Supporting Online Material for

Standardized Tests Predict Graduate Students’ Success

Nathan R. Kuncel* and Sarah A. Hezlett

*To whom correspondence should be addressed. E-mail: kunce001@umn.edu

Published 23 February, Science 315, 1080 (2007)
DOI: 10.1126/science.1136618

This PDF file includes:

SOM Text
Table S1
References
Table 1

Correlations and odds ratios between 7 standardized tests and 8 measures of student success in graduate school

GRE-T = Graduate Record Examination Total, GRE-S = Graduate Record Examination Subject, LSAT = Law School Admission Test, GMAT = Graduate Management Admission Test, MAT = Miller Analogies Test, MCAT = Medical College Admission Test, PCAT = Pharmacy College Admission Test, rc = corrected correlation, r_obs = observed correlation, N = total sample size, k = number of samples in meta-analysis, OR = odds ratio, -- = not relevant or no data.

<table>
<thead>
<tr>
<th>Test</th>
<th>1st Year GPA</th>
<th>Graduate GPA</th>
<th>Qualifying Exams</th>
<th>Degree Completion</th>
<th>Research Products</th>
<th>Publication Citations</th>
<th>Faculty Ratings</th>
<th>Licensing Exams</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRE-T</td>
<td>rc (r_obs)</td>
<td>.41 (.27)</td>
<td>.37 (.25)</td>
<td>.40 (.30)</td>
<td>.22 (.16)</td>
<td>.11 (.08)</td>
<td>.23 (.17)</td>
<td>.50 (.27)</td>
</tr>
<tr>
<td></td>
<td>N (k)</td>
<td>45,617 (1,231)</td>
<td>14,291 (103)</td>
<td>1,196 (11)</td>
<td>6,304 (32)</td>
<td>3,328 (18)</td>
<td>2,306 (12)</td>
<td>4,939 (34)</td>
</tr>
<tr>
<td></td>
<td>OR</td>
<td>5.2</td>
<td>4.2</td>
<td>4.9</td>
<td>2.3</td>
<td>1.5</td>
<td>2.4</td>
<td>8.1</td>
</tr>
<tr>
<td></td>
<td>N (k)</td>
<td>10,225 (98)</td>
<td>2,413 (22)</td>
<td>534 (4)</td>
<td>2,575 (11)</td>
<td>3,058 (18)</td>
<td>2,306 (12)</td>
<td>879 (12)</td>
</tr>
<tr>
<td></td>
<td>OR</td>
<td>6.2</td>
<td>5.1</td>
<td>8.6</td>
<td>4.6</td>
<td>2.2</td>
<td>2.5</td>
<td>8.1</td>
</tr>
<tr>
<td>LSAT</td>
<td>rc (r_obs)</td>
<td>.53 (.38)</td>
<td>-- (.40)</td>
<td>-- (.09)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>-- (.30)</td>
</tr>
<tr>
<td></td>
<td>N (k)</td>
<td>259,640 (154)</td>
<td>22,218 (142)</td>
<td>24,938 (142)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OR</td>
<td>9.7</td>
<td>4.9</td>
<td>1.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GMAT</td>
<td>rc (r_obs)</td>
<td>.41 (.29)</td>
<td>.35 (.25)</td>
<td>--</td>
<td>.13 (.10)</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>N (k)</td>
<td>48,837 (315)</td>
<td>5,538 (28)</td>
<td>680 (4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OR</td>
<td>5.2</td>
<td>3.9</td>
<td>1.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAT</td>
<td>rc (r_obs)</td>
<td>.41 (.29)</td>
<td>.39 (.27)</td>
<td>.58 (.47)</td>
<td>.21 (.15)</td>
<td>.19 (.13)</td>
<td>--</td>
<td>.37 (.25)</td>
</tr>
<tr>
<td></td>
<td>N (k)</td>
<td>2,999 (34)</td>
<td>11,368 (70)</td>
<td>987 (10)</td>
<td>3,963 (20)</td>
<td>314 (4)</td>
<td></td>
<td>1,909 (25)</td>
</tr>
<tr>
<td></td>
<td>OR</td>
<td>5.1</td>
<td>4.6</td>
<td>13.2</td>
<td>2.2</td>
<td>2.0</td>
<td></td>
<td>4.2</td>
</tr>
<tr>
<td>MCAT</td>
<td>rc (r_obs)</td>
<td>.59 (.44)</td>
<td>.46 (.32)</td>
<td>--</td>
<td>-- (.14)</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>N (k)</td>
<td>4,076 (28)</td>
<td>4,706 (28)</td>
<td>28,050 (126)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OR</td>
<td>14.2</td>
<td>6.5</td>
<td>1.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCAT</td>
<td>rc (r_obs)</td>
<td>.50 (.45)</td>
<td>.41 (.38)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>N (k)</td>
<td>2,829 (22)</td>
<td>1,132 (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OR</td>
<td>8.1</td>
<td>5.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supporting Online Materials
Kuncel & Hezlett (2007)

Only summary results were reported in this review. The results that appear in Figure 1 are presented in Table S1 with additional detail. Each study summarized in Table 1 also contains considerably more information such as additional validity estimates for other criteria (e.g., specific course grades) and for more specific disciplines. These studies can be consulted for specific details but the patterns reported here for grades and other outcomes still hold even when sub-analyses were conducted. It is important to note that many of the outcome measures occur years after the student was tested. That is, standardized tests are predicting important human behaviors years after they are taken. In addition, all of these outcomes, even grades, are based on hours of student work. Performance on classroom examinations and licensing examinations are the end product of a great deal of prior behavior. In the following sections we discuss specifics of the graduate school success measures and describe the sources and methods used to produce the overall summary results presented in Figure 1 and Table S1. We then expand on the article by reviewing additional corroborating research on the predictive validity of tests, present a deeper treatment of the utility of standardized tests, and finally discuss additional information on what is termed predictor bias or differential prediction.

Details on the Measures of Graduate School Success

The predictive validity of tests for first year and graduate GPA is not identical across all tests but are consistently positive and larger than other predictors including undergraduate record. However, it cannot be concluded on the basis of these data that one test is better than another due to their use in different disciplines. Fields vary in the variability of grades they assign as well as the content and structure of the courses. In fact, given the differences in grading standards, content, and instruction across the fields, the consistency of the results is striking. Only the GRE-Subject and GRE General exams can be directly compared. The subject tests are used in some of the same fields as the general tests and, in head to head comparisons; the Subject test is consistently a better predictor with the combination of the 3 yielding even better results (1).

The degree completion criterion differed somewhat across studies. In all cases, it was a contrast between students who completed versus those who did not, at a given time point. However, some results excluded students who did not complete on time. Overall, the association between test scores and degree attainment is strongest when the contrast is between students who leave for academic reasons versus those who do not.

Some studies reported correlations for test relationships with qualifying examinations while other reported correlations for what were labeled comprehensive examinations. Given the ambiguity in the literature, these results were combined to examine the relationship between standardized tests and field specific examinations of knowledge that are required for progress in a degree program.

Results for research products were based on studies that examined some combination of publications, conference papers, and conference proceedings. Publication citations were only for published research.

Data were not available for all outcome measures for all tests. In some cases the outcome measure does not apply for most or all of the fields which use the admissions test. For example,
law school students do not take comprehensive examinations and most academic fields do not have licensing examinations. In other cases, no multi-sample studies have been done to our knowledge. We note a few isolated primary studies that help provide some information about these outcomes.

Methods for Obtaining Correlations

Information for the correlations between Graduate Record Examination (GRE-Total) and the Graduate Record Examination Subject (GRE Subject) tests and all outcome measures was estimated from a meta-analysis (1). The predictive power of total scores was calculated using three pieces of information: 1. the correlation between the verbal and quantitative exams, 2. the meta-analytically estimated correlation between GRE-V and the outcome, and 3. the meta-analytically estimated correlation between the GRE-Q and the outcome. This information was used to compute a unit-weighted composite which is described in more detail later in the supporting materials. It should be noted that the correlation between the GRE Verbal and qualifying/comprehensive examinations is substantially larger than the correlation for GRE Quantitative likely due to the highly verbal nature of these written examinations in many fields. Combining the GRE-Verbal and GRE-Quantitative exams into a unit-weighted (not optimally weighted) total score reduces predictive power over what is possible with the GRE-Verbal alone.

Data for the correlation between the Law School Admission Test (LSAT) and 1st year grades with and without corrections was available from the meta-analysis we used as the source of data for our figure and table (2). Data for the correlations between the LSAT and cumulative law school GPA at 142 law schools was presented in another study (3). These results are ultimately based on 142 independent samples. These data had been analyzed in 6 clusters of law schools that share similar characteristics (3). We aggregated these 6 correlations into a meta-analysis by weighting each correlation by sample size. The same study also contained means and standard deviations for students who graduated as well as those who failed to graduate from law school for each of the 6 law school clusters. These data were converted into point-biserial correlations to make the results comparable to the other data presented this paper. These correlations were then aggregated, weighting by sample size, into a single overall value. The correlation between the LSAT and passing the bar exam on the first try were presented in a different study of 163 law schools (4). It is important to note that this study was very representative of the population of law schools. It includes the data from 95% of U.S. mainland law schools which are accredited by the American Bar Association. Range restriction corrections for GGPA were not conducted as the necessary information was not available.

The correlations between the Graduate Management Admission Test (GMAT Total) score and all outcome measures were estimated from a meta-analysis (5). Correlations between the Total scores (Verbal + Quantitative) and outcomes were calculated using the correlation between the Verbal and Quantitative exams and each of their associated correlations with the outcome or success measures.

The correlations between the Miller Analogies Test (MAT) and all subsequent outcome measures were obtained from a meta-analysis (6). The MAT does not have multiple scales; results are reported with a single score. Therefore, no compositing of scale validities was necessary. The results for the MAT are based on studies that are, on average, older than the other data presented here. Given changes in grading and curricula, the correlations may differ today. However, research on other tests has not found any substantial effects for time on the
relationship between test scores and grades in graduate school (1). It should be noted that different estimates of the reliability of faculty ratings were used in the research on the MAT and the GRE with the former being more conservative than the later. As can be seen in the table, the uncorrected correlations are similar.

Finally, all outcome correlations with the PCAT were obtained from a single meta-analysis (7). The PCAT is composed of multiple scales and estimates of the validity of a total score was not always available. As with the GRE-T and GMAT, we estimated the validity of the total score composite using the formula described in a later section. The intercorrelations among the PCAT scales were obtained from its technical manual (8). It should be noted that the licensing examination examined in this study is the previous version of the current licensing examination for pharmacy. The two licensing examinations share a lot in common but no estimates of predictive power were available for the updated licensing examination. Unlike the other studies that examined cumulative graduate grade point average, the data for graduate grade point average for the PCAT are for 3rd year, non-cumulative grades. Data are available in the original study for both 2nd and 3rd year non-cumulative grade point averages. Correlations are very similar for each year. Results for all PCAT correlations with outcomes were not corrected for restriction of range because the information was not available in sufficient quantity to conduct these corrections. The results are, therefore, quite likely to be conservative underestimates as pharmacy programs select students using the PCAT.

Results from the Medical College Admission Test (MCAT) came from multiple sources. For graduation versus dropout the result presented in our table is based on data from the 1982-1983 cohorts across 126 medical schools (9). Data were only presented in a format that could be converted into a point-biserial correlation for the Reading and Chemistry scales. However, a figure in the study indicates that the other scales were also predictive such that students with scores below 4 on the MCAT had academic troubles 45 to 50 percent of the time in contrast to those with above average scores having difficulty less than 5 percent of the time (9). Therefore, the estimate of the overall predictive power of the MCAT for the graduation outcome that we present is an underestimate. Additionally, these estimates were not corrected for range restriction, further reducing the observed correlation. We aggregated the data we had using the unit-weighted composite described below. For grade outcomes and medical boards, a separate study of cohorts from 14 medical schools provided the data (10).

Corrections for Unreliability and Restriction of Range

Corrections for statistical artifacts are done to reduce the bias in the observed correlation between predictors and outcomes, and so the corrected estimates are the focus of this paper. Nevertheless, it is critical to note that even the uncorrected estimates are large enough to have practical utility for admissions. That is, the overall conclusions presented in the article are not dependent on statistical artifact corrections.

Estimates of measurement unreliability are based on studies that quantify human rating errors over time or the inconsistency of multiple judges who are rating the same target. These estimates were used to correct the correlation for measurement unreliability in the outcomes except for degree attainment, research productivity, citation counts, and qualifying and comprehensive examinations. To the extent that there is measurement error in these outcomes due to transcriptions errors or other sources, the results we present will, on average, underestimate the relationship. Corrections are not typically used for the predictor as it is used
for decision making. That is, the effects of unreliability in the outcome measure are corrected to better reflect students’ real achievement, free from measurement error. Since the admissions decisions must be made with the predictors “as is” no corrections are made for unreliability in predictors. Correction formulas have been developed to address these issues and have often been applied to meta-analytic findings.

The corrections for unreliability used in our review are based on Classical Test Theory and use estimates of reliability based on correlations among scores. The three major types of reliability estimates used in testing are test retest, alternate form, and internal consistency. Test retest are scores from a test administered at two different points in time, often multiple days or months apart. Standardized ability tests show strong test-retest reliabilities with correlations of .86 to .93 (11). Alternate form reliabilities are correlations between scores on two versions of the same test (i.e., different items). Finally, internal consistency reliabilities are a class of reliability estimates based on the intercorrelation between one half of a test with another. The correlation between the odd and even items is called a split half reliability. The most common estimate internal consistency reliability estimate is the Cronbach’s alpha which is the average of all possible split halves. Each of these estimates has a set of assumptions and limitations (12). It would be preferable to correct each correlation with its own, sample specific, estimate of grade reliability. This is not the case in the research we review as few reliability estimates exist for graduate grades in different fields and discipline. Instead available estimates were used in an artifact distribution method (12). Overall reliability corrections resulted in less than a 10% increase in the correlation coefficients.

In selecting students from a population of applicants, variability in the predictor (e.g., test scores or prior grades) in the selected or admitted sample is reduced. Because the computation of the correlation is affected by the variability of the predictor, this reduction in the predictor’s variability artificially reduces the magnitude of the correlation between the predictor and a measure of performance. This is typically called restriction of range. Since the principle research question of interest is the extent to which using a test improves the performance of the selected group over the applicant group, the attenuated observed correlation is an underestimate of the relationship of interest: the correlation between the test and the performance measure for the applicant group. Most of the estimates we provide have been corrected for restriction of range. When estimates were left uncorrected it was because the necessary information was not available. Ideally these corrections would be done with multivariate range restriction corrections, however, the necessary information is nearly never available. Instead univariate corrections are commonly used and these provide good, if somewhat conservative, corrections (12). It is widely acknowledged that correcting correlations for unreliability in the criterion and for range restriction results in more accurate estimates. The Standards for Educational and Psychological Testing were developed by the National Council on Measurement in Education, the American Psychological Association, and the American Education Research Association (13). These standards endorse the value of corrections.

Converting Correlations to Odds Ratios

Table 1 also contains Odds Ratios for each pairing of test with outcome. This was done to aid interpretation for those who are less familiar with Pearson correlations. The Odds Ratio is the ratio of the odds of one event A/(A+B) to the odds of another event C/(C+D). In the case of tests and outcomes it is the odds of above-average success for above-average test scorers to the
odds of above-average success for below-average scorers. The correlation is more suited to the analysis of continuous data than the Odds Ratio but the Odds Ratio does provide a separate statistic that is likely to be more interpretable to many scholars. The Odds Ratio can be converted into a standardized mean difference ($d$) via the following formula:

$$d = \frac{\sqrt{3}}{\pi} \log OR$$

Where $d$ is the standardized mean different and OR is the odds ratio (14). In turn, the standardized mean difference can be converted into a point-biserial correlation by the following formula:

$$r_{pb} = \frac{d}{\sqrt{4 + d^2}}$$

Where $d$ is the standardized mean difference and $r_{pb}$ is the point biserial correlation (12). Note that this formula assumes, as we are, that groups are being split into equal proportions (.5 and .5).

Computing Total Score Correlations

Not all meta-analyses presented the predictive validity of the overall or total test scores. Therefore, some of the estimates presented in the table required combining validity for two or more scales (e.g., Verbal and Quantitative for the GRE). This was accomplished using unit-weighted composites that estimate the correlation between the average of standardized scores and the criterion. The correlation between the composite of the two or more predictors will differ from the average correlation in all cases except when the predictors are perfectly correlated with each other. For example, the composite for GRE General tests (called GRE-T in our figure and table) is the correlation between the criterion (e.g., grades, faculty ratings) and the average of scores on the Verbal and Quantitative sections once standardized. This was estimated using the correlation between each predictor and the criterion as well as the correlations among all of the predictors. The intercorrelations used in these estimates were either obtained from the meta-analyses or from the official technical manual for the test. The correlation for the total score provides a summary estimate of the validity that would be obtained if the two scales are used together to make admission decision and given equal weight. We chose to present unit-weighted composites as they more accurately reflect how sub scales are actually used. We are aware of few departments that make use of empirically developed weighting schemes when combining predictors. The unit-weighted composite estimate is nearly always lower than what would be obtained using optimal weights. The formula for the correlation between a unit-weighted composite and another variable is:

$$r_{co} = \frac{r_{po}}{\sqrt{1 + \frac{k-1}{k} r_{pp}}}$$
Where \( r_{co} \) is the correlation between the composite and the outcome, \( r_{po} \) are the correlations between the predictors and the outcome (which are then averaged), \( k \) is the number of predictors, and \( r_{pp} \) are the correlations among the predictors (also averaged) (15).

Other Validity Evidence

A number of other large scale (independent samples from multiple institutions) quantitative analyses have examined the predictive validity of standardized tests in graduate admissions. We focused on both the largest and most recent studies that presented information on criteria other than grades. However, the other review studies in the literature are consistent with the results we present in the review and should be discussed.

For the Graduate Record Examination a number of other large scale examinations have been conducted. Some of these are based on a small number of studies or are for a single outcome measure (16-19). A few reviews have examined the predictive power of the GRE for specific disciplines (18, 19). Studies from all but one of these are included in the meta-analytic results presented in Table S1 and Figure 1 (16-18). In all cases the relationship between GRE scores and outcome measures is positive. Finally, the GRE, unlike the MCAT, PCAT, and LSAT is not typically used for professions with licensing examinations. One exception is in the state of Texas within the United States where school principals take a state licensing test. Based on currently a small amount of evidence the positive pattern remains with a observed correlation of .69 (N=53) in one study between the GRE and performance on the licensing examination (20). A second study reported positive relationships between the GRE and licensing examination scores that reached significance for some samples (21).

For the GMAT two large scale meta-analysis have also been conducted on validation data from the Graduate Management Admission Council’s internal validation service (22, 23). These findings parallel the results we report in Figure 1 and Table S1. The results we report are based on a meta-analysis of the publicly available literature. On other criteria, a study on degree attainment in business school doctoral programs has also been conducted (24). Although this study was conducted on 25 schools, no relationship between the GMAT and degree attainment was observed in contrast to the other studies reviewed here. Other research investigating doctoral business schools has reported that although the GMAT is a valid predictor of grades earned, undergraduate record is an even better predictor with both GMAT and UGPA providing the best information (25). Finding that undergraduate grades are the best predictor and that the GMAT did not predict degree attainment is both inconsistent with the overall pattern across fields as well as the data we present for MBA programs. Doctoral programs in business schools appear to differ from other fields as well as MBA programs. Whether these inconsistent findings are in fact true of business school Ph.D. programs should be further investigated. Unfortunately, doctoral students are comparatively rare in business schools compared to many other fields.

The Law School Admission’s Council has done a series of reports on the predictive validity of the LSAT for grade outcomes from several different time periods (26-28). In addition, other externally conducted meta-analysis have been conducted on the relationship between the LSAT and grade point average in law school with nearly identical findings to those reported in Table 1 (29).

We are aware of no other large scale evaluations of the PCAT or the MAT other than the validation data collected in their technical manuals (8, 30). It should be noted that this
information was incorporated into the meta-analyses we review and synthesize for these tests. Thus, the technical manuals do not provide a unique source of information.

The predictive validity of the MCAT has also been examined in a separate summary study (31). The MCAT has been modified in the past. Although the previous versions were valid predictors, comparisons of results suggest that the newer version is a somewhat better predictor. We are aware of only a few older studies that have examined the relationship between MCAT scores and graduation versus dropout for academic reasons. Other studies corroborate the positive relationships between MCAT scores and outcomes including degree attainment that we present in Figure 1 and Table S1 (32-36). These data are based on the earlier and, arguably, slightly less effective MCAT. Based on prior research and the overall findings in testing, the new MCAT is likely to have positive relationships with degree attainment. It is important to note that few low scoring students are admitted into medical schools making the base rate of drop out very low. This in turn reduced the magnitude of the correlation.

Predictive Validity of Tests for Non-Native English Speakers

The predictive power of standardized tests has been examined meta-analytically for non-native English speaking students for the GRE and the GMAT (1, 5). In both cases, test scores were positively correlated with grade point averages earned in graduate school. Additionally, both meta-analyses report stronger correlations for the quantitative scales than the verbal scales. The exact cause of this difference is not known. One possible explanation is the tendency for these students to enroll in quantitative courses and fields (37). Another possibility is that the Verbal scales results in a less accurate assessment of skills due to language differences. Research has provided partial support for a moderating influence of basic English proficiency on the predictive power of the GRE (38).

Utility of Standardized Tests

Even tests with small correlations with outcomes can have utility for admission purposes, especially when the program can be selective. Correlation coefficients can be used to estimate improvements in correct versus incorrect decisions. These estimates are based on the base rate of success that would obtained from the applicant group, the selection ratio (the proportion of the applicant group that is selected or admitted), and the correlation. These estimates can make the benefit of a particular predictor far more apparent than a correlation. For example, assuming that among applicants to a highly selective school (10 percent admitted) 50 percent would be rated as satisfactory or better by faculty, the use of the GRE General Exam alone would increase the base rate of success from 50 percent to 84 percent. The utility of tests holds true even under unfavorable conditions when the base rate of success is high or low (attempting to predict rare events), schools are less selective, and correlations with success are lower. For example, first-time bar exam passage rates are just under 89 percent nationally (4). Failure is a relatively rare event. Although LSAT scores are restricted by selection, if we were to accept the observed correlation of .30 without corrections, the LSAT alone will, on average, increase the bar pass rate at a selective school (20 percent admitted) by 6 percent to a 95 percent passing the first time. Consistent with this estimate, the top group of law schools have an average pass rate of 94 percent and, on average, admit about 1 in 5 applicants based on a combination of LSAT scores and other predictors that are less predictive of bar passage than the LSAT (4). The conversion of
a correlation to correct and incorrect decisions requires the assumptions that the relationship between the predictor and the outcome is linear and that the relationship between the predictor and outcome are bivariate normal (15).

A final concern about the utility of standardized tests in admission is the extent to which scores do not matter at all. It is certainly true that as the range of scores gets more and more restricted, other characteristics of the individual that are orthogonal to test scores will increasingly matter. However, it does not appear to be the case that standardized tests have an asymptotic relationship with subsequent performance or success either in education or at work. Research on the extreme ends of talent as measured by standardized tests demonstrates relationships with success even among highly select groups. For example, twenty years after taking a standardized test, outcome differences are still observed among children who scored in the top 1% (39). Additionally, research in the work domain found that the relationship between scores and performance does not asymptote for above average scores (40).

Predictor Bias and Differential Item Functioning

The apparent phenomena of under prediction of the performance of women in college but not graduate school warrants some additional discussion. Some, if not all, of under prediction can be attributed to differences between men and women in young adulthood in terms of their responsibility and study behaviors as well as the influence of genders differentially enrolling in majors and courses that vary in their grading severity (41, 42). Therefore, once these other factors are considered the relationship between test and subsequent performance becomes nearly identical for the two genders. Additional studies have documented the general finding of grade selection effects (i.e., the tendencies for groups to differentially enroll in courses across disciplines) on under prediction of grades for women and minorities (43).

This highlights a broader set of research questions surrounding omitted variables. The slope and intercept of the regression lines is influenced by multiple predictors as well as the nature of the outcome variable. Observed similarities or differences in regression equations can change when other factors are considered. Other predictors of student performance include interests and personality traits. Differences between ethnic and racial groups on these variables are few and small suggesting that they would not change conclusions about test scores (44). Nonetheless, other unknown variables could potentially have an influence on these conclusions.

An issue related to research on predictor bias is the phenomena of stereotype threat. In laboratory experiments this effect has been shown to further decrease the performance of some minority students on standardized tests questions. Often this effect has been misinterpreted as accounting for all of the commonly observed difference in scores obtained by majority and minority groups (45). Subsequent research has examined aspects of this effect in operational settings and the effect has failed to replicate under actual testing conditions with very large samples (46). Other work, again with very large samples has created high fidelity testing situations and has also failed to replicate (47). Additional studies have examined predictions made from stereotype threat in simulated work settings (48). Differences in levels of threat did not yield changes in test scores.

Finally, some research has attempted to investigate applied effects of threat in work and academic settings. Threat is said to reduce performance on tests below what the person is actually capable of doing. On the basis of this, one can predict that tests scores would have differential relationships with subsequent performance at school or work for groups that have
experienced or not experienced threat. This question was examined outside of the laboratory in actual academic settings using the SAT and large U.S. Army samples using the ASVAB (49). If threat produces lower performance on a test, then subsequent performance in a non-threat work or educational setting should be higher than would be expected from the threat attenuated test scores. That is, the relationship between the test and an outcome measure should be different for those experiencing stereotype threat than for those (of a majority group) who do not experience stereotype threat. Instead, they found that performance in a non-threat work or educational settings is consistent with test scores across gender and racial groups. For example, operationally obtained SAT Math and Verbal scores for women (a threat situation) were not differentially related for women to performance in English classes (a no threat performance domain as no stereotype exists that women are worse at English composition).

In research conducted on Differential Item Functioning (DIF) on college and graduate admissions tests most items do not display DIF but some patterns have emerged over time. Generally, men tend to do slightly better on items with content focused on either sciences or activities that are stereotypically male while women tend to do slightly better on items with content oriented toward interpersonal relationships and the arts (50). Items based on English homographs, words with two distinct meanings, tend to favor European American test takers over those who self identify as Hispanic (51). In contrast, items that contain correct responses with words that are frequently used true Spanish cognates tend to favor Hispanic test takers (52). Items are identified in applied settings as demonstrating DIF based on null hypothesis tests often in very large samples. This results in a number of items demonstrating minute indications of DIF for both majority and minority groups. Again the cumulative effect of these differences on actual decision making is effectively zero (53).
References and Notes


45. P. R. Sackett, C. M. Hardison, M. J. Cullen, Amer. Psych. 59, 7.
47. P. Loveless, “Does self-stereotyping affect performance on the ACT assessment?”
48. for a review of these 4 studies see Sackett, P. R. Human Perf. 16, 295.
50. K. A. O’Neill, W. M. McPeek, in Differential Item Functioning, P. W. Holland, W.
51. A. P. Schmitt, N. J. Dorans, “Differential item functioning for minority examinees on the
52. A. P. Schmitt, in Differential item functioning on the Scholastic Aptitude Test, A. P.
   Schmitt, N. J. Dorans, Eds. (ETS Research Memorandum No. 87-1, Educational Testing