GMAC®

Graduate
Management
Admission
Council®
Creating Access to Graduate Business Education®

Predicting Success in Graduate Management Doctoral Programs

Kara O. Siegert

GMAC® Research Reports • RR-07-10 • July 12, 2007

Abstract

An integral part of the test evaluation and improvement process involves providing evidence that valid inferences can be made based on test scores. Additionally, it is imperative to provide evidence that validity results can be generalized to all potential populations being administered an exam. This study evaluated the predictive validity of the Graduate Management Admission Test® (GMAT®) for doctoral students enrolled in 18 different graduate business-oriented programs. Results indicated that the GMAT® exam was a better predictor of first-year performance than was previously reported. Moreover, GMAT® scores were better predictors of grades than undergraduate grade point average. Results for gender, English skill, and program concentration subgroups are also described.

Purpose

Growing media attention surrounds the use, reliability, and validity of test scores for admission into higher education. With recent concerns over scoring errors and misuse of test scores, admission testing is subjected to an ever-increasing level of public scrutiny. It is not surprising then that admission testing programs are under pressure to provide evidence that they are fulfilling their missions by using instruments that enable valid and reliable inferences about future student performance. A battery of analyses and strategies are used to continually evaluate a test. Standards specify that validity and reliability be investigated by multiple methods, on numerous occasions, and for all groups or subgroups of the population that are administered the test (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Though it is often difficult to gather data on all subgroups, a concentrated effort should be made to provide comprehensive evidence about whether inferences can be validly and reliably drawn.

The purpose of the current study is to examine the validity of inferences made from exam scores on the Graduate Management Admission Test® (GMAT®) for graduate management doctoral programs. Although the GMAT® exam is administered to more than 200,000 examinees a year, a small but important percentage of these examinees are not applying to Master of Business Administration (MBA) programs. One specific examinee subgroup completes the GMAT® exam in hopes of entering into doctoral graduate study in management education. However, much of the research for the GMAT® exam has been based on the validity of inferences for mid-program performance in MBA programs, not doctoral-level study (Graham, 1991; Hecht & Schrader, 1986; Kuncel, Credé, & Thomas, 2004; Olsen, 1957; Talento-Miller & Rudner, 2005). The current study applied a similar methodology for predicting first-year success for students enrolled in doctoral programs in business-related fields. Validity was also examined separately for different gender, English ability, and program concentration subgroups.

Theoretical Framework

The GMAT® exam, which has been in existence since 1954, currently measures verbal, quantitative, and analytical writing skills using three separate sections (Graduate Management Admission Council® (GMAC®), 1999). The three sections that comprise the GMAT® exam are officially titled the GMAT® Verbal (V), GMAT® Quantitative (Q), and GMAT® Analytical Writing Assessment (AWA). The V and Q sections are administered in multiple-choice formats, whereas the

AWA section requires examinees to respond to writing prompts by composing two essays. Upon completing the exam, examinees are provided with four separate scores—one each for the V, Q, and AWA sections and one Total score. The Total score is a combination of the V and Q section scores, but it is rescaled to a different metric. The validity of the inferences made from the GMAT® assessment has been researched continually during the past 50 years (Graham, 1991; Hecht & Schrader, 1986; Kuncel et al., 2004; Olsen, 1957; Talento-Miller & Rudner, 2005). Although studies such as these have shown that GMAT® scores are useful in predicting graduate performance in Master's-level programs, it is necessary to determine if inferences can be extrapolated to other program types.

In 1993, Zwick conducted the first validity study examining the relationship between GMAT® exam scores and performance for doctoral-level management and business students. The study examined V and Q sections of the GMAT® exam along with undergraduate grade point average (UGPA) as predictors of success for 5,219 doctoral programs representing 36 different schools. Success was measured as first-year grade point average (FYGPA) and final grades. The results suggested that UGPA was a better predictor of FYGPA and final grades than prediction equations including V and Q scores. However, the best prediction resulted when UGPA, V, and Q scores were combined as predictors. Overall the results revealed lower predictive validity values for doctoral students than those previously reported for MBA students.

Unfortunately, the AWA was not included as a predictor in the Zwick (1993) study because it had not yet been added to the GMAT® exam. Additionally, the correlations used to predict the relationship between the predictors and criteria were not corrected for restriction of range. Restriction of range occurs when enrolled students' scores do not span the range of all possible scores examinees can receive. With most graduate business programs, the admitted students are at the upper end of the score range for both GMAT® scores and UGPA. After all, excellence in prior performance is a large part of why students are admitted to graduate programs. Additionally, FYGPA and final grades often lack variability because most admitted students are well-prepared for graduate study and thus perform at the upper-end of the grading scale. The effects of restriction of range and low variance on validity estimates are often lower prediction values.

The purpose of the current study is to collect further evidence of GMAT® exam validity for inferences made about success in doctoral business programs. To add to the previous results of Zwick's (1993) study, AWA scores were included as an additional predictor of success. A correction for restriction of range was also introduced and applied to offer more accurate estimates of predictive validity. Finally, the current study expanded on Zwick's study by providing results separately for gender, English skill, and program concentration groups.

Method

Sample

To solicit participation, electronic and postal mailings were sent out to approximately 265 doctoral business program faculty members, representing programs in 12 different countries. A GMAC® database was used to solicit all doctoral faculty members who were listed as contacts within their programs for GMAT® exam-related inquiries. Members of the group DocNet, which is made up of faculty and staff from doctoral degree programs in business and management and facilitates the improvement of doctoral business education programs, were also solicited. A total of 18 programs, with sample sizes ranging from 24–100, ultimately submitted their data for participation in the study. All programs were located within the United States.

Participants were asked to input their student-level data into a provided template and e-mail the template to GMAC® for analysis. The programs provided information on predictors such as UGPA, V, Q, AWA, and Total scores, and the criteria of FYGPA and second-year grades (SYGPA). Student or individual-level (IL) data such as these are commonly collected during the admission process and the progression of students through program coursework. For their participation, programs each received an individualized program-level (PL) report detailing the validity of their admission procedures.

A few programs submitted additional student-level data. For instance, some programs reported prior graduate degree attainment (yes or no) as an added predictor, and criteria variables that did not focus on student grades, such as number of publications, graduation within five years of program entry (yes or no), and job placement aligned with the program's objectives (yes or no), were also added. However, because only a limited number of programs provided student-level data on these variables, graduate degree attainment, number of publications, graduation within five years of program entry, and job placement aligned with the programs' objectives were not included as variables in the analyses.

Analysis

Program Level

For this study, data were analyzed using two different methodologies, PL and IL. The first method used the results from the 18 PL studies that were conducted and compiled into reports for each program. When results from all 18 PL reports were examined, there were a total of 1,006 doctoral students with scores for both the GMAT® exam and FYGPA in the program. Because data provided by each program represented only students admitted to the program, rather than all applicants, average exam scores and subsequent grades are higher than what would be expected from a sample of all applicants to the program. Often, very few students who have low GMAT® exam scores and/or GPA values are included in the data, thus limiting the variability of the data. As a result, validity estimates based on admitted students are often lower than what would be expected had a sample of all program applicants been used. Thus, restriction of range corrections were performed at the PL using the formula proposed by Hunter and Schmidt (1990):

$$r_{ij}^{*} = \frac{Ur_{ij}}{\sqrt{\left(U^{2}-1\right)r_{ij}^{2}+1}}$$

Each program's correction was based on the ratio of variance in the IL observed data provided by the program to the variance in the population of applicants. The population variance for a given program was based on data collected from all examinees who sent GMAT® scores to that particular school during the 2002–2004 testing years. Thus, each program's correction differed depending on the sample of GMAT® examinees electing to send their scores to that school during the specified time period. It has been argued that examinees' decisions to send scores to certain institutions can be interpreted as interest in applying to those schools (Stolzenberg & Relles, 1985). If this is true, then these scores more accurately represent the applicant population than do the restricted scores that come only from students who are enrolled. On average, bivariate correlations between GMAT® scores and UGPA with grades increased less than .06 when PL correlations were corrected for range restriction.

The PL results summarize predictive validity estimates across the 18 programs. Talento-Miller, Rudner, and Owens (2006) described multiple methods for evaluating predictive validity across studies that included the same predictors (i.e., GMAT® scores and UGPA) and outcomes (i.e., program grades). PL results were summarized by calculating the mean and median validity estimates across the 18 programs. The median values provided estimates not biased by extremely low or high validity values. Based on previous research (Talento-Miller et al., 2006), this method yields appropriate average estimates of PL validity. It should be noted, however, that it is inappropriate to summarize PL subgroup (e.g., gender and ethnicity) validity by calculating the mean and median validity estimates obtained from subgroup analyses across multiple studies. As a result, IL analyses were used to examine predictive validity for various subgroups of doctoral students.

Individual Level

For the IL analysis, students were pooled across all 18 participating programs for a total of 1,148 students. The IL results were used to obtain validity estimates for gender, English skill, and program concentration groups. At the program level, there were not enough data available for most of the individual programs to provide accurate validity estimates for the various groupings. As a result, data were pooled across all programs to provide one large sample of doctoral students.

Similar to those for the PL data, correlations between the admission variables and program performance were corrected for restriction of range for the IL data. For the IL analysis, restriction of range was corrected using the same Hunter and Schmidt (1990) formula proposed earlier. However, for the IL data, the population data used to make the corrections was based on all GMAT® examinees who elected to send their GMAT® scores to at least one of the 18 participating programs and indicated that they were interested in applying to doctoral programs. To control for PL variance, FYGPA and SYGPA were standardized within each program, and dummy coding was used for all subsequent regression analyses using the IL data. Dummy codes allowed for program effects or differences to be accounted for when predicting performance for students from various programs.

Additionally, the results of the current study were examined separately for various subgroups of the doctoral student population. The IL data were evaluated for gender, English skill, and program concentration subgroup differences. Sample sizes for the various subgroups were smaller than anticipated because some programs did not identify subgroups for all of their students. As a result, the ability to generalize from some of the IL subgroup analyses is limited.

Results

Program Level

FYGPA

A table identifying the abbreviations used in all subsequent tables in this paper, as well as bivariate and multiple correlations for each of the 18 programs, can be found in the Appendix. Bivariate and multiple correlations (R) with FYGPA averaged across the 18 PL validity studies are reported in Table 1. On average, combinations that included all of the predictors had the highest validity values. Specifically, the combination that included V, Q, AWA, and UGPA, had the highest predictive validity value (median = .447). Predictive validity values between .30–.40, which account for roughly 9–14% of variance (R^2) in performance, are usually considered acceptable in admission testing (Kaplan & Sacuzzo, 1997). For the present study, the combination that yielded the highest estimate of predictive validity accounted for 20% of the variance in FYGPA for doctoral students. Additionally, GMAT® Total (median = .342) and the combination of V and Q scores (median = .337) were better predictors than UGPA (median = .243). These estimates are higher than those found with the combination of V and Q scores (median = .20) and UGPA (median = .25) in Zwick's (1993) study. However, Zwick's values were not corrected for restriction of range, whereas the values for the current study were.

Table 1. Summary of First-Year Predictive Validity across All Programs

	V	Q	A	U	VQ	T	VQAU	TAU
N	18	18	16	17	18	18	18	18
Mean	.132	.178	.116	.257	.360	.307	.490	.466
SD	.175	.319	.186	.116	.218	.241	.203	.144
Median	.104	.156	.096	.243	.337	.342	.447	.432
25th	.009	.082	.002	.213	.136	.158	.334	.367
75th	.251	.431	.265	.295	.499	.501	.633	.533

SYGPA

Predictive validity for SYGPA was summarized across all programs and presented in Table 2. Only 11 of the 18 participating programs provided student data on performance during the second year of coursework. As a result, the generalizability of the findings is limited. The highest median multiple correlation resulted when V, Q, and AWA scores were combined with UGPA (R = .616). This combination predicts quite well, accounting for approximately 40% of the variation in SYGPA.

Table 2. Summary of Second-Year Predictive Validity across All Programs								
	V	Q	A	U	VQ	T	VQAU	TAU
N	11	11	9	10	11	11	11	11
Mean	.156	.227	.101	.232	.499	.405	.618	.547
SD	.264	.425	.290	.221	.251	.261	.281	.211
Median	.171	.087	.076	.231	.508	.377	.616	.505
25th	.079	–.063	–.044	.086	.244	.207	.344	.370
75th	.331	.587	.323	.357	.732	.673	.916	.721

When compared with the results for FYGPA, validity estimates were higher when predicting performance beyond the first year. AWA score was only slighted correlated with first- and second-year performance, R = .096 and R = .076, respectively. GMAT® Total score demonstrated a strong relationship with FYGPA (R = .342) and SYGPA (R = .377). Differences between FYGPA and SYGPA were apparent when estimates for V and Q were compared. Typically V score was a better predictor of SYGPA (median R = .171) than FYGPA (median R = .104). Yet, Q score was a better predictor of FYGPA (median R = .156) than SYGPA (median R = .087).

Individual Level

Though the PL results allow for interpretation of predictive validity for various doctoral programs, the IL data were aggregated to examine prediction across a sample of doctoral students. In a recent study (Talento-Miller et al., 2006), several different methods for analyzing validity data, similar to data presented for this study, were compared. The results revealed that pooling IL student data across several programs yielded average predictive validity values that were lower than average values estimated by pooling data at the PL. Thus, it would be expected that the IL results for this study would be lower than the PL results in the previous section. Table 3 presents the corrected bivariate correlations between the variables included in the study. Please note that PL effects are not controlled for in this matrix. Thus, the relationship between the variables is likely lower than anticipated.

Table 3. Individual Level Corrected Correlations							
	FYGPA	SYGPA	V	Q	A	U	T
FYGPA	1.000 (1068)						
SYGPA	.621 (634)	1.000 (710)					
V	.133 (945)	.146 (633)	1.000 (1046)				
Q	.124 (945)	.093 (633)	.022 (945)	1.000 (1044)			
A	.173 (642)	.112 (380)	.389 (644)	–.062 (644)	1.000 (721)		
U	.181 (869)	.176 (568)	.155 (801)	.158 (799)	.077 (528)	1.000 (971)	
T	.204 (946)	.162 (636)	.702 (944)	.699 (944)	.236 (646)	.211 (800)	1.000 (1046)

FYGPA

Once the bivariate correlations were corrected, these corrected correlations were used in subsequent regression analyses. As mentioned previously, regression analyses for the IL data included dummy coded variables to account for variability among the programs. Variance accounted for by the dummy codes was included in the correlations that are presented. Table 4 presents the corrected multiple correlations with PL variance accounted for by dummy codes. Additionally, the relative importance of variables is presented in Table 4 as calculated using the Pratt Index formula presented here.

$$PI_i = \frac{r^*_{ij}\beta_i}{R^2}$$

In this formula, r^*_{ij} is the adjusted bivariate correlation of predictor variable *i* with criterion variable *j*; β_i is the standardized beta weight for variable *i*; and R^2 is the squared multiple correlation of the set of variables with *j*. Pratt index values are representative of the percentage of variance accounted for by each variable in the multiple correlation. The combination including GMAT® Total, AWA, and UGPA yielded the strongest relationship with FYGPA [R = .377, F (17, 509), p < .00]. Table 5 presents the results from the simultaneous regression analysis for this combination. Combined, GMAT® Total, AWA, and UGPA accounted for approximately 14% of the variability in FYGPA for doctoral management education students. All of the predictors significantly contributed to prediction, with GMAT® Total yielding the strongest unique relationship.

Table 4. Individual Level Prediction of First-Year Grades

	N	Correlation with FYGPA	Contribution to Prediction				
			V	Q	A	U	T
V	945	.199					
Q	945	.179					
A	642	.229					
U	870	.210					
VQ	946	.263	.56	.44			
T	947	.290					
VQAU	526	.342	.17	.22	.30	.30	
TAU	527	.377			.21	.27	.52

Table 5. Simultaneous Regression Analysis Predicting First-Year Grades from GMAT® Total, AWA, and UGPA

Variable	$R^2 = .142$*			
	B	*SEB*	β	sr^2
Total	.005	.001	.314*	.061
AWA	.183	.053	.152*	.020
UGPA	.429	.101	.184*	.031

Note. Results for PL dummy codes are not presented because PL variance is treated as error.
SEB = standard error of *B*; sr^2 = squared semi-partial.
*p < .01.

When compared to the best predictor combination for the PL results, the best predicting combination for IL resulted, as expected, in lower predictive validity, R = .432 vs. R = .377, respectively. It should also be noted that for the PL results, the best predicting combination included V, Q, AWA, and UGPA. For IL results, the combination that included T, AWA, and UGPA resulted in the best prediction. The use of additional dummy codes for the IL data may help account for additional PL differences.

SYGPA

Table 6 presents the corrected correlations for SYGPA, as well as information on the contribution of the different variables to prediction. Again, PL variability was accounted for by using dummy codes. It should be noted that the sample size for SYGPA analyses were considerably smaller than those used for FYGPA analyses. As a result, the ability to generalize second-year findings is restricted. As with the results presented for FYGPA, the combination including GMAT® Total, AWA, and UGPA yielded the strongest relationship with SYGPA [R = .341, F (14, 281), p <.00]. The results from this simultaneous regression analysis can be found in Table 7. When beta (β) values are examined, it is clear that only GMAT® Total and UGPA significantly contributed to the prediction of SYGPA. GMAT® AWA scores do not appear to meaningfully influence performance during the second year.

Table 6. Individual-Level Prediction of Second-Year Grades

	N	Correlation with SYGPA	Contribution to Prediction				
			V	Q	A	U	T
V	633	.194					
Q	633	.149					
A	380	.167					
U	568	.195					
VQ	633	.241	.66	.34			
T	636	.249					
VQAU	294	.330	.25	.25	.10	.41	
TAU	296	.341			.10	.37	.54

Table 7. Simultaneous Regression Analysis Predicting Second-Year Grades from GMAT® Total, AWA, and UGPA

Variable	R^2 = .116*			
	B	*SEB*	β	sr^2
Total	.004	.001	.276*	.043
AWA	.103	.073	.085	.006
UGPA	.419	.138	.180*	.029

Note. Results for PL dummy codes are not presented because PL variance is treated as error.
SEB = standard error of *B*; sr^2 = squared semi-partial.
*p < .01.

Similar to the results for FYGPA, the V, Q, and AWA scores had higher estimates of validity for IL data than they had for the PL data for SYGPA. However, UGPA; GMAT® Total; the combination of V and Q; the combination of V, Q, AWA, and UGPA; and the combination of GMAT® Total, AWA, and UGPA resulted in higher predictive validity for the PL analyses. This is likely due to the high variation in V, Q, and AWA scores at the individual level, which is balanced out at the program level. Applicants or students often perform better on one section of the GMAT® exam, resulting in a wider range of V and Q scores at the IL. Also, programs typically select applicants based on GMAT® Total scores, not V or Q. As a result, variability in Total scores is minimized at the IL, whereas variability among V and Q scores is increased.

One advantage of examining IL data, rather than PL data, is the ability to study relationships for different subgroups of doctoral students. For this study, predictive validity was examined for various groups based on: gender, English language ability, and program concentration. Since a limited number of programs submitted data on SYGPA and the separation of data into subgroups resulted in even smaller sample sizes, subgroup results are only presented for FYGPA.

Subgroups

Gender. Variation in predictive validity was found when males and females were compared. Table 8 presents the results of separate regression analyses conducted for the two gender subgroups. Predictive validity was slightly higher for females than it was for male doctoral students. However, GMAT® Total was the strongest individual predictor for both groups with regard to FYGPA.

Table 8. First-Year Predictive Validity by Gender Subgroup

	Females *R*(*N*)	Males *R*(*N*)
V	.321 (382)	.247 (544)
Q	.313 (382)	.261 (544)
A	.337 (267)	.263 (365)
U	.349 (355)	.226 (493)
VQ	.381 (382)	.336 (544)
T	.423 (382)	.344 (545)
VQAU	.548 (209)	.350 (287)
TAU	.545 (218)	.394 (297)

Although the combination that included V, Q, AWA, and UGPA [$R = .548$, $F(18, 190)$, $p < .00$] produced results similar to those for the GMAT® Total, AWA, and UGPA combination [$R = .545$, $F(17, 200)$, $p < .00$] for females, FYGPA was better predicted by GMAT® Total, AWA, and UGPA for males [$R = .394$, $F(17, 279)$, $p < .00$]. These results can be found in Tables 9 and 10.

When examining predictive validity for gender subgroups, it can be seen that the distinction in predictive validity estimates was most prevalent when AWA and UGPA were added to the mix of predictors. It appears that these variables were related to a much greater increase in prediction for female students than for their male counterparts.

Table 9. Simultaneous Regression Analysis Predicting First-Year Grades from GMAT® Verbal, GMAT® Quantitative, AWA, and UGPA for Gender Subgroups

Variable	*B*	*SEB*	β	sr^2	R^2
Females					.300*
V	.060	.013	.375*	.075	
Q	.042	.011	.263*	.049	
AWA	.219	.082	.182*	.027	
UGPA	.311	.163	.128	.013	

Table 9. Simultaneous Regression Analysis Predicting First-Year Grades from GMAT® Verbal, GMAT® Quantitative, AWA, and UGPA for Gender Subgroups					
Variable	*B*	*SEB*	β	sr^2	R^2
Males					.123*
V	.026	.012	.163	.015	
Q	.039	.013	.212*	.030	
AWA	.154	.079	.130	.013	
UGPA	.341	.137	.153	.020	

Note. Results for PL dummy codes are not presented because PL variance is treated as error.
SEB = standard error of *B*; sr^2 = squared semi-partial.
*p < .01.

Table 10. Simultaneous Regression Analysis Predicting First-Year Grades from GMAT® Total, AWA, and UGPA for Gender Subgroups					
Variable	*B*	*SEB*	β	sr^2	R^2
Females					.297*
Total	.007	.001	.434*	.113	
AWA	.257	.077	.214*	.040	
UGPA	.432	.158	.177*	.026	
Males					.155*
Total	.005	.001	.351*	.069	
AWA	.180	.072	.151	.019	
UGPA	.376	.130	.169*	.025	

Note. Results for dummy codes are not presented because PL variance is treated as error.
SEB = standard error of *B*; sr^2 = squared semi-partial.
*p < .01.

English ability. Programs were asked to provide information on students' written and spoken English ability by indicating if there was a concern regarding the student's English skills (questionable skills) or if there were no concern (acceptable skills). Similar to the other regression analyses, various predictors and combinations of predictors were examined to determine the admission variable(s) that provided the greatest amount of information about FYGPA for these subgroups. The large discrepancy in sample sizes for the subgroups should be noted when interpreting predictive validity values. There were very few students for whom programs indicated that there was some concern over their spoken or written English ability. Thus, caution should be used when interpreting the findings, as they may not replicate with a larger sample.

Table 11 presents the validity estimates for these two subgroups. GMAT® Total was the best individual predictor of success for students with English skills that were deemed acceptable. However, AWA scores were the most effective individual predictor of FYGPA for students with questionable skills. Additionally, V scores highly correlated with first-year success for students with questionable English ability. These findings may indicate that the AWA and V sections are particularly useful predictors of success for non-native English speakers and some international students.

Table 11. First-Year Predictive Validity by English Ability Subgroup		
	English Acceptable *R*(*N*)	English Questionable *R*(*N*)
V	.197 (727)	.517 (90)
Q	.211 (727)	.372 (90)
A	.211 (452)	.563 (79)
U	.219 (640)	.279 (80)
VQ	.280 (727)	.530 (90)
T	.331 (758)	.451 (95)
VQAU	.347 (401)	.695 (74)
TAU	.352 (386)	.687 (70)

As with the individual predictors, the combinations revealed a distinction between estimates of predictive validity for the two subgroups. Predictive validity for the subgroup with questionable English skills was again higher than it was for the subgroup with acceptable English ability. Although the combination with section scores was a better predictor for the subgroup with questionable skills [$R = .695$, F (10, 63), $p < .00$], the combination with GMAT® Total was more effective for the subgroup with acceptable skills [$R = .352$, F (15, 370), $p < .00$]. The regression analysis results for these combinations and subgroups can be found in Tables 12 and 13. Again, the small sample size for the questionable skills subgroup makes it difficult to determine whether the findings would replicate.

Table 12. Simultaneous Regression Analysis Predicting First-Year Grades from GMAT® Verbal, GMAT® Quantitative, AWA, and UGPA for English Skill Subgroups					
Variable	*B*	*SEB*	β	sr^2	R^2
Acceptable					.120*
V	.011	.010	.068	.003	
Q	.030	.010	.181*	.023	
AWA	.190	.065	.158*	.020	
UGPA	.504	.112	.227*	.046	
Questionable					.483*
V	.086	.019	.495*	.171	
Q	.002	.024	.009	.000	
AWA	.495	.131	.401*	.116	
UGPA	–.431	.291	–.155	.018	

Note. Results for PL dummy codes are not presented because PL variance is treated as error.
SEB = standard error of *B*; sr^2 = squared semi-partial.
*p < .01.

Table 13. Simultaneous Regression Analysis Predicting First-Year Grades from GMAT® Total, AWA, and UGPA for English Skill Subgroups					
Variable	*B*	*SEB*	β	sr^2	R^2
Acceptable					.124*
Total	.004	.001	.240*	.034	
AWA	.160	.062	.133*	.016	
UGPA	.501	.114	.226*	.046	
Questionable					.472*
Total	.011	.002	.570*	.228	
AWA	.420	.135	.340*	.085	
UGPA	–.236	.295	–.085	.006	

Note. Results for dummy codes are not presented because PL variance is treated as error.
SEB = standard error of *B*; sr^2 = squared semi-partial.
*p < .01.

Program concentration. Programs were also asked to indicate the program concentration in which students were enrolled from a list of seven options. The options listed were: finance, organizational behavior, accounting, marketing, operations, other with a quantitative focus, and other with a non-quantitative focus. When data were aggregated across the 18 programs, there were six program concentrations with sample sizes greater than 50. Program concentrations with sample sizes of 50 or smaller were not reported because the small sample sizes might not have been representative of the larger population of students enrolled in those concentrations.

Table 14 presents the correlations between the various predictors and FYGPA. The strongest individual predictor varied among the different program concentrations. For instance, the AWA was the best single predictor for accounting and organizational behavior, whereas for marketing and finance students, FYGPA was best predicted by Q scores. Quantitative and operations students were even more unique: The best individual predictors for performance in these concentrations were V or UGPA, respectively. Unlike the results from the previous subgroup analyses, individual predictors were almost as effective at predicting success as combinations of variables for some of the program concentration subgroupings.

Table 14. First-Year Predictive Validity by Program Concentration Subgroup						
	Accounting *R*(*N*)	Marketing *R*(*N*)	Finance *R*(*N*)	Quantitative *R*(*N*)	Organizational Behavior *R*(*N*)	Operations *R*(*N*)
V	.391 (119)	.589 (62)	.387 (142)	.623 (158)	.699 (123)	.477 (91)
Q	.391 (119)	.674 (62)	.621 (142)	.541 (158)	.692 (123)	.531 (91)
A	.485 (87)	—	.366 (96)	.475 (113)	.708 (77)	.583 (64)
U	.468 (98)	—	.525 (128)	.327 (181)	.485 (117)	.617 (100)
VQ	.391 (119)	.830 (62)	.623 (142)	.686 (158)	.724 (123)	.710 (92)
T	.318 (129)	.762 (62)	.442 (153)	.564 (158)	.635 (133)	.645 (98)

Table 14. First-Year Predictive Validity by Program Concentration Subgroup						
	Accounting *R*(*N*)	Marketing *R*(*N*)	Finance *R*(*N*)	Quantitative *R*(*N*)	Organizational Behavior *R*(*N*)	Operations *R*(*N*)
VQAU	.550 (60)	—	.655 (76)	.629 (94)	.667 (63)	.625 (57)
TAU	.533 (61)	—	.582 (78)	.651 (94)	.660 (63)	.652 (57)

Note. Program concentrations with a sample size less than 50 were not reported. For some of the program concentrations, V+Q was a better predictor than V+Q+A+U or T+A+U. Most of the programs that participated did not have complete data on all predictors for all students; this resulted in reduced sample sizes for some analyses. It is likely that the data used for the V+Q+A+U and T+A+U combinations are sub-samples of the data used for single predictors and the V+Q combination. As such, the results from these analyses may not be directly comparable.

Tables 15 and 16 provide the regression analysis results for the five program concentrations that had more than 50 students for the combination of GMAT® scores, AWA, and UGPA. For all five of these program concentrations—accounting, finance, quantitative, organizational behavior, and operations—the combination of predictors accounted for at least 28% of the variability in FYGPA.

Table 15. Simultaneous Regression Analysis Predicting First-Year Grades from GMAT® Verbal, GMAT® Quantitative, AWA, and UGPA for Program Concentration Subgroups					
Variable	*B*	*SEB*	β	sr^2	R^2
Accounting					.303
V	–.023	.025	–.143	.013	
Q	.040	.027	.256	.033	
AWA	.293	.181	.246	.040	
UGPA	.573	.288	.258	.060	
Finance					.429*
V	–.020	.019	–.123	.010	
Q	.092	.027	.387*	.105	
AWA	–.209	.152	–.157	.018	
UGPA	.411	.227	.197	.031	
Quantitative					.396*
V	.045	.020	.270	.040	
Q	.065	.019	.369*	.091	
AWA	.110	.123	.089	.006	
UGPA	.795	.219	.350*	.100	
Organizational Behavior					.445*
V	.045	.024	.135	.009	
Q	.045	.027	.271	.032	
AWA	.554	.179	.432*	.108	
UGPA	.683	.322	.266	.051	

Table 15. Simultaneous Regression Analysis Predicting First-Year Grades from GMAT® Verbal, GMAT® Quantitative, AWA, and UGPA for Program Concentration Subgroups					
Variable	*B*	*SEB*	β	sr^2	R^2
Operations					.391
V	.035	.026	.204	.026	
Q	.044	.024	.295	.047	
AWA	.396	.178	.353	.070	
UGPA	–.060	.294	–.028	.001	

Note. Results for PL dummy codes are not presented because PL variance is treated as error.
SEB = standard error of *B*; sr^2 = squared semi-partial.
*p < .01.

Table 16. Simultaneous Regression Analysis Predicting First-Year Grades from GMAT® Total, AWA, and UGPA for Program Concentration Subgroups					
Variable	*B*	*SEB*	β	sr^2	R^2
Accounting					.284
Total	.002	.002	.115	.007	
AWA	.204	.163	.171	.023	
UGPA	.638	.287	.287	.073	
Finance					.339*
Total	.002	.002	.148	.014	
AWA	–.303	.154	–.227	.040	
UGPA	.434	.238	.208	.035	
Quantitative					.424*
Total	.009	.002	.543*	.143	
AWA	.081	.114	.066	.004	
UGPA	.819	.212	.360*	.106	
Organizational Behavior					.436*
Total	.004	.003	.246	.024	
AWA	.523	.163	.408*	.116	
UGPA	.673	.318	.262	.051	
Operations					.425*
Total	.007	.002	.484*	.102	
AWA	.361	.154	.322	.072	
UGPA	–.039	.271	–.018	.000	

Note. Results for dummy codes are not presented because PL variance is treated as error.
SEB = standard error of *B*; sr^2 = squared semi-partial.
*p < .01.

All subgroups. Using data from all students, a regression equation was built using the V, Q, AWA, UGPA combination and the combination including Total, AWA, and UGPA. Additionally, residuals were calculated for each combination as FYGPA minus the predicted value. Thus, positive average residuals indicate the group performed better than predicted, and negative values indicate they did not perform as well as predicted. The results of this analysis are presented in Table 17.

The magnitude of standardized residual values can be interpreted as effect size values (i.e., .2 = small; .5 = med; .8 = large) according to Cohen (1969). The results indicated that most subgroups performed as expected, with no absolute value exceeding .23.

Table 17. Average Standardized Residuals and Scaled Difference for First-Year Grades

	V+Q+A+U		T+A+U	
	Average Residual (*N*)	Scaled Difference	Average Residual (*N*)	Scaled Difference
Gender				
Female	.093 (222)	.074	.087 (223)	.069
Male	–.077 (295)	–.061	–.073 (295)	–.058
English Skills				
Acceptable	–.021 (387)	–.017	–.019 (388)	–.015
Questionable	.126 (70)	.100	.117 (70)	.093
Concentrations				
Accounting	–.038 (60)	–.030	–.046 (60)	–.036
Finance	–.088 (70)	–.070	–.093 (72)	–.074
Quantitative	.129 (94)	.102	.122 (94)	.097
Organizational Behavior	.228 (63)	.181	.217 (63)	.172
Operations	–.053 (56)	–.042	–.056 (56)	–.044

Positive = overpredicted; Negative = underpredicted.

The standardized residuals were rescaled to determine the difference that would be observed for a four-point grading scale using the average standard deviation of grades observed across the studies (SD = 0.793). On average, the absolute difference equated to approximately 0.074 on a 4-point scale, indicating that there is little practical difference in predictive validity by gender, English skill, or program concentration subgroup. An analysis of variance of the residuals performed separately for gender, English skill, and program concentration subgroups indicated that the omnibus tests were not significant at a $p = .01$ level for all analyses. Thus, a prediction equation based on all doctoral students could be used for applicants from the different subgroups identified in the current study.

Conclusion

The present study expanded on previous research in the area of test validity by exploring the validity of inferences made about first-year success in doctoral business programs based on both performance on an admission exam and performance during undergraduate education. These findings indicate that the GMAT® exam is a better predictor of FYGPA for doctoral students than was previously reported (Zwick, 1993). Additionally, it provides evidence that the GMAT® Total score and the combination of GMAT® V and GMAT® Q scores are better predictors of performance than UGPA. However, previous research (i.e., Zwick) did not correct for restriction of range or include the AWA as a predictor. Moreover, the current study examined predictive validity for three different doctoral student subgroups based on gender, English skill, and program concentration enrollment.

Though significance testing indicated that there were no statistically significant differences among estimates of predictive validity for the various subgroups, there was some variation in prediction for different groups. For instance, estimates were higher for females and students with questionable English skills. However, these findings should be validated with a larger sample size. Group membership overall did not result in consistent over- or under-prediction of performance for any particular groups.

It should be noted that there is self-selection bias in terms of the types of students who enter into doctoral programs in management education. These students are not likely to represent the general population, especially with respect to the subgroups examined in this study. Thus, it would not be appropriate to generalize these subgroup findings to the same subgroupings in the general population. Furthermore, a number of self-selection factors that could influence the prediction of performance were not included in these analyses and should be considered for future research.

Although predictive validity for doctoral programs was not as high as estimates revealed for full-time or executive MBA programs in previous research (Talento-Miller & Rudner, 2005), these results are not surprising. Doctoral programs often are more selective than full-time or executive MBA programs in terms of the types applicants admitted. Students frequently score at the upper-end of the GMAT® exam and UGPA scales and perform very well in their graduate-level courses. Typically, programs are more interested in predicting performance for applicants or potential students than for the students actually admitted or enrolled. In forming better predictions, the program hopes to minimize the numbers of enrolled students who drop out or perform poorly. This study revealed that for those students enrolled in doctoral programs, the GMAT® exam, used alone or in combination with UGPA, was an effective predictor of performance.

Future research should explore additional predictors of performance that may be gathered during the admission process, such as interview ratings or letters of recommendation. Alternative indicators of success, such as end-of-program grades and comprehensive exam or project performance should also be considered for further examination. Unfortunately, programs that participated in the current study did not include this information.

Contact Information

For questions or comments regarding study findings, methodology or data, please contact the GMAC® Research and Development department at research@gmac.com.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (1999). *Standards for Educational and Psychological Testing*. Washington, D. C.: American Educational Research Association.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

GMAC®. (1999). *Guide to the use of GMAT® scores.* [Brochure]. Princeton, NJ: Author.

Graham, L. D. (1991). Predicting academic success of students in a master of business administration program. *Educational and Psychological Measurement*, 51, 721–727.

Hecht, L. W. & Schrader, W. B. (1986, June). *Graduate Management Admission Test: Technical Report on Test Development and Score Interpretation for GMAT® User*. McLean, VA: Graduate Management Admission Council®.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.* Newbury Park, CA: Sage Publications.

Kaplan, R., & Sacuzzo, D. (1997). *Psychological Testing: Principles, Applications, and Issues* (4th ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

Kuncel, N. R., Crede´, M., & Thomas, L. T. (2004, April). *A Comprehensive Meta-Analysis of the Predictive Validity of the Graduate Management Admission Test (GMAT®) and Undergraduate Grade Point Average (UGPA).* Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Chicago, IL.

Olsen, M. (1957, January). *The Admission Test for Graduate Study in Business as a predictor of first year grades in business schools, 1954–1955.* Princeton, NJ: Educational Testing Service®.

Stolzenberg, R. & Relles, D. (1985). *Calculation and practical application of GMAT® predictive validity measures. GMAC® Occasional Papers.* McLean, VA: Graduate Management Admission Council®.

Talento-Miller, E. & Rudner, L. M. (2005, October). *GMAT® validity study summary report for 1997 to 2004.* McLean, VA: Graduate Management Admission Council®.

Talento-Miller, E., Rudner, L. M., & Owens, K. (2006). *Assessing Validity by Combining Examinees Across Programs* (GMAC® Research Report RR-06-14). McLean, VA: Graduate Management Admission Council®.

Zwick, R. (1993). The validity of the GMAT® for the prediction of grades in doctoral studying business and management: An empirical Bayes approach. *Journal of Educational Statistics*, 18(1), 91–107.

Appendix

Table A1. Variables Identified

Variable(s)	Description
FYGPA	First year average
V	GMAT® Verbal score
Q	GMAT® Quantitative score
A	GMAT® Analytical Writing Assessment score
U	Undergraduate GPA
VQ	GMAT® Verbal score & GMAT® Quantitative score
T	GMAT® Total score
VQAU	GMAT® Verbal score, GMAT® Quantitative score, GMAT® Analytical Writing Assessment score, & Undergraduate GPA
TAU	GMAT® Total score, GMAT® Analytical Writing Assessment score, & Undergraduate GPA

Table A2. First-Year Predictive Validity by Program									
	N	V	Q	A	U	VQ	T	VQAU	TAU
A	50	–.025	.382	.045	.278	.406	.353	.475	.426
B	53	.087	–.173	.309	.247	.193	–.031	.427	.410
C	31	–.101	–.021	–.353	.211	.101	–.096	.441	.436
D	42	.475	.235	.130	.157	.479	.588	.493	.596
E	50	.071	.158	–.004	.197	.164	.238	.242	.302
F	47	.071	.154	.063	.312	.160	.327	.329	.415
G	24	.378	.751	–.026	.614	.759	.700	.940	.920
H	85	.159	.425		.247	.427	.307	.453	.354
I	100	.228	–.077	.225	.318	.269	.202	.430	.371
J	48	.028	–.095	.205	.243	.109	.028	.336	.300
K	45	.175	.134	.279		.194	.498	.292	.505
L	19	–.132	.497	.341	.249	.560	.509	.735	.560
M	97	.161	.025	.058	.031	.162	.338	.165	.341
N	86	.319	–.404	–.026	.238	.643	.345	.745	.434
O	80	.121	.450	.385	.230	.454	.445	.599	.524
P	60	–.043	–.216	.221	.215	.218	–.169	.432	.430
Q	39	.414	.266	.006	.226	.430	.509	.507	.572
R	50	–.004	.715		.353	.761	.427	.784	.498

Note. The interaction between outliers and the correction for restriction of range likely resulted in some of the negative correlations revealed for some programs.

Table A3. Second-Year Predictive Validity by Program									
	N	V	Q	A	U	VQ	T	VQAU	TAU
A	46	–.257	.587	–.062	.223	.732	.414	.760	.505
G	16	.513	.744	.076	.675	.787	.813	.977	.889
H	85	.162	.528		.428	.528	.377	.616	.512
J	48	.173	–.063	.245	.158	.204	.207	.298	.287
K	24	.534	–.249	.521		.697	.672	.754	.721
L	13	–.313	.890	–.026	.240	.891	.806	1.00	.880
M	90	.171	.047	.049	.130	.171	.346	.215	.370
N	53	.131	–.409	–.475	.334	.508	.072	.916	.659
O	65	.079	.398	.184	.262	.399	.384	.470	.427
Q	36	.193	–.063	.400	–.087	.244	.058	.452	.434
R	50	.331	.087		–.450	.332	.310	.344	.334

Note. The interaction between outliers and the correction for restriction of range likely resulted in some of the negative correlations revealed for some programs.