Evaluating the Predictive Validity of Graduate Management Admission Test Scores

Stephen G. Sireci and Eileen Talento-Miller

Educational and Psychological Measurement 2006; 66; 305
DOI: 10.1177/0013164405282455

Validity Studies

Educational and Psychological Measurement
Volume 66 Number 2
April 2006 305-317

10.1177/0013164405282455
http://epm.sagepub.com
hosted at
http://online.sagepub.com

Evaluating the Predictive Validity of Graduate Management Admission Test Scores

Stephen G. Sireci
University of Massachusetts Amherst
Eileen Talento-Miller
Graduate Management Admission Council, McLean, Virginia

Admissions data and first-year grade point average (GPA) data from 11 graduate management schools were analyzed to evaluate the predictive validity of Graduate Management Admission Test® (GMAT®) scores and the extent to which predictive validity held across sex and race/ethnicity. The results indicated GMAT verbal and quantitative scores had substantial predictive validity, accounting for about 16% of the variance in graduate GPA beyond that predicted by undergraduate GPA. When these scores and undergraduate GPA were used together, they accounted for approximately 25% of the variation in first-year graduate GPA. Correcting correlations for restriction of range improved the predictive power. No statistical differences were found across examinee groups defined by race/ethnicity and sex, which suggests a lack of bias in these scores. The predictive utility of GMAT analytical writing scores was relatively low, accounting for only about 1% of the variation in graduate GPA, after accounting for undergraduate GPA and GMAT verbal and quantitative scores.

Keywords: *admissions testing; differential predictive validity; predictive validity; validity*

According to *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999), validity "refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). A primary use of the Graduate Management Admission Test® (GMAT®) is to assist admissions officers of graduate management schools in making admissions decisions. Therefore, the validity of GMAT scores depends in large part on the degree to which they predict students' performance in graduate school.

Since the GMAT first came into use in 1954, researchers have been assessing its utility in predicting grades. A study by Olsen (1957) looked at the predictive ability of the GMAT exam, then known as the Admission Test for Graduate Study in Business, and undergraduate grades in relationship to first-year grades at 10 business schools. She found that the use of GMAT scores in admissions improved predictions over using undergraduate grades alone. Over the years, the results of hundreds of validity studies conducted by the Graduate Management Admission Council® (GMAC®) have been summarized showing that combining GMAT scores and undergraduate grades results in better prediction of performance in business school than any of the predictors alone (GMAC, 2003; Olsen, 1957; Powers & Moss, 1980; Wightman & Leary, 1985; Zhao et al., 2000).

In this study, we retrospectively compare contemporary management students' performance in graduate school with their GMAT scores. First, we investigate the degree to which GMAT scores and undergraduate grade point average (UGPA) predict graduate management school grades. Next, we look at the degree to which the predictive power of GMAT scores holds up over specific subgroups of examinees defined by sex or race/ethnicity. By looking at both overall results and results for specific subgroups, a broad picture of the fairness and utility of GMAT scores is provided.

Method

Student Data

Schools that had the highest numbers of minority student applicants to their graduate management programs were invited to participate in this study. Eleven of these schools participated and sent data on students' sex, race/ethnicity, GMAT scores, UGPA, and first-year graduate school GPA. The GMAT scores used were GMAT Quantitative (GMAT-Q), GMAT Verbal (GMAT-V), and GMAT Analytical Writing (GMAT-AW). Across the 11 schools, data were sent on 5,076 students. Data on biological sex were available for all students, and about 71% were male. Data on students' race/ethnicity were available for only 3,175 (62.5%) students. Of the students who had race/ethnicity data, the largest racial/ethnic group was White (72.4%), followed by Asian American (16.2%), African American (6.2%), Hispanic/Latino(a) (4.9%), and American Indian (0.3%). The sample sizes for these student groups are presented in Table 1, stratified by school.

As can be seen in Table 1, the numbers of American Indian students were too few to analyze their data as a separate group and achieve reliable results. Thus, these students were included in the overall predictive validity analyses but not in the analyses that looked at consistency of the prediction across subgroups. It can also be seen that there were relatively small numbers of non-White students within each school. These small sample sizes limited the types of analyses that could be conducted, as we discuss in a subsequent section.

Table 1
Within-School Frequencies for Selected Examinee Subgroups

		Sample Sizes for Selected Groups						
School	Total *N*	Men	Women	White	Asian American	African American	Hispanic/ Latino(a)	American Indian
1	215	151	64	92	33	1	6	0
2	476	302	174	205	57	27	13	1
3	468	365	103	197	53	10	8	0
4	636	456	180	272	99	17	23	1
5	171	135	36	91	7	5	4	0
6	272	192	80	104	54	14	10	0
7	430	310	120	195	31	33	17	1
8	782	535	247	265	70	31	24	1
9	233	165	68	121	19	8	7	1
10	750	581	169	430	45	7	23	3
11	643	415	228	326	45	45	20	3
Total	5,076	3,607	1,469	2,298	513	198	155	11

Data Analyses

Multiple regression analyses were used to evaluate the degree to which admission variables predicted students' graduate GPA. In the context of this study, we looked at four predictors: GMAT-Q, GMAT-V, GMAT-AW, and UGPA, the measures commonly used to evaluate applicants for admission to graduate schools in the United States. Three of the 11 participating schools did not provide data on students GMAT-AW scores, and so we looked at the predictive relationship with and without this variable.

In predictive validity studies, it is desirable to find large and statistically significant multiple correlation coefficients, which account for a substantial amount of variation in the predictor. Previous studies on the predictive validity of admissions test scores have found average multiple correlations between predictor variables and GPA in the range of .40 to .50. For example, Linn and Hastings (1984) found a multiple correlation of .46 for Law School Admission Test (LSAT) scores combined with undergraduate grades in predicting first-year law school grades. Summaries of recent studies conducted by GMAC on the validity of the GMAT show a median multiple correlation of .45 for the combination of GMAT scores and UGPA (GMAC, 2003; Zhao et al., 2000).

In addition to gauging the strength of the predictors considered simultaneously, multiple regression can also be used to evaluate the utility of each predictor. The regression slopes provide one indicator of the relationship between a predictor and the criterion. If the regression coefficient for a particular variable were statistically significant, it can be concluded the variable is important for accounting for variation in the criterion. However, it is difficult to isolate the relative contributions of each predictor from the regression coefficients. The squared *semipartial* correlation ($r^2_{Y(i.jkl)}$) is better suited for this purpose. This statistic represents the proportion of variance in the criterion that is not accounted for by the other predictor variables. For example, if the squared semipartial correlation for GMAT-Q were .10, it would mean that GMAT-Q accounts for 10% of the variation in students' GPA above and beyond the variation accounted for by GMAT-V, GMAT-AW, and UGPA. In this study, we report the squared multiple correlations for each analysis, as well as the zero-order correlations, standardized regression weights, and semipartial correlations for each predictor variable.

Using multiple regression to assess differential predictive validity. Multiple regression can also be used to evaluate the consistency of the predictive relationship across different groups of examinees. If the utility of a set of predictors does not hold up over particular groups of examinees (e.g., men and women), the predictors are said to have *differential predictive validity*, which is undesirable because test scores and other prediction criteria should not interact with sex or cultural heritage. Differential predictive validity can be evaluated by fitting separate regression lines for each group and then testing for statistically significant differences between the slopes and intercepts (Linn, 1984; Pedhazur & Schmelkin, 1991; Wainer & Sireci, 2005). Such an approach is preferable when there are sufficient data for each group of interest (minimally, 100 students, preferably more). When sample sizes do not permit fitting separate regression equations, a single equation is fitted to the data for all examinees, and the residuals are analyzed for patterns of *overprediction* (predicted scores are higher than criterion scores, i.e., $Y - \hat{Y} < 0$) or *underprediction* (predicted scores are lower than criterion scores, i.e., $Y - \hat{Y} > 0$; Braun, Ragosta, & Kaplan, 1988; Koenig, Sireci, & Wiley, 1998).

Given that graduate school GPA is not the same across all graduate management schools, and given the relatively small sample sizes for non-White students within each school, it was not possible to derive separate regression equations for each racial/ethnic group within each school. Instead, GPA was standardized *within each school* to have a mean of zero and a standard deviation of one. This standardization assigned the mean GPA of each school a value of zero and assigned rescaled GPAs for each student in terms of her or his deviation from that mean. This standardized GPA (Z-GPA) was then used to derive separate regression equations for each group of interest across all schools. That is, the regression equation for African Americans used the Z-GPAs for all African Americans in any of the 11 schools. The regression coefficients and intercepts from the group-specific regression equations were then tested for statistically significant differences across groups using the *t* test for independent regression coefficients (Cohen & Cohen, 1983, p. 11; Howell, 1997, p. 258):

$$t = \frac{b_1 - b_2}{\sqrt{se_{b1}^2 + se_{b2}^2}}, \tag{1}$$

with degrees of freedom equal to $n - 2$. Statistical significance was evaluated at $p < .05$, but we focused on effect sizes to gauge the magnitude of any differences and to account for the effects of sample size on statistical significance.

Differential predictive validity was also evaluated by fitting a single regression equation across all students within each school using Z-GPA (i.e., differentiating only by school but not by ethnicity or sex within a school) and then looking at the average residuals for each student group across all schools. That is, after fitting a regression equation to the data for all students in each school, the residuals for each group of students defined by ethnicity or sex were pooled across schools. In this way, general patterns of overprediction or underprediction could be discovered.

Results

Predictive Validity

Our first set of analyses focus on the general predictive validity question: To what degree do GMAT scores predict first-year GPA in graduate school? To answer this question, we first calculated the zero-order correlations between the predictors and first-year GPA. Next, separate regression analyses were conducted within each school to predict GPA from GMAT scores and UGPA. First, only UGPA was used to predict graduate GPA. Next, the GMAT scores were entered into the equation. This two-step process allowed us to measure the "value added" of GMAT scores above and beyond UGPA. We define value added as the increase in R^2 due to the addition of the new variables into the prediction equation.

Zero-order correlations. The correlations between the predictors and the criterion (i.e., first-year GPA in graduate school) were computed within each school. In addition to computing these observed correlations (i.e., zero-order correlations), we also present the correlations corrected for restriction of range. Because these data are based on samples of students already admitted into graduate school, they underestimate the true correlations that would be obtained if all students for whom GMAT data were available were admitted into the school. Therefore, we also present the "corrected" correlations, which in this case are corrected for restriction of range on the GMAT scores (population data on undergraduate GPA were not available to correct for restriction of range on that predictor). To compute these corrected correlations, we used Equation (6.8.5) from Lord and Novick (1968, p. 143), which uses the ratio of the observed variance of the predictor in the sample to the observed variance of the predictor in the population. The equation is

$$\rho = \sqrt{\frac{1}{1 + \frac{s_x^2}{\sigma_x^2}\left(\frac{1}{r_{yx}^2} - 1\right)}}, \tag{2}$$

Table 2
Zero-Order Predictor/Criterion Correlations

	Observed				Corrected[a]		
Predictor	Min., Max.	Median	Weighted *M*	R^2	Min., Max.	Weighted *M*	R^2
UGPA	.18, .36	.285	.299	.089			
GMAT-Q	.08, .53	.309	.305	.093	.19, .80	.603	.363
GMAT-V	.19, .35	.303	.273	.075	.38, .66	.524	.275
GMAT-AW[b]	.03, .27	.177	.145	.021	.04, .33	.194	.038

Note: UGPA = undergraduate grade point average; GMAT-Q, GMAT-V, and GMAT-AW = Graduate Management Admission Test Quantitative, Verbal, and Analytical Writing, respectively. R^2 was calculated by squaring the weighted mean multiple correlation (R).
a. Corrected for restriction of range on the predictor. Data were not available to correct for UGPA.
b. Based on only eight schools with AW data.

where s_x^2 is the observed variance in the sample and σ_x^2 is the observed variance in the population. The population data were taken from GMAC (2002).

A summary of these correlations is presented in Table 2. The minimum and maximum uncorrected and corrected correlations are presented. The median and weighted mean correlations across the schools are presented for the uncorrected correlations, and the weighted mean is presented for the corrected correlations. The weighted means were calculated by weighting the correlations by the sample size within each school. The weighted mean correlations were squared to provide an estimate of the average percentage of variance in graduate GPA accounted for by the predictors.

The correlations between the predictors were generally small (average correlations were near 0 between GMAT-Q and GMAT-V and around .10 between the GMAT scores and UGPA). One might expect multiple measures of cognitive ability, such as these, to be at least moderately correlated; however, the graduate school selection process yields unusual correlations between these factors. Because the factors are considered in conjunction, a high score in one area may end up compensating for a low score in another area. For instance, according to GMAC staff, among all test takers in 2003, the correlation between the verbal and quantitative subtest scores was approximately .34 (personal communication, January 26, 2005); however, among individuals who matriculated into the 11 graduate management programs studied here, the correlations were at best very small positive values or at worst moderate *negative* correlations. Because the admission process yields a group whose high scores in one area can compensate for low scores in another, these correlations are likely not unusual based on the selectivity of the school. Relationships such as these argue well for the use of the subtest scores separately rather than a composite measure such as GMAT total score to predict since they emphasize real variation between candidates.

Before correction, UGPA, GMAT-Q, and GMAT-V each accounted for about 8% to 9% of the variation in first-year graduate GPA. GMAT-AW accounted for only about 2%. After correcting GMAT scores for restriction of range, the percentage of variation accounted for more than tripled for GMAT-Q (36.3%) and GMAT-V (27.5%). The percentage of variance accounted for by GMAT-AW nearly doubled but was still quite small (3.8%). These findings suggest that UGPA, GMAT-Q, and GMAT-V are useful predictors of graduate management school GPA.

Multiple regression results. Although the correlations between predictor variables and the criterion are interesting, they do not provide the full picture of how well the *combination* of predictors predicts success in graduate management school. To evaluate the utility of the combination of predictors, subsets of the predictor variables were entered into regression equations. First, only UGPA was entered. Next, GMAT scores were entered. This approach was taken to evaluate the "value added" of GMAT scores above and beyond UGPA. Because only 8 schools provided GMAT-AW data, we first conducted the analyses for all 11 schools using only GMAT-Q, GMAT-V, and UGPA. We then reran the analyses for the subset of 8 schools using all four variables. In the second set of analyses, we also estimated the value added for GMAT-AW, above and beyond UGPA and the other GMAT scores.

A summary of these regression results is presented in Table 3. This summary includes the minimum and maximum values of R and R^2, the median R, the weighted mean R, and an estimate of the average R^2, which was calculated by squaring the weighted mean R.

The results in Table 3 highlight several findings. The most obvious finding is that GMAT-Q and GMAT-V significantly increase the predictability of GPA above and beyond that predicted from UGPA alone. In fact, the amount of variance in graduate management school GPA accounted for by UGPA, GMAT-Q, and GMAT-V is almost triple the amount accounted for by UGPA alone. A second finding is that, when all three predictors are used, there is substantial prediction within every school (i.e., minimum R is .31 for the 11 schools). A third finding is that, for the 8-school subset, the value added for GMAT-AW, above and beyond GMAT-Q, GMAT-V, and UGPA, is relatively small (accounting for about 1% unique variance on average).

In interpreting these results, it should be kept in mind that the R and R^2 results are not corrected for restriction of range. To estimate what the average multiple correlation would be after accounting for restriction of range, we conducted a regression analysis using the weighted restriction-corrected correlations for GMAT-Q and GMAT-V reported in Table 2 (.603 and .524, respectively) along with the weighted correlation for UGPA (.299), and the weighted correlations between the predictor variables using the matrix syntax in SPSS. The resulting multiple R was .849, which is a substantial increase from the observed weighted mean multiple R using these three predictors (.489; see Table 3). These results suggests the predictive power of GMAT-Q, GMAT-V, and UGPA taken together is far greater than observed when looking only at data from students who gained entry into graduate management school.

Differential Predictive Validity

Descriptive statistics for subgroups. Before presenting the regression results with respect to subgroup differences, the mean differences across subgroups on the predic-

Table 3
Summary of Within-School Regression Analyses: R and R^2

	Predictor(s)	Min., Max. $R(r)$	Median $R(r)$	Weighted M $R(r)$	R^2	Value Added
11-school sample	UGPA	.18, .36	.285	.298	.089	
	UGPA, GMAT-Q, GMAT-V	.31, .71	.489	.499	.249	.160
8-school sample	UGPA	.21, .41	.282	.282	.079	
	UGPA, GMAT-Q, GMAT-V	.44, .70	.508	.505	.255	.176
	UGPA, GMAT-Q, GMAT-V, GMAT-AW	.44, .70	.524	.514	.264	.009

Note: UGPA = undergraduate grade point average; GMAT-Q, GMAT-V, and GMAT-AW = Graduate Management Admission Test Quantitative, Verbal, and Analytical Writing, respectively. R^2 was calculated by squaring the weighted mean multiple correlation (R).

Table 4
Mean (Standard Deviation) for Predictor and Criterion Variables

Variable	All	Men	Women	White	Asian American	African American	Hispanic/Latino(a)
GMAT-Q	45.21 (4.15)	45.70 (3.94)	44.03 (4.40)	44.10 (3.91)	45.83 (3.40)	39.87 (4.80)	41.97 (4.12)
GMAT-V	38.99 (4.99)	39.10 (5.01)	38.72 (4.92)	40.72 (4.13)	38.83 (4.67)	36.87 (4.61)	38.16 (4.38)
GMAT-AW	4.78 (0.73)	4.75 (0.74)	4.84 (0.72)	4.98 (0.68)	4.84 (0.69)	4.64 (0.78)	4.80 (0.69)
UGPA	3.39 (0.37)	3.37 (0.38)	3.45 (0.34)	3.41 (0.35)	3.39 (0.34)	3.12 (0.40)	3.29 (0.37)
Z-GPA	0.00 (1.00)	0.09 (0.99)	–0.22 (0.98)	0.15 (0.94)	–0.13 (0.95)	–0.90 (0.90)	–0.42 (0.98)

Note: GMAT-Q, GMAT-V, and GMAT-AW = Graduate Management Admission Test Quantitative, Verbal, and Analytical Writing, respectively. UGPA = undergraduate grade point average; Z-GPA = graduate school grade point average, standardized within school.

tor and criterion variables must be considered. These statistics are summarized in Table 4. When subgroup differences appear on both the predictor and criterion variables, expected relationships between the residuals occur. For example, if a subgroup were lower on both predictor and criterion variables, there would be relatively less room to underpredict criterion performance and the errors of prediction would show a pattern of overprediction.

An inspection of the means and standard deviations in Table 4 illustrates that men and women are similar on GMAT scores and UGPA, but women earned a noticeably lower average first-year GPA (about 1/3 of a standard deviation lower). With respect to race/ethnicity, Whites scored highest on all variables except GMAT-Q, where Asian Americans scored higher. African Americans scored about one standard deviation lower than Whites on GMAT-Q, GMAT-AW, and standardized first-year GPA (Z-GPA) and about three quarters of a standard deviation lower on GMAT-V and UGPA.

Multiple regression results. Given that 3 of the 11 schools did not report data on GMAT-AW, and given the relatively small sample sizes for some student groups, the analysis of differential predictive validity focused only on GMAT-Q, GMAT-V, and UGPA.

To evaluate differential predictive validity, first-year GPA was standardized within each school and separate regression equations were conducted for each subgroup by aggregating the data across schools. Although this approach is not ideal, it allowed for aggregating the data for minority students across schools. Furthermore, the appropriateness of using the standardized GPA across schools can be evaluated by calculating the regression coefficients for predicting Z-GPA based on the entire sample and comparing them with the average values from the within-school analyses (reported previously in Table 3).

The regression results for the entire sample based on Z-GPA are presented in Table 5. The R from this analysis was .462, which yielded an R^2 of .213. These values are similar to, but slightly lower than, those obtained using the weighted mean in the within-school analyses (.499 and .249, respectively; see Table 3). This finding suggests about 3.6% of the variation in GPA was lost when collapsing Z-GPA across schools.

Given that caveat, we turn to the results from the group-specific regression equations, which are summarized in Table 6. The multiple R is large and statistically significant for all groups, which supports the conclusion of adequate predictive validity for all groups. The proportion of variance in GPA accounted for by GMAT-Q, GMAT-V, and UGPA ranged from about 14% (Hispanic/Latino[a]) to 27% (African American). Statistical tests for differences between the regression coefficients (including the intercepts) for men and women were conducted, as well as for differences between Whites and the other racial/ethnic subgroups. Across all comparisons, no statistically significant differences were found (i.e., all $p > .05$), which supports the hypothesis of no differential prediction across majority and minority groups.

Given the sample sizes associated with these statistical tests, the power to detect a statistically significant difference at $p < .05$ was very high for all groups, assuming a

Table 5
Summary of Across-School Regression Analyses Using Z-GPA

Predictor	B	$r^2_{Y(i.jkl)}$
GMAT-Q	.285	.081
GMAT-V	.239	.056
UGPA	.234	.064
Model	R =.462, R^2 = .213	

Note: Z-GPA = graduate school grade point average, standardized within school; B = standardized regression coefficient. $r^2_{Y(i.jkl)}$ = squared semipartial correlation for predictor; GMAT-Q, GMAT-V, and GMAT-AW = Graduate Management Admission Test Quantitative, Verbal, and Analytical Writing, respectively. All statistics are statistically significant at $p < .0001$.

Table 6
Summary of Within-Group Regression Analyses

			Standardized Regression Coefficients			
Group	R	R^2	GMAT-Q	GMAT-V	UGPA	Intercept
Men	.469	.220	.259	.244	.260	–7.030
Women	.423	.179	.265	.207	.223	–6.601
African American	.522	.272	.402	.204	.236	–7.132
White	.466	.217	.304	.161	.263	–6.918
Asian American	.393	.155	.287	.139	.181	–6.612
Hispanic/Latino(a)	.373	.139	.263	.073	.264	–6.025

Note: GMAT-Q, GMAT-V and GMAT-AW = Graduate Management Admission Test Quantitative and Verbal, and Analytical Writing, respectively; UGPA = undergraduate grade point average. All R, R^2, and intercepts are statistically significant at $p < .0001$. All regression coefficients are statistically significant at $p <$ .0001 except for GMAT-Q and UGPA for Hispanic/Latino(a), which was statistically significant at $p = .002$; and GMAT-V for Hispanic/Latino(a), which was not statistically significant ($p = .380$).

medium effect size (i.e., standardized mean difference = .5) existed in the population (all power calculations were greater than .99 using these assumptions). Assuming a *small* effect size in the population (i.e., standardized mean difference = .2), the power to detect such a difference given these sample sizes was very high (greater than .97) for all analyses except White–Hispanic/Latino(a) (power = .60) and White–African American (power = .72).

Residual analyses. A summary of the residual analyses is presented in Table 7. The mean and standard deviations of the residuals for each group, expressed on the standardized GPA scale ($M = 0$, $SD = 1$), are presented. As expected, the groups that had higher mean scores on the predictors and criterion variables (i.e., men and Whites) were, on average, underpredicted (i.e., positive residuals), and those groups lower on

Table 7
Summary of Residual Analysis

Group	*n*	*M*	*SD*
Men	2,947	0.000	0.866
Women	1,225	0.000	0.879
African American	174	–0.278	0.786
Asian American	474	–0.192	0.876
Hispanic/Latino(a)	131	–0.141	0.933
White	2,165	0.116	0.936

Note: Predictors are undergraduate grade point average (UGPA), Graduate Management Admission Test Quantitative (GMAT-Q), and Graduate Management Admission Test Verbal (GMAT-V). Criterion was standardized ($M = 0$, $SD = 1$) first-year graduate GPA.

these variables were overpredicted (i.e., negative residuals). However, the magnitude of these overprediction and underprediction errors is small, the largest being about one quarter of a standard deviation.

To put the magnitude of prediction error in perspective, the mean residual for African Americans in one of the schools with the largest number of these students was transformed onto the GPA scale. The amount of overprediction for these students was about .05 on a GPA scale ranging from 0 to 4.0. For this school, the mean GPA for African Americans predicted from UGPA, GMAT-Q, and GMAT-V was 3.32 and the actual GPA was 3.27. Given that African Americans had the largest average prediction error, it appears that the degree of prediction error for all groups on the typical 0 to 4.0 GPA scale is small.

Discussion

The results of this study support several conclusions about the validity of GMAT scores for use as admissions criteria for graduate management schools. Overall, the results suggest that GMAT-Q and GMAT-V scores are good predictors of first-year GPA in graduate school, even after accounting for students' undergraduate GPA (see Tables 2 and 3). In addition, the predictive utility of these scores seems to hold across subgroups of examinees defined by sex and race/ethnicity (see Table 6). In fact, when separate regression equations were fit for women and men, no statistically significant differences in regression coefficients were found.

The same finding of no statistical significance across regression coefficients for GMAT-Q and GMAT-V was found when comparing coefficients derived from the data for White students to those derived from the data for African American, Asian American, and Hispanic/Latino(a) students. Our power analyses indicate we had substantial power to detect differences in predictive validity across these student subgroups, if the difference in the population represented a moderate effect size (i.e., 0.5 standard deviations). However, our power analyses also revealed that if there were small (i.e., 0.2 standard deviations), but true, differences in predictive validity, our sample sizes for African Americans and Hispanics/Latinos(as) may have been to small to detect the difference using a criterion of statistical significance. Thus, although the results of our analyses generally support consistency of predictive validity across subgroups, future studies using larger samples of African American and Hispanic/Latino(a) students will be better able to detect small differences across these groups, should they exist.

It is noteworthy that when we looked at small differences in prediction across groups by comparing patterns of the errors of prediction (residuals), women and minorities tended to be overpredicted, which means their GMAT scores predicted they would have *higher* GPAs than they actually earned.

With respect to GMAT-AW, the results suggest that, on average, the predictive utility of GMAT-AW scores is relatively low (see Tables 2 and 3). After accounting for GMAT-Q, GMAT-V, and UGPA, GMAT-AW accounts for only about an additional 1% of the variation in graduate school GPA. Given that not all of the participating schools reported GMAT-AW scores, analyses of the differential prediction when including these scores was not conducted.

Limitations of the Study

Like all studies, ours has limitations. First, only 11 schools participated in the study, and the degree to which the findings generalize to other schools is unknown. Second, it should be kept in mind that the validity of admissions test scores is not purely an issue of predictive validity. Criteria such as GPA are not perfect measures of success in graduate school, and many other variables that play a role in graduate school success, such as diligence, cannot be measured on a standardized test. Future research should investigate other aspects of the utility and validity of GMAT scores such as consistency of factor structure across subgroups of examinees (construct validity) and the perceived utility of the scores by admissions officers and faculty. Future research should also strive to gather more data on underrepresented minority groups such as Native Americans, African Americans, and Hispanics/Latinos(as) and acquire more data on GMAT-AW.

Another limitation of the study is that 38% of the students did not report race/ethnicity. Categories such as "race" and "ethnicity" are imperfect descriptors of students, and increasing numbers of students choose not to self-report themselves into such categories. The degree to which nonreporting of race/ethnic information affected the results of this study is unknown.

The study was also limited by the constraints of a small set of participating schools. Given that some variation across schools was noted, it would be interesting to conduct a mixed-effects model, such as hierarchical linear modeling, where we could explicitly study factors that affect the regression coefficients across schools. However, such analyses will require larger numbers of participating schools (e.g., at least 30; see Bryk & Raudenbush, 1992).

Concluding Remarks

Notwithstanding the aforementioned limitations, this study provides important information in support of the validity of GMAT scores. The results indicate GMAT-Q and GMAT-V scores are valid predictors that do not appear to be biased against women or minorities. Given the great care that goes into the development of GMAT tests, including the quality control checks such as item bias reviews and analysis of differential item functioning (Hecht & Schrader, 1986), this finding is not surprising. The data analyzed in this study lead to the conclusion that GMAT scores can be interpreted similarly across men, women, minority, and nonminority students. The utility of GMAT-AW needs further study and would require a larger number of schools that could report these scores.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Braun, H., Ragosta, M., & Kaplan, B. (1988). Predictive validity. In W. W. Willingham, M. Ragosta, R. E. Bennet, H. Braun, D. A. Rock, & D. E. Powers (Eds.), *Testing handicapped people* (pp. 109-132). Needham Heights, MA: Allyn & Bacon.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Graduate Management Admission Council. (2002). *Guide to the use of GMAT scores: Effective October 1, 2002*. Princeton, NJ: Educational Testing Service.

Graduate Management Admission Council. (2003). *GMAT validity: A three-year summary, 1999-2000 through 2001-2002*. McLean, VA: Author.

Hecht, L., & Schrader, W. (1986). *Graduate Management Admission Test: Technical report on test development and score interpretation for GMAT users*. Princeton, NJ: Educational Testing Service.

Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.). Belmont, CA: Wadsworth.

Koenig, J. A., Sireci, S. G., & Wiley, A. (1998). Evaluating the predictive validity of MCAT scores across diverse applicant groups. *Academic Medicine, 73*, 65-76.

Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement, 21*, 33-47.

Linn, R. L., & Hastings, C. N. (1984). A meta-analysis of the validity of predictors of performance in law school. *Journal of Educational Measurement, 21*, 245-259.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Olsen, M. (1957). *The admission test for graduate study in business as a predictor of first-year grades in business school, 1954-1955.* Princeton, NJ: Educational Testing Service.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

Powers, D., & Moss, P. (1980). *A summary of the results of the Graduate Management Admission Council (GMAC) Validity Study Service for 1979-1980.* Princeton, NJ: Educational Testing Service.

Wainer, H., & Sireci, S. G. (2005). Item and test bias. In *Encyclopedia of social measurement* (Vol. 2, pp. 365-371). San Diego, CA: Elsevier.

Wightman, L., & Leary, L. (1985). *GMAC Validity Study Service: A three-year summary (1981-82 through 1983-84).* Princeton, NJ: Educational Testing Service.

Zhao, J. C., Patsula, L., Stone, E., Paez, R., Babula, J., Tye, E., et al. (2000). *GMAT Validity Study Service: A three-year summary (1996-97 through 1998-99).* McLean, VA: Graduate Management Admission Council.