandate them. In the meantime, the Board owes it to women and to the educational community to provide all institutions employing its tests an unambiguous, highly visible "users' warning label" that these tests contain a gender bias and that their appropriate use requires the kind of qualitative approach recommended above. If the College Board does not undertake these measures itself, we imagine that it will not be long before the courts order it to do so.

REFERENCES

Burton, N. (1995) "How Have Changes in the SAT Affected Women's Mathematics Performance?" Paper presented to the annual conference of the American Educational Research Association, San Francisco.

California Postsecondary Education Commission (1992) "Eligibility of California's 1990 High School Graduates for Admission to the State's Public Universities: A report of the 1990 High School Eligibility Study" (Sacramento: California Postsecondary Education Commission).

Cleveland, W.S. (1979) "Robust Locally Weighted Regression and Smoothing Scatterplots," Journal of the American Statistical Association, Vol. 74, No. 368: 829-36.

College Entrance Examination Board (1984) "College Board Technical Manual of the Scholastic Aptitude Test and Achievement Tests" (New York: College Entrance Examination Board).

College Entrance Examination Board (1991) "Testing 101: A Short Course on the SAT" (New York: College Entrance Examination Board).

Elliott, 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, Vol. 25, No. 4, pp. 333-47.

Holden, Constance (1989) "Court Ruling Rekindles Controversy Over SATs," Science, Vol. 243, pp. 885-7.

Jencks, Christopher and James Crouse (1982) "Should we Relabel the SAT ... or Replace It?" In W. Schrader (ed.), New Directions for Testing and Measurement: Measurement, Guidance, and Program Improvement (San Francisco: Jossy-Bass), pp. 33-49.

Lavin, D.E. (1965) The Prediction of Academic Performance (New York: Russell Sage Foundation).

Linn, Marcia (1992) "Gender Differences in Educational Achievement," in Sex Equity in Educational Opportunity, Achievement and Testing: Proceedings of the 1991 ETS Invitational Conference (Princeton: Educational Testing

Service), pp. 11-50.

Linn, R.L. (1973) "Fair Test Use in Selection," Review of Educational Research, Vol. 43: 139-61.

Rigol, Gretchen Wyckoff (1989) "Why Do Women Score Lower than Men on the SAT?" College Prep, No. 4; Reprinted by the College Entrance Examination Board (New York).

Stricker, Lawrence, Donald Rock and Nancy Burton (1991) "Sex Differences in SAT Predictions of College Grades," College Board Report No. 91-2 (New York: College Entrance Examination Board).

Thorndike, R.L. (1963) The Concepts of Over- and Underachievement (New York: Bureau of Publications, Teachers College, Columbia University).

Wagner, M.E. and E. Strabel (1935) "Homogeneous Grouping as a Means of Improving the Prediction of Academic Performance, Journal of Educational Psychology, Vol. 19: 426-46.

Wainer, Howard and Linda S. Steinberg (1991) "Sex Differences in Performance on the Mathematics Section of the Scholastic Aptitude Test: A Bidirectional Validity Study," Research Report No. 91-45 (Princeton: Educational Testing Service).

Wellesley College Center for Research on Women (1992), The AAUW Report: How Schools Shortchange Girls: A Study of Major Findings on Girls and Education (Washington, D.C.: AAUW Educational Foundation and National Education Association).

Willingham, W.W., C. Lewis, R. Morgan, and L. Ramist (1990) Predicting College Grades: An Analysis of Institutional Trends Over Two Decades (Princeton: Educational Testing Service).