ED 397 074 TM 025 069

AUTHOR	Bridgeman, Brent; Lewis, Charles
TITLE	Gender Differences in College Mathematics Grades and SAT-M Scores: A Reanalysis of Wainer and Steinberg.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-95-17
PUB DATE	Jul 95
NOTE	34p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Calculus; *College Students; *Grades (Scholastic); Higher Education; *Mathematics Achievement; *Prediction; Predictive Validity; Scores; *Sex Differences; Test Results
IDENTIFIERS	*Scholastic Aptitude Test

ABSTRACT

H. Wainer and L. Steinberg (1992) showed that within broad categories of first-year college mathematics courses (e.g., calculus), men had substantially higher average scores on the mathematics section of the Scholastic Aptitude Test (SAT-M) than women who earned the same letter grade. However, Wainer and Steinberg's analysis may lead to unwarranted conclusions in that they: (1) focused primarily on differences in SAT-M scores given course grades when the more important question for admissions officers is the difference in course grades given scores on the predictor; (2) failed to account for differences among calculus courses; and (3) focused on the use of the SAT-M as an isolated indicator. Reanalysis, taking distinctions among courses into consideration, suggests that Wainer and Steinberg's estimates were too large by about 10 points in calculus courses, although estimates for precalculus courses are confirmed. The reanalysis, which concentrated on 43 colleges, indicated that a more appropriate composite indicator made up of both SAT-M and high school grade point average demonstrated minuscule gender differences for both calculus and precalculus courses. (Contains 10 tables and 9 references.) (Author/SLD)

RESEARCH REPORT

GENDER DIFFERENCES IN COLLEGE MATHEMATICS GRADES AND SAT-M SCORES: A REANALYSIS OF WAINER AND STEINBERG

Brent Bridgeman
Charles Lewis

Educational Testing Service
Princeton, New Jersey
July 1995

Gender Differences in College Mathematics Grades and SAT-M Scores: A Reanalysis of Wainer and Steinberg

Brent Bridgeman
and
Charles Lewis

Educational Testing Service

Abstract

Wainer and Steinberg (1992) showed that within broad categories of first-year college mathematics courses (e.g., calculus) men had substantially higher average scores on the mathematics section of the SAT (SAT-M) than women who earned the same letter grade. In calculus courses, the median difference over the five letter grades was about 38 points on the 200-800 SAT-M scale. However, three aspects of their analyses may lead to unwarranted conclusions. First, they focused primarily on differences in SAT-M scores given course grades when the more important question for admissions officers is the difference in course grades given scores on the predictor. Second, they failed to account for differences among calculus courses. Because different calculus courses may be differentially selected by men and women (e.g., calculus for engineers vs calculus for liberal arts students), and because these courses may have different grading standards, the way Wainer and Steinberg aggregated information without regard to the specific course taken could exaggerate gender differences. The reanalysis, taking distinctions among courses into consideration, suggests that their estimates were too large by about 10 points in calculus courses, although their estimates for pre-calculus courses were confirmed. Most important, Wainer and Steinberg focused on the use of SAT-M as an isolated indicator; such use is contrary to professional recommendations. The reanalysis indicated that a more appropriate composite indicator made up of both SAT-M and high school grade point average demonstrated minuscule gender differences for both calculus and pre-calculus courses.

Gender Differences in College Mathematics Grades and SAT-M Scores: A Reanalysis of Wainer and Steinberg

In a clear and provocative article, Wainer and Steinberg (1992) showed that men and women who earned the same grade in first-year college mathematics courses had SAT-M[1] scores that differed, on average, by about 33 points. For example, men who received a B grade in a calculus course had an a mean SAT-M score of 615 while women with the same grade had a mean SAT-M score of 580, or a mean difference of 35 points. They concluded that "it is a capital mistake to use the SAT-M in isolation for decisions involving comparisons between the sexes" (p. 334). This conclusion is probably correct and is consistent with previous findings (e.g., Bridgeman & Wendler, 1991). Nevertheless, a casual reading of Wainer and Steinberg can easily lead to a misunderstanding of the nature and extent of the problem. Three aspects of their analysis require reexamination. First, they focused most attention on differences in SAT-M scores given grades. Although they acknowledged that admissions officers are more interested in predictions in the other direction (i.e., differences in course grades given scores on a predictor), they provided no estimates of the size of these grade differences. Second, they aggregated data across courses in a way that could artificially exaggerate gender differences. Third, and probably most important, they showed the consequences of less than optimal use of SAT-M as a sole predictor but provided no information on gender differences when SAT-M was used appropriately in the context of other information. They recognized the possibility of the latter two problems (in the discussion on p. 330 and the footnote on p. 331), but choose not to present any data that would illuminate them.

In their footnote on p. 331, Wainer and Steinberg acknowledge the data aggregation problem as follows:

> Another possibility is a statistical artifact (the previously mentioned Simpson's paradox) caused by aggregation over different colleges and/or different classes. Suppose, for example, women predominate at one college in the sample and men at another; suppose further that higher grades are easier to obtain in calculus at the former school than at the latter. Aggregating over the two schools could yield the results obtained here. The same effect might occur at the classroom level as well. We investigated this possibility to some extent. One school in the sample provided a large proportion of the students. The results of that school alone are consistent with those from an aggregation over the schools in the rest of the sample. Making similar comparisons school-by-school, or worse, classroom-by-classroom, is difficult because of the shrinking statistical power associated with the shrinking sample sizes. The obvious statistical tests are not too helpful since we would be interested in accepting the null hypothesis. At this initial stage, we must content ourselves with the possibility that inappropriate aggregation is a conceivable explanation, but we believe that it is not a particularly likely one.

Thus, they considered the possibility of classroom-level aggregation problems but decided not to pursue them because they saw no evidence of school-level effects and because of reduced power in small samples. However, both of these reasons for deciding not to pursue classroom or course-level effects are questionable.

Failure to find a school-level aggregation problem is only marginally relevant to possible course-level problems. There is little reason to believe that women predominate at colleges with low grading standards while men predominate at colleges with high standards. But there is reason to expect that within colleges men and women may be differentially attracted to different courses and that these courses may differ considerably in the strictness of their grading standards (Ramist, Lewis, & McCamley-Jenkins; Elliott & Strenta, 1988). Many colleges have one calculus course for engineering and science majors, and a different course for other students. Furthermore, the science/engineering calculus course is likely to have both a higher proportion of men and stricter grading standards. This phenomenon can be seen in Table 5 (p. 281) of Bridgeman and Wendler (1991). In the two colleges that identified separate calculus courses for engineering and science majors, men predominated in those courses (over 2.5 to 1), mean grades were lower than in the other calculus courses, and mean SAT-M scores (of both men and women) were higher. Considering both the science/engineering calculus courses and the other calculus courses as a single course (as Wainer and Steinberg did) would tend to exaggerate SAT-M gender differences for each letter grade.

Low statistical power might be a problem for analysis of a particular course within a single college. However, when within-course gender differences are aggregated across over 100 courses, substantial power remains and the results can be reported with considerable confidence.

Although Wainer and Steinberg note that the stated purpose of the SAT is to supplement the school record, they focus the SAT-M as a sole predictor. This analysis is instructive, but gives the reader no idea how SAT-M is related to gender differences when it is used as intended.

The current reanalysis augments the Wainer and Steinberg results by analyzing gender differences within individual courses within colleges, by evaluating gender differences when the SAT-M is used as intended in combination with the high school record[2], and by emphasizing the analysis that must actually be used in any selection or placement decision--namely, predicting college grades from information available at the time of the decision, such as high school grades and test scores.

It should be noted that the relationships between SAT-M scores (and HSGPA) and grades in specific mathematics courses are discussed here only to illustrate how gender differences in test scores are related to gender differences in grades. Although it is informative to focus on predictions in specific courses, this should not suggest that SAT-M should be used for course-level placement decisions. SAT-M was designed to supplement other information for predicting overall success in the first year of college, not to make placements into individual mathematics courses. Other measures are probably better suited to the placement function (Bridgeman & Wendler, 1989).

Method

Subjects

Most of the data analyzed by Wainer and Steinberg came from a data set of colleges collected and described by Ramist, Lewis, and McCamley (1991). Data from the freshman classes of 1982 and 1985 were available from 38 colleges, and an additional 7 colleges provided data only for the freshman class of 1985. For the reanalysis, we focused on only the 1985 cohort, thus avoiding any potential complications from shifting grading standards over time. This data set was ideal because it separated course grades not only into the broad categories used by Wainer and Steinberg (e.g., calculus, pre-calculus) but also into specific course titles (e.g., calculus for engineers, or calculus for liberal arts students). For 12 of the colleges, course grades were reported separately for different sections of the same course. In order to simplify the analyses and discussion, only courses in the two largest categories (calculus and pre-calculus) were considered. These courses account for about 69% of the students in the sample who take mathematics in the freshman year. The two colleges in the sample that enrolled only women were dropped, for a final sample of 43 colleges.

In addition to SAT scores and course grades, the data set included high school grade point averages based on self-reported grades in a set of core academic courses; these grades were reported by students in the Student Descriptive Questionnaire which students completed when they registered to take the SAT. Previous research suggests that students report their grades fairly accurately, and underreporting or overreporting is unrelated to gender (Freeberg, 1988).

Procedures

First, in a replication of the Wainer/Steinberg procedure, the mean SAT-M of all women who got an A in any calculus course was subtracted from the mean SAT-M of all men who got an A in any calculus course[3]. This procedure was repeated for grades of B, C, D, and F; for these analyses, pluses or minuses appended to the letter grades were ignored. Next, gender differences were computed separately within each calculus course in each college (or, for the 12 colleges where the data were available, within each section of each course). The mean SAT-M for women who got an A in the course was subtracted from the mean SAT-M of men who got an A in the course. These differences were weighted by the total number of students with As in the course (or course section) and these weighted differences were averaged across courses. This procedure was repeated for the other letter grades.

For comparison, the above procedures were repeated using the high school grade point average (HSGPA) instead of SAT-M. To make the units comparable, HSGPA was standardized to the same mean and standard deviation as SAT-M[4]. This was done separately for students in calculus and pre-calculus courses. Thus, a 10 point difference on the standardized HSGPA scale is in this sense comparable to a 10 point difference on SAT-M.

Next, the across and within college analyses were repeated with a composite score that was formed from SAT-M and HSGPA. Within each course, a regression equation was used to find the optimal weighting of SAT-M and HSGPA for predicting course grades. The regression weights were averaged across courses to create a single regression equation. This equation was used to create a composite score for each student. As with the HSGPA, the composite scores were standardized to be comparable to SAT-M.

The above analyses were then repeated for pre-calculus courses[5].

The first set of analyses investigated gender differences in various scores (SAT-M, HSGPA, and composite) for each letter grade. The next set of analyses used essentially the same methods to investigate differences in grade averages at various levels of the predictors. That is, how different are the grades of men and women who have approximately the same SAT-M scores or the same scores on the composite of SAT-M and HSGPA? The former set of analyses parallels the retrospective analyses presented first by Wainer and Steinberg; the latter set is the one of primary relevance for the use of scores and grades for admissions purposes.

Results and Discussion

Sample sizes and mean SAT-M scores for calculus courses are presented in Table 1. Grades of D and F were relatively rare; fewer than 20% of both men and women received those grades. In a number of courses no grades of D or F were given, and sample sizes for the within-course analyses were reduced for those grades because computation of a within-course gender difference required at least one man and one women at a given letter grade level.

The first column of numbers in Table 2 shows the difference in SAT-M scores of men and women for each calculus grade when differences are computed across courses (i.e., the Wainer/Steinberg method). Except for minor fluctuations caused by differences in the samples, these numbers are directly comparable to the second column of numbers in Wainer and Steinberg's Table 3 (p. 328); the median of 38 is identical to the median in the Wainer/Steinberg table. The next column in Table 2 shows the gender difference in SAT-M scores when the difference is computed within each calculus course and these differences are averaged over courses. The averaged within course differences were 8 to 14 points smaller than the differences computed by the Wainer/Steinberg method. Thus, about one-quarter of the gender difference observed by Wainer and Steinberg appears to be an artifact of the way men and women sorted themselves into courses with different grading standards. However, because about three-quarters of the difference persisted, the current results indicate that even within individual courses, on average, men have higher SAT-M scores than women with the same course grades. A supplemental analysis of just those colleges that provided information on grades within different sections of the same course mirrored the results of the within-course analyses. Thus, although section-level data may be important for other purposes, course-level information appears to be sufficient (and necessary) for estimating gender differences.

The next two columns in Table 2 show the analogous results for the standardized high school grade point average. The negative signs indicate that HSGPAs of women are higher than those of men with the same calculus grades. Because these differences are of nearly the same size as the differences for SAT-M, but in the opposite direction, the composite of SAT-M and HSGPA might be expected to show virtually no difference between men and women at each grade level. The last column of the table shows that is exactly what happened. Note that these numbers are on the same type of 200-800 scale as SAT-M, and are therefore indeed minuscule. These results highlight Wainer and Steinberg's contention that it is a "capital mistake to use SAT-M in isolation" (p.334), but they also indicate that it might be an equally capital mistake to use the HSGPA in isolation. It is only the combination of SAT-M and HSGPA that produces the near-zero differences.

Table 3 reports results on SAT-M (comparable to the first two columns of data in Table 2) for three levels of college selectivity, where selectivity is defined in terms of average scores on the combined verbal and mathematical portions of the SAT. These levels were originally defined to split the sample of colleges roughly into thirds (see Ramist, Lewis, & McCamley, 1991). Results mirror those in Table 2, and there did not appear to be any important differences associated with selectivity.

Table 4 reports results on the standardized composite for colleges in the same three selectivity levels. For the within-course analyses, gender differences were very small for all three groupings of colleges. Within each letter grade level in the least selective colleges, women had slightly higher scores than men on the standardized composite. This may in part reflect the greater potential of high school grades to indicate gender differences in these colleges; in the most selective colleges, grades of both men and women were near ceiling levels. In the most selective colleges, 47% of the men and 57% of the women had HSGPAs of 3.75 or higher. In the least selective colleges, the comparable percentages were 18% for men and 23% for women.

Sample sizes for the pre-calculus courses are presented in Table 5. Numbers of students and numbers of courses were considerably lower than they were for calculus courses, and Ds and Fs were slightly more prevalent. Also, unlike the calculus sample, women slightly outnumbered men.

Table 6 is the pre-calculus version of Table 2. The most striking difference for the pre-calculus courses, compared to the calculus courses, is the essential equivalence of the across-course and within-course analyses. This probably reflects the fact that special pre-calculus courses for science and engineering students are rare; most science/engineering students would have taken such a course in high school. If men and women are not differentially selecting different courses, across-course and within-course differences would be more comparable. As with the calculus courses, gender differences for HSGPA were of about the same size, but opposite direction, as the SAT-M differences, and the differences for the standardized composite were quite small.

Sample sizes and mean grades in calculus courses for the prospective analyses are presented in Table 7. The table shows the calculus course grades for different levels of SAT-M, standardized HSGPA, and the standardized composite. These analyses are most relevant for the usual prediction problem-- predicting course grades from test scores and high school grades. Although it might seem that gender differences in course grades given predictor scores should be comparable to differences in predictor scores given course grades, there is no guarantee that the size, or even the direction, of such differences will be comparable. Recall that, for any two variables X and Y that are not perfectly related (such as test scores and grades), the regression of *Y* on *X* is not the same as the regression of *X* on *Y*. This means that there are two distinct regressions for female students and two for male students. In general, the relationship of the regressions of *X* on *Y* for female and male students says nothing about the relationship between the regressions of *Y* on *X* for the two groups. Consequently, if the primary concern is with underprediction of grades from test scores for female students, nothing relevant is learned by considering the postdiction of test scores from grades.

The relatively large number of students in the top HSGPA category reflects the skewed distribution of high school grades in this select sample; in the original (0–4) grade scale, over 8000 students reported HSGPAs of 3.9 or higher. Also note the differences in the relative numbers of men and women in the extreme groups when extremes are defined by HSGPA instead of SAT-M. Men predominate in the top group defined by SAT-M and in the bottom group defined by HSGPA.

The grade averages were computed across courses (without regard to course distinctions). These numbers should be interpreted cautiously because courses with different grading standards are grouped together. Thus, for example, it could be argued that within the 650-699 score range women might have higher grades than men because they are more likely to enroll in courses with easier grading standards. However, the averaged within-course gender differences, which are not subject to this potential bias, were nearly as large as the across-course differences (see Table 8 [the prospective analog of Table 2]); for the highest and lowest categories the within-course differences were actually slightly larger. Thus, differential enrollment patterns that seriously biased across-course estimates of score differences given grades had a minimal impact on grade differences given scores. This analysis again emphasizes the non-symmetric nature of these relationships; what is true for score differences given grades is not necessarily true for grade differences given scores.

The center columns of Table 8 show differences based on the standardized HSGPA. All differences were small and some were positive (higher grades for males) while others were negative. The right columns show the differences for the standardized composite. Although all differences were negative, they were about half the size of the differences based on the SAT-M alone, and were fairly small in absolute terms (median difference of less than one-tenth of a point on the 5-point [0-4] grade scale).

Although gender differences in the selected sample would be minimized if decisions were based on the HSGPA alone, a mere lack of gender difference is not a rational basis for selecting a prediction instrument. For both men and women, the composite score provides better prediction than either SAT-M or HSGPA in isolation. We are in agreement with Wainer and Steinberg that if balanced gender representation is a societal value, it should be recognized by explicit acknowledgement in the selection process, and not by using a less than optimal selection score simply because it yields the desired gender balance in the selected sample.

Although small, the fact that the differences in the last column of Table 8 are not zero may be troublesome to a few readers. These differences indicate that women's grades are underpredicted, that is, women get slightly higher grades than would be predicted from the composite score. However, given only that the two groups differ on the criterion (i.e., women have higher grades than men), underprediction is virtually inevitable for any predictor with a gender difference of about the same size as the gender difference in criterion scores. Although the reasons for this underprediction are complex (Cole & Moss, 1989; Humphreys, 1987; Linn, 1984), it should suffice to note that similar differences could be found with such noncontroversial measuring instruments as tape measures. For example, in a group in which men are taller than women on average, a tape measure would underpredict height for men. If the tape measure were not very precise (say it measured to only the nearest inch) the underprediction could be quite noticeable. Using the typical regression approach, the difference in the predicted scores of two groups is obtained simply by multiplying the score difference on the predictor by the correlation between predictor and criterion[6]. Suppose the correlation between predictor (a measurement made with a tape measure) and criterion (a second measurement made with a different tape measure) were .5. Then, if men and women differed, on average, by two inches on the initial assessment, they would differ by only 1 inch (.5 x 2) on predicted scores on the criterion. If the actual average criterion scores of men and women differed by two inches (just as the original predictor measurements had), there would be a 1 inch discrepancy between the difference in mean predicted scores and the difference in mean actual scores. That is, men's scores would be "underpredicted" and women's scores would be "overpredicted." With a correlation of .5 between the predictor and criterion (which is fairly typical of the size of the correlation of test scores and grades), underprediction could be avoided only if differences between gender groups (in standard score units) were twice as large on the predictor as on the criterion[7]. This is equally true for tape measures or test scores. Thus, the near-zero gender differences reported for the various levels of the standardized HSGPA do not suggest that gender differences are the same on HSGPA as on course grades; indeed, the standardized difference (favoring females) is three times as large for HSGPA as it is for course grades (the standardized difference was -.10 for course grades and -.30 for HSGPA)[8]. For comparison, the standardized difference was .43 for SAT-M and .07 for the composite.

Tables 9 and 10 are the pre-calculus versions of Tables 7 and 8. The score categories were adjusted (top category 600+ and bottom category <450) to reflect thegenerally lower scores in this group. Once again, gender differences were smallest for the HSGPA categories and substantially smaller for the composite than they were for SAT-M by itself. The median gender difference for the composite score was -.25 in the least selective colleges and -.06 in the most selective colleges, though it should be noted that relatively few students were enrolled in pre-calculus courses at the most selective colleges (244 men and 417 women). Overall, the standardized gender difference for course grades (-.13) was very close to the standardized difference for the composite score (-.08). As in the calculus sample, the standardized difference for HSGPA favored women (-.44) while the difference for SAT-M favored men (.38).

Table 9 dramatically illustrates the differences in using SAT-M by itself or the composite for selecting students. Suppose outstanding students were to be selected and a cut score of 600 or above on SAT-M were used. The 949 students in the selected group would be 61% male, and the mean grades ultimately earned would be 2.60 for the men and 2.86 for the women. If instead the cut score were set at 600 on the standardized composite, the 1249 students selected would be 44% male. Even though a larger group was selected, this group would ultimately be more successful (mean grade of 2.86 for the men and 2.94 for the women) than the group selected by SAT-M alone. Although the relative sizes of the groups selected by the two methods would vary with different cut scores, the basic relationships should hold: the group selected by the composite score would have a higher proportion of women, a higher proportion of students getting high course grades, and a smaller difference between the grades of men and women. Although these generalizations should be equally true for calculus and pre-calculus courses, note that the above illustration assumes that grades for men and women have comparable meaning. This seems to be a reasonable assumption for pre-calculus courses; for calculus courses where this assumption may be in doubt , a similar analysis could be run within individual courses or in courses grouped by grading standards.

Conclusion

The current results show that the gender differences found by Wainer and Steinberg were somewhat inflated by their failure to account for differences among calculus courses, but they support their overall conclusion that SAT-M should not be used in isolation. Although Wainer and Steinberg noted that such isolated use is contrary to the stated purpose of the SAT, they provided no information on gender differences when the SAT-M score is appropriately combined with the high school grade point average to form a composite score. The current results indicated that gender differences in the composite score for each course grade level were minuscule with some favoring men and others favoring women. In the more relevant prospective analyses, focusing on differences in course grades at each level of the composite score predictor, gender differences were also quite small, both relative to differences for the SAT-M alone and in absolute terms.

References

Bridgeman, B., & Wendler, C. (1989). *Prediction of grades in college mathematics courses as a component of the placement validity of SAT-Mathematics scores*. College Board Report No. 89-9. New York: College Entrance Examination Board.

Bridgeman, B. & Wendler, C. (1991). Gender differences in predictors of college mathematics performance and grades in college mathematics courses. *Journal of Educational Psychology, 83,* 275-284.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: Macmillan.

Elliott, R., & Strenta, A. C. (1988). Effects of improving the reliability of GPA on prediction generally and on comparative predictions for gender and race particularly. *Journal of Educational Measurement, 25,* 333-347.

Freeberg, N. E. (1988). *Analysis of the revised Student Descriptive Questionnaire, Phase I: Accuracy of student-reported information*. College Board Report No. 88-5. New York: College Entrance Examination Board.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology, 71,* 327-333.

Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement, 21,* 33-48.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups*. College Board Report No. 93-1. New York: College Entrance Examination Board.

Wainer, H. & Steinberg, L. (1992). Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: A bidirectional validity study. Harvard Educational Review, 62, 323-336.

Author Note

Thanks to Laura Jenkins for managing the data base and performing the analyses, and to Leonard Ramist for supervising the original construction of the data base. Thanks also to Nancy Burton and Howard Wainer who provided a number of useful comments on an earlier draft of this manuscript. The financial support of Educational Testing Service is also gratefully acknowledged. The conclusions and opinions are those of the authors and do not necessarily represent official positions of Educational Testing Service.

Correspondence concerning this article should be addressed to Brent Bridgeman, 09-R, Educational Testing Service, Princeton, NJ 08541. Electronic mail may be sent via internet to bbridgeman@ets.org.

Footnotes

[1] SAT-M refers to the mathematics section of the Scholastic Aptitude Test. This test has recently been revised and renamed and is now a part of the Scholastic Assessment Tests; the research reported here used the older version of the test.

[2] Typically, both SAT-V and SAT-M are used in combination with the high school record to predict overall freshman grade point average. However, for predictions of grades in college mathematics courses, SAT-V does not add useful information once high school grade point average and SAT-M score have been included in the prediction equation.

[3] Many authors prefer to do the subtraction in the other direction (F-M); we have used the M-F direction to facilitate comparisons with Wainer/Steinberg.

[4] In the calculus courses, the SAT-M mean was 598 with a standard deviation of 87; in the pre-calculus courses the mean was 515 with a standard deviation of 76.

[5] The equation, with standardized regression coefficients, was: -2.50 + (SAT-M x .32) + (HSGPA x .33). In the pre-calculus courses, the regression equation was: -2.21 + (SAT-M x .29) + (HSGPA x .34).

[6] If scores on predictor and criterion are measured in different units, they should first be standardized to the same mean and standard deviation. In this example, both predictor and criterion are measured with the same units (inches).

[7] The size of the gender difference on the predictor needed to avoid over- or underprediction is a function of the size of the gender difference on the criterion and the correlation between predictor and criterion. These differences are illustrated in the table below:

Gender difference on criterion in standard deviation (SD) units	Correlation between predictor and criterion	Gender difference on predictor in SD units needed to avoid over/underprediction
1	.3	3.3
1	.5	2.0
1	.8	1.3
1	1.0	1.0
.5	.3	1.7
.5	.5	1.0
.5	.8	.6
.5	1.0	.5

Footnotes (con't)

[8] Standardized scores were computed within each course by subtracting the mean for the females from the mean for the males and dividing the result by the square root of the male variance plus the female variance divided by two. These standardized differences were weighted by the number of students in the course and averaged over courses.

Table 1

Sample Sizes and Mean SAT-M Scores for Calculus Courses

	Males		Females		Number of Courses for Gender Difference[a]
Grade	n	Mean SAT-M	n	Mean SAT-M	
A	4333	647	3460	613	266
B	6009	625	4449	587	284
C	5229	604	3695	567	250
D	1850	586	1354	544	166
F	1862	588	898	544	121
Total	19,283	617	13,856	581	

[a] Number of courses with at least one male and one female at given grade level.

Note.--*N*s reflect total number of course grades; students who took more than one calculus course in their freshman year are counted separately for each course taken.

Table 2

Gender Difference at Each Grade Level on SAT-M, Standardized HSGPA, and Standardized Composite Scores Computed Across and Within Calculus Courses

	Mean Score for Males minus Mean for Females					
	SAT-M		STD HSGPA		STD Composite	
Grade	Across	Within	Across	Within	Across	Within
A	35	21	-17	-23	11	-2
B	38	28	-21	-24	10	2
C	37	29	-20	-21	11	5
D	42	33	-30	-31	7	1
F	44	35	-30	-29	8	4

Note.--Standardized (STD) scores are adjusted to the mean and standard deviation of SAT-M. Composite is weighting of SAT-M and High School Grade Point Average that produced the best prediction of calculus grade.

Table 3

Difference Between SAT-M Scores of Men and Women Computed Across and Within Calculus Courses for Three Levels of College Selectivity

	Mean SAT-M for Males minus Mean for Females					
	Least Selective		More Selective		Most Selective	
Grade	Across	Within	Across	Within	Across	Within
A	31	18	32	21	35	23
B	40	28	40	30	32	25
C	35	26	42	33	31	23
D	37	34	47	34	41	29
F	46	28	47	33	45	43

Note.--Selectivity is defined in terms of mean total SAT score (SAT-V + SAT-M) for the college. Least selective is colleges with total less than 986; more selective is total between 986 and 1121 inclusive; most selective is total greater than 1121. Sample sizes were smallest in the least selective colleges (ranging from 244 people in 26 courses with a grade of F to 720 people in 55 courses with a grade of B), and largest in the more selective colleges (ranging from 1755 people in 61 courses with a grade of F to 6377 people in 106 courses with a grade of B).

Table 4

Difference Between Standardized Composite Scores of Men and Women Computed Across and Within Calculus Courses for Three Levels of College Selectivity

	Mean STD Composite Scores for Males minus Mean for Females					
	Least Selective		More Selective		Most Selective	
Grade	Across	Within	Across	Within	Across	Within
A	-2	-14	9	-2	12	3
B	8	-4	12	3	8	2
C	6	-5	15	8	6	2
D	-4	-6	13	3	9	-1
F	16	-4	10	-1	16	19

Note.--Selectivity is defined in terms of mean total SAT score (SAT-V + SAT-M) for the college. Least Selective is colleges with total less than 986; More Selective is total between 986 and 1121 inclusive; Most Selective is total greater than 1121. Standardized Composite is optimal weighting of SAT-M and High School Grade Point Average adjusted to mean and standard deviation of SAT-M.

Table 5

Sample Sizes and Mean SAT-M Scores for Pre-Calculus Courses

	Males		Females		Number of Courses for Gender Difference[a]
Grade	n	Mean SAT-M	n	Mean SAT-M	
A	596	556	824	538	64
B	1071	538	1244	510	75
C	1092	526	1203	491	83
D	558	509	600	476	61
F	545	507	410	472	44
Total	3862	529	4281	502	

[a] Number of courses with at least one male and one female at given grade level.

Note.--*N*s reflect number of course grades; students who took more than one pre-calculus course in their freshman year are counted separately for each course taken.

Table 6

Gender Difference at Each Grade Level on SAT-M, Standardized HSGPA, and Standardized Composite Scores Computed Across and Within Pre-Calculus Courses

	Mean Score for Males minus Mean for Females					
	SAT-M		STD HSGPA		STD Composite	
Grade	Across	Within	Across	Within	Across	Within
A	18	17	-30	-30	-11	-11
B	28	25	-27	-25	-2	-3
C	35	34	-29	-23	1	4
D	33	29	-32	-24	-3	1
F	35	40	-46	-35	-12	0

Note.--Standardized (STD) scores are adjusted to the mean and standard deviation of SAT-M. Composite is weighting of SAT-M and High School Grade Point Average that produced the best prediction of pre-calculus grade.

Table 7

Calculus Course Grades at Selected Levels of SAT-M, Standardized HSGPA, and Standardized Composite

		STD Score Based on SAT-M		STD Score Based on HSGPA		STD Score Based on Composite	
STD Score	Gender	n	Calculus Grade	n	Calculus Grade	n	Calculus Grade
700+	M	3322	2.94	4316	2.98	2488	3.15
	F	1085	3.08	4098	3.03	1266	3.25
650-699	M	3790	2.63	3204	2.71	4059	2.73
	F	2123	2.91	2644	2.69	2916	2.89
600-649	M	4468	2.44	3545	2.50	4057	2.48
	F	2918	2.72	2539	2.44	3154	2.67
550-599	M	3792	2.31	3109	2.34	3323	2.24
	F	3163	2.56	2118	2.35	2631	2.40
500-549	M	1856	2.10	2613	2.12	2199	2.06
	F	2059	2.36	1419	2.19	1761	2.22
<500	M	1502	1.95	2269	1.84	2008	1.79
	F	2113	2.09	917	1.95	1464	1.87

Note.--Calculus grade is averaged across courses.

Table 8

Gender Difference in Calculus Course Grades at Each Score Level on SAT-M, Standardized HSGPA, and Standardized Composite Computed Across and Within Calculus Courses

	Mean Calculus Grade for Males minus Mean for Females					
STD	SAT-M		STD HSGPA		STD Composite	
Score	Across	Within	Across	Within	Across	Within
700+	-.14	-.16	-.05	-.05	-.10	-.09
650-699	-.28	-.26	.02	.05	-.16	-.13
600-649	-.28	-.26	.06	.09	-.19	-.13
550-599	-.25	-.22	-.01	.03	-.16	-.09
500-549	-.26	-.21	-.07	-.02	-.16	-.09
<500	-.14	-.18	-.11	-.04	-.08	-.05

Note.--Standardized (STD) scores are adjusted to the mean and standard deviation of SAT-M. Composite is weighting of SAT-M and High School Grade Point Average that produced the best prediction of calculus grade.

Table 9

Pre-Calculus Course Grades at Selected Levels of SAT-M, Standardized HSGPA, and Standardized Composite

		STD Score Based on SAT-M		STD Score Based on HSGPA		STD Score Based on Composite	
STD Score	Gender	n	Pre-Calculus Grade	n	Pre-Calculus Grade	n	Pre-Calculus Grade
600+	M	581	2.60	444	2.77	547	2.86
	F	368	2.86	890	2.76	702	2.94
550-599	M	911	2.36	725	2.52	788	2.40
	F	829	2.71	999	2.53	934	2.61
500-549	M	649	2.08	785	2.25	893	2.21
	F	1033	2.44	925	2.23	1028	2.30
450-499	M	490	1.98	676	2.03	612	1.93
	F	674	2.28	563	1.99	592	2.08
<450	M	649	1.80	995	1.70	760	1.54
	F	1132	1.88	584	1.85	755	1.66

Note.--Pre-calculus grade is averaged across courses.

Table 10

Gender Difference in Pre-Calculus Course Grades at Each Score Level on SAT-M, Standardized HSGPA, and Standardized Composite Computed Across and Within Pre-Calculus Courses

	Mean Pre-Calculus Grade for Males minus Mean for Females					
STD	SAT-M		STD HSGPA		STD Composite	
Score	Across	Within	Across	Within	Across	Within
600+	-.26	-.29	.01	-.03	-.08	-.14
550-599	-.35	-.35	-.01	-.03	-.21	-.22
500-549	-.36	-.37	.02	.03	-.09	-.09
450-499	-.30	-.29	.04	.06	-.15	-.11
<450	-.08	-.08	-.15	-.11	-.12	-.05

Note.--Standardized (STD) scores are adjusted to the mean and standard deviation of SAT-M. Composite is weighting of SAT-M and High School Grade Point Average that produced the best prediction of pre-calculus grade.