Graduate Management Admission Council®

Differential Validity and Differential Prediction of the GMAT® Exam

Eileen Talento-Miller

GMAC® Research Reports • RR-17-01 • April 11, 2017

Because group differences are often observed on scores of standardized tests, the fairness of using scores for admissions selection is often questioned. Yet, simple group differences do not present a complete picture for the use of test scores. Validity studies for different groups of test takers provide evidence for the appropriate interpretation of test scores to promote responsible and fair use of scores (AERA, APA, NCME, 2014; Cleary, 1968; Halpern, 2000; Hecht, Manning, Swinton, & Braun, 1989; Kuncel & Hezlett, 2007). These studies demonstrate whether the differences in test scores represent differences in ability as evidenced by later performance.

Over time, the changing demographics of the applicant pool for higher education necessitate continual evaluation of admissions factors, notably standardized test scores. As an example, within the past decade, the mix of candidates taking the Graduate Management Admission Test® (GMAT®) exam has shifted: Although the test was developed in the United States, the majority of examinees since 2008 has been non-U.S. citizens. The GMAT exam is used for admission to graduate management programs in hundreds of countries around the world. As such, studies of differential validity and prediction for GMAT scores that reflect the diversity of the examinee pool and the graduate programs should be conducted.

Several prior studies and meta-analyses have examined the validity of GMAT scores for predicting grades in graduate business programs for all students (Kuncel, Crede, & Thomas, 2007; Qian, Trang, & Kingston, 2016; Oh, Schmidt, Shaffer, & Le, 2008; Talento-Miller & Rudner, 2008) and for specific groups of students (Crooks & Heuvelmans, 1999; Hecht, et al., 1989; Sireci & Talento-Miller, 2006; Talento-Miller, 2008). The differential validity and differential prediction studies conducted to date tended to have a limited focus, such as race/ethnicity subgroups within U.S. programs (Sireci & Talento-Miller, 2006), or citizenship groupings within programs outside the United States. (Crooks & Heuvelmans, 1999; Talento-Miller, 2008). Sample-size limitations often preclude the examination of differential validity and differential prediction within a single program's validity study. In addition, inconsistency in data collection can make it difficult to combine data across programs for group analysis.

By collecting data from different schools and programs and matching the school-provided information with the GMAT exam database to obtain consistent demographic information for the students, several groups can be formed with robust sample sizes.

The current study looks at the question of differential validity and differential prediction of GMAT scores in addition to the other major factor in admissions—undergraduate grade point average (UGPA)—relative to grades in graduate business programs.

The GMAT exam is used as part of the admissions process for thousands of graduate business programs worldwide. The test was launched in 1954, but has undergone many changes throughout the decades since its introduction. Notably, over time, the GMAT exam has expanded to include two sections beyond the original Verbal and Quantitative (Quant) reasoning sections. The Analytical Writing Assessment (AWA) was added in 1994, and the Integrated Reasoning (IR) section was added in 2012. The Total Score, to be consistent with its original design, still includes performance from only the Verbal and Quant sections. Each of the five scores—Total, Verbal, Quant, AWA, and IR —was evaluated for differential validity. Differential prediction analyses included the four section scores along with UGPA to determine how well grades were predicted for each examinee group.

Related Literature

Gender and race / ethnicity. Numerous studies of differential validity and differential prediction for higher education admission tests have examined gender and racial or ethnic groups. For instance, the study by Young and Kobrin (2001) reviewed decades of research on the two primary undergraduate admission tests used in the United States, namely the SAT® exam and the ACT® exam. Findings were relatively consistent over time. Regarding gender, validity coefficients tended to be higher for female examinees, but prediction equations tended to underpredict grades. In other words, although the test scores were more strongly related to grades, when a common prediction is used, female students tended to have slightly higher grades than would be predicted. For graduate school admissions, there appears to be no underprediction for female students (Kuncel & Hezlett, 2007). A study conducted by Wright and Bachrach (2003) that looked at the GMAT exam suggested there might be underprediction of grades for females with high GMAT scores, but the study did not use a regression approach. Other studies based on data combined across schools suggested either no difference in grade prediction by gender (Sireci & Talento-Miller, 2006; Talento-Miller, 2009) or overprediction of grades (Talento-Miller, 2008).

For the race and ethnicity comparisons conducted for undergraduate admission tests administered in the United States, the validity coefficients for underrepresented U.S. minority groups, including African American and Hispanic American students, tended to be lower compared with White students, and the common prediction equation tended to overpredict their grade point average (Young & Kobrin, 2001). Instead of disadvantaging these groups, the common equation suggested that their grades would be higher than what they would subsequently achieve. For the GMAT exam, study results were similar to those seen for the undergraduate tests, showing overprediction for the African American and Hispanic American groups (Sireci & Talento-Miller, 2006).

Language and citizenship. The undergraduate admission tests reviewed in Young & Kobrin (2001) and the graduate admission tests (except for the GMAT exam) reviewed in Kuncel & Hezlett (2007) are used primarily in the United States. As such, language and citizenship differences among examinees are limited. A study of differential validity and prediction of the SAT exam reported that 93% of the participants indicated English was their best language (Mattern, Patterson, Shaw, Kobrin, & Barbuti, 2008). This study showed lower validity and underprediction for the examinee group whose best language was not English.

Unlike the study for the SAT, studies for the GMAT exam suggested similar validity coefficients for native English speakers and native speakers of other languages (Crooks & Heuvelmans, 1999; Talento-Miller, 2008). When looking at differential prediction, Talento-Miller (2008) showed that results for the two non-native English groups differed for language groups: For the Western European language group, grades were underpredicted, but for the language group that included all the other non-English languages, grades were overpredicted.

Other than language differences, there may be cultural differences among students who sit for the GMAT exam that affect performance. The study by Stolzenberg and Relles (1991) showed that both English fluency and country of origin affected the prediction of foreign students' performance in U.S. graduate business programs, with small effects on prediction of graduate grades both by language and country of origin. Though not specifically addressing differential prediction, one finding suggested underprediction of grades for graduate business school students whose undergraduate study was done outside of the United States. Another study suggested overprediction of grades for non-U.S. citizens enrolled in U.S. graduate business programs (Talento-Miller, 2009). Koys (2005) showed the usefulness of GMAT scores and UGPA for predicting performance in graduate business programs in three different countries outside the United States, but did not specify whether the students were local citizens. Two meta-analyses of graduate business programs outside the United States looked at validity values by nationality groups. In the Crooks and Heuvelmans (1999) study, nationality groups were classified into five world regions that showed differences in GMAT score validity values. For example, the highest validity values were observed for the Asia-Pacific and Western European nationalities for GMAT Total Score, Central Eastern Europe and Africa and Middle East for Quant scores, and Asia-Pacific for Verbal scores. The correlation of GMAT Quant score with grades was near zero for the Asia-Pacific nationalities group. Although this might be hypothesized to be a restriction of range issue, if many of the Asia-Pacific group had extremely high Quant scores, the information could not be found in the study because the descriptive values were not reported by group. Interestingly, in the Talento-Miller (2008) study, the lowest multiple correlation was observed for the Americas nationality group, which included students from North America, Central America, and South America. It may be that differences in grades for the Americas nationality group would be better explained by UGPA, which was not included in the study. Differential prediction results showed similar findings to the language results, with underprediction for Western European nationalities and overprediction for Asia-Pacific nationalities.

Age. Because some graduate management programs require prior work experience as a condition for admission, there may be more drastic differences in the age of students within these programs compared with other types of undergraduate or graduate programs. Some programs, such as a Master of Accounting, may draw an applicant pool from among students coming directly out of undergraduate courses, whereas programs such as the executive MBA may require up to 10 years of prior work experience from potential applicants. As with gender and race/ethnicity comparisons, differences in test scores by age may prompt questions of fairness, as scores tend to be lower for older candidates. Using UGPA as an admissions factor may also be problematic if the applicant is many years out from the grades earned on the undergraduate transcript (Talento-Miller, Guo, & Siegert, 2008). Studies looking at program differences suggest that non-MBA programs, which tend to draw younger students, show generally higher validity values across UGPA and GMAT scores (Talento-Miller, 2009). Comparisons of validity values for executive MBA programs with other programs suggest lower values for UGPA, but higher values for GMAT scores (Siegert, 2008; Talento-Miller & Rudner, 2008).

Although suggestive, these program differences do not directly indicate whether differential validity or differential prediction exists by age. The study by Crooks & Heuvelmans (1999), which examined four age groupings, suggested validity values for GMAT scores decreased as age increased. In fact, the value for Quant scores at the highest age group (35+) was effectively zero, but compared with the other categories, the sample size for this group was limited with fewer than 100 observations. The study by Hecht et al. (1989) suggested that the relationship of GMAT scores with graduate grades is stronger for older students compared with younger students,

though the opposite was true for UGPA. There appears to be underprediction for the older students, which was exacerbated when UGPA was introduced to the model, but the residuals were based on a prediction model that was fitted to the younger students.

Differential validity and differential prediction studies appeared to be consistent for some groups, but not so for others. Results by gender for graduate programs in general, and for GMAT scores in particular, suggest that female applicants are not disadvantaged by the admissions factors. Race/ethnicity findings for graduate programs mirror those for undergraduate programs showing overprediction for underrepresented minorities, again suggesting no disadvantage for these groups. Results for language and citizenship show differences depending on how the groups are defined. Less information is available for age comparisons, but the combination of studies of programs and studies of age suggest there may be differences in validity or prediction.

Prior research addresses differential validity and differential prediction of GMAT scores for different groups but it is limited. For instance, meta-analyses that examined citizenship were based on programs located exclusively within the United States or exclusively outside the United States. Findings from program types used as a proxy for age groups contradict some of the findings from the age studies, such as the usefulness of GMAT Quant scores. Some of the studies did not include AWA because of the availability of the scores, or UGPA because of the differences in the variable across countries. None of the studies included IR, since all studies predate the introduction of the IR section to the GMAT exam. Therefore, an updated comprehensive review of differential validity and differential prediction of GMAT scores and UGPA for graduate business programs is needed.

Methodology

Data Collection

Data collected from different business schools for individual validity studies were combined to enable group analyses with robust sample sizes. Schools either were invited to participate in a special study to collect data on IR scores, or they volunteered the data on their own. Two options were provided for data submission: (a) the school submitted their students' personal information, which was then matched to the GMAT test database, or (b) the school submitted anonymized data. Data pulled from the GMAT database included each of the GMAT scores and background information that examinees provided. Demographic information provided by examinees included gender, citizenship, native language, and date of birth. Academic information included undergraduate major, highest degree received, and self-reported undergraduate GPA (on a 4.0 scale). For schools that provided anonymized data, some included demographic and academic information, but the format and categories may have differed from those in the GMAT database. Where possible, the school-provided data were converted to comparable categories.

For each business school program, an individualized study was created and a report submitted. Many programs, however, did not have enough cases in different groups to allow for robust analyses. Combining data across programs allows for analyses of groups that would not be possible at the individual program level. But differences among programs, including program type, school location, grading scales, course expectations, and students' prior experience, could lead to difficulties in interpreting results. Therefore, admissions and program grade information was standardized within each school before combining data. Because each school's standard deviation was forced to be one for each variable, no adjustments for restriction of range were attempted on the data.

Variables

Dependent Variable

Graduate GPA (GGPA). Graduate grades were standardized within programs including all cases provided, regardless of whether other data for that student, such as GMAT scores, were available. Some schools provided cases with very low averages, including a few cases with a GGPA of 0.00. Cases such as these may have withdrawn from courses for reasons other than academic ability. Therefore, the distribution of standardized GGPA was examined. Outliers, defined as cases greater than three standard deviations from the mean, were removed from further analyses.

Independent Variables

GMAT scores. GMAT scaled scores formed the basis for the analyses. For data that came directly from the GMAT database, some examinees had multiple scores from taking the exam more than once. To replicate the process used for admission to many programs, the highest score from each section was used as opposed to using a set of scores from a single examination instance.

Undergraduate GPA (UGPA). Data from the GMAT database included self-reported GPA on a 4.0 scale. If school-reported UGPA was available, those data were used in place of self-reported data, unless the value was not on a 4.0 scale. If no additional information was available to convert the UGPA to the common scale, the data were set to missing.

Grouping Variables

Student data. Information available for students included gender, citizenship, native language, and date of birth. Race and ethnicity information was available only for U.S. citizens. Although some individual categories for citizenship and language had sufficient cases for analyses, data were grouped to facilitate comparisons. Standardized GPA and GMAT exam variables were compared across and within groupings to ensure the appropriateness of combining data. The goal was to have a minimum of 75 complete cases in each group—equivalent to 15 times the number of variables in the multiple regression.

Four categories were created for the race and ethnicity comparison: White (non-Hispanic), Asian American, African American, and Other. The Other race and ethnicity category included designations such as Mexican American, Puerto Rican, American Indian, Multiethnic, Multiracial, and others. Because the GMAT exam is administered only in English and is recommended for programs taught in English, only two groups were created for native language. Age was calculated roughly from year of birth to the year the data were submitted for analysis. For the age variable, groups were defined roughly to correspond to different work experience requirements. For instance, the youngest group would likely be in programs that require little to no work experience, such as Master of Accounting programs; whereas the oldest group might attend programs with extensive work experience required, such as executive MBA programs.

Analyses

Differential Validity. Correlation and regression analyses were computed to determine differential validity. For each of the admissions variables, the relationship with GGPA was calculated using Pearson correlations. No adjustments were made for restriction of range or attenuation for any of the dependent or independent variables. Results for the full data set were compared with previous meta-analytic research reports of unadjusted correlations. Correlations were calculated by group and compared. Regression analyses compared the predictive validity of the combined admissions variables across groups. GMAT Total Score, which represents the combined performance from the Verbal and Quant sections, was not included in the regression analyses.

Differential Prediction. Analyses for differential prediction involved determining a single prediction equation for all students and then looking at differences in average residuals among groups to see whether a particular group might be disadvantaged. Because residuals are calculated as observed GPA minus predicted GPA, positive values indicate underprediction, meaning the group received higher grades on average than what was expected based on their admissions variables. For the same reason, negative residuals indicate overprediction, suggesting the group received lower grades than expected based on their admissions variables.

An important factor to consider when interpreting residual analyses is that differences in variables among groups will influence over- or underprediction. That is, if a group has a lower GPA on average, then regression to the mean suggests that the group's predicted values will be higher than their observed values, leading to a pattern of overprediction.

Results

For the current study, 28 graduate business programs provided data for validity studies that included enough information to classify students into groups. Of these, 22 (79%) were from programs located in the United States. Twenty MBA programs (71% of total) were represented in the data set, one of which was an executive MBA program. Of the remaining data sets, four (14%) were from other business master's programs, and four (14%) did not specify program type. **Table 1** shows the descriptive admissions information across programs and **Table 2** shows descriptive information for the groups of students included in the study. The *Appendix* lists the individual categories within regional citizenship groups. Compared with the demographics of all GMAT examinees in the 2013 *Profile of Graduate Management Admission Test Candidates* (GMAC, 2013), the students within these combined programs were older, less likely to be female, and were more likely to be from the United States. Among U.S. citizens, more student data in the study fell into the Asian American race/ethnicity category compared with the GMAT examinee database.

Table 1. Program Descriptive Data

	K	Minimum	Maximum	Median	Combined Data	GMAT Examinees 2013–2015
N		27	740	191	5968	757,035
Mean Total Score	28	529.85	729.82	641.25	659.39 (86.13)	551.94 (120.88)
Mean Verbal	28	25.93	42.44	33.30	36.52 (6.99)	26.80 (9.21)
Mean Quant	28	34.30	48.87	44.31	43.97 (6.66)	38.91 (10.87)
Mean AWA	28	4.36	5.56	5.02	5.18 (0.72)	4.37 (1.19)
Mean IR	28	4.07	6.87	5.44	5.76 (1.90)	4.23 (2.15)
Mean UGPA	27	3.11	3.70	3.36	3.42 (0.40)	Not reported

Table 2. Group Descriptive Data

Group	K	%	Min N	Max N	Total	Verbal	Quant	AWA	IR	UGPA
Gender										
Female	27	32.6%	1210	1694	643.84 (86.72)	35.35 (6.91)	42.93 (7.40)	5.14 (0.74)	5.50 (1.89)	3.47 (0.36)
Male	27	67.4%	2112	3499	652.71 (85.93)	35.89 (6.93)	43.75 (6.59)	5.11 (0.72)	5.84 (1.90)	3.35 (0.40)
Race/Ethnicity										
White (non-Hispanic)	20	70.2%	979	1681	661.26 (75.10)	38.39 (5.28)	42.56 (6.22)	5.38 (0.60)	6.27 (1.69)	3.42 (0.36)
Asian American	20	14.0%	179	336	663.60 (81.44)	36.99 (6.50)	44.34 (5.79)	5.44 (0.60)	6.02 (1.96)	3.39 (0.35)
African American	20	5.1%	97	121	608.43 (90.34)	35.25 (5.83)	39.17 (8.11)	5.28 (0.63)	5.43 (1.78)	3.23 (0.37)
Other	20	10.7%	150	257	647.86 (77.35)	37.47 (6.06)	41.94 (6.18)	5.35 (0.61)	5.88 (1.77)	3.35 (0.34)
Language										
English	21	59.8%	1398	2343	661.50 (76.90)	38.15 (5.56)	42.91 (6.25)	5.40 (0.60)	6.22 (1.70)	3.40 (0.38)
Other	21	40.2%	953	1578	647.06 (93.25)	32.96 (7.39)	45.65 (6.51)	4.82 (0.70)	5.49 (1.88)	3.40 (0.44)
Citizenship										
UCAP	26	63.9%	1949	3490	660.64 (83.35)	38.25 (6.01)	42.64 (6.65)	5.40 (0.62)	5.98 (1.84)	3.43 (0.37)
ESA	26	14.1%	483	769	665.11 (80.74)	32.67 (7.04)	47.88 (4.79)	4.78 (0.68)	5.69 (1.74)	3.44 (0.39)
CSA	26	9.0%	225	489	661.25 (87.02)	34.13 (7.48)	46.44 (5.01)	5.02 (0.67)	5.44 (1.82)	3.32 (0.52)
WE	26	5.0%	142	271	630.52 (113.62)	34.75 (8.55)	42.00 (8.38)	4.88 (0.81)	5.35 (2.17)	3.46 (0.52)
MCLA	26	4.3%	150	234	638.03 (98.52)	34.50 (7.62)	43.43 (7.09)	4.67 (0.80)	5.15 (1.95)	3.34 (0.47)
AEEME	26	3.8%	116	206	622.09 (112.79)	32.82 (8.34)	42.75 (8.04)	4.71 (0.77)	4.78 (2.30)	3.43 (0.40)
Age										
≤ 25	22	17.6%	599	745	664.19 (69.63)	35.78 (6.00)	45.05 (6.14)	5.14 (0.68)	6.16 (1.68)	3.54 (0.34)
26–30	22	55.4%	1434	2370	681.05 (64.94)	38.44 (5.26)	45.04 (5.26)	5.35 (0.62)	6.23 (1.65)	3.41 (0.36)
≥ 31	22	27.0%	460	1156	637.54 (90.58)	34.62 (7.16)	42.97 (7.25)	5.00 (0.70)	5.65 (1.84)	3.35 (0.42)

Note: UCAP = United States, Canada, Australia and Pacific Islands; ESA = East and Southeast Asia; CSA = Central and South Asia; WE = Western Europe; MCLA = Mexico, Caribbean, and Latin America; AEEME = Africa, Eastern Europe, and Middle East.

Group differences were observed across many of the admissions variables. **Table 3** provides the summary of the standardized variables from the complete cases used in the regression analyses. Because the variables are standardized, effect sizes can be compared to get a sense of group differences. For the gender groups, graduate GPA (GGPA) and most GMAT scores tended to be higher for the male students, with UGPA and AWA higher for the female students, but effect sizes were small (D = 0.10–0.30).

For the race/ethnicity comparisons, the White (non-Hispanic) group had the highest averages on all variables except for Quant and AWA, which were highest for the Asian American group. Effect sizes were largest for the White (non-Hispanic) versus African American comparison (D = 0.14–1.64).

For the language groups, the most notable differences were seen with the Verbal and AWA scores favoring native English speakers (D = 0.56 and 0.55, respectively) and the Quant scores favoring other native language speakers (D = 0.73). When examining citizenship groups, the Western European (WE) world region had the highest graduate and undergraduate grade averages, in addition to the highest Verbal and IR scores. The highest Quant average was observed for the East and Southeast Asia (ESA) group, followed closely by the Central and South Asia (CSA) group, which had the highest average for Total Score. The United States, Canada, Australia and Pacific Islands (UCAP) group had the highest average for AWA. For the age categories, the two younger groups had the advantage for the comparisons, but effect sizes were small (D = 0.14–0.34).

Table 4 displays the summary of the simple correlations for the combined data. The comparisons show that the data standardized within schools produce results similar to the weighted mean and median values of the correlations produced by the school-level analyses. Consistent with previous research, validity values for GMAT Total Scores tend to be higher than those for UGPA, whereas the separate section scores from which Total is based—Verbal and Quant—tend to have correlations similar to those for UGPA. Interestingly, validity values for the IR section, the relatively recent addition to the GMAT exam, compare favorably with those of the Verbal and Quant sections, even though the IR section is much shorter. The AWA validity value is considerably lower than the other variables, but the variability shows that the scores may be a valuable predictor for some programs.

It is important to note that both the AWA and the Quant predictors observed negative correlations for some of the individual program results. The notion that better test performance leads to worse program performance is anomalous and typically results from a data set with severe range restriction and/or small sample sizes. Peculiarities such as these in data sets may explain some unusual findings across groups that are not necessarily related to fairness of scores.

Current results can be compared with previous meta-analyses to determine whether changes over time or changes to the test may be affecting validity. Unadjusted values from the Talento-Miller and Rudner (2008) study were used as the basis for the comparisons. The current validity values for UGPA, Total Score, and Verbal scores are higher than in the previous study, whereas the values for the Quant scores are nearly identical to what was observed in the previous meta-analysis.

Table 3. Listwise Standardized Mean by Group

Group	N	GGPA	Total	Verbal	Quant	AWA	IR	UGPA
All	2966	-0.03	-0.13	-0.02	-0.14	0.03	0.02	-0.03
Gender								
Female	1029	-0.15	-0.31	-0.11	-0.33	0.09	-0.14	0.14
Male	1848	0.05	-0.03	0.03	-0.03	-0.01	0.11	-0.12
Race/Ethnicity								
White (non-Hispanic)	936	0.15	-0.07	0.27	-0.31	0.23	0.15	0.03
Asian American	166	-0.20	-0.25	-0.16	-0.11	0.32	-0.06	-0.14
African American	89	-0.92	-1.71	-0.76	-1.52	0.09	-0.53	-0.67
Other	136	-0.33	-0.57	0.04	-0.73	0.18	-0.06	-0.22
Language								
English	1315	0.01	-0.21	0.17	-0.39	0.24	0.10	-0.05
Other	809	-0.09	-0.02	-0.39	0.34	-0.31	-0.08	0.00
Citizenship								
UCAP	1843	-0.03	-0.22	0.13	-0.38	0.21	0.07	-0.07
ESA	406	-0.16	0.16	-0.46	0.59	-0.38	-0.04	0.01
CSA	169	-0.05	0.30	-0.13	0.56	0.06	0.02	-0.06
WE	105	0.45	0.03	0.32	-0.21	0.08	0.14	0.33
MCLA	126	0.06	-0.43	-0.45	-0.08	-0.74	-0.18	-0.05
AEEME	88	-0.02	-0.39	-0.51	-0.09	-0.52	-0.28	0.12
Age								
≤ 25	552	0.10	0.01	0.02	0.00	0.06	0.11	0.16
26–30	1296	-0.06	-0.18	-0.03	-0.19	0.08	0.00	-0.08
≥ 31	385	-0.08	-0.17	-0.12	-0.10	-0.19	-0.05	-0.18

Note: UCAP = United States, Canada, Australia and Pacific Islands; ESA = East and Southeast Asia; CSA = Central and South Asia; WE = Western Europe; MCLA = Mexico, Caribbean, and Latin America; AEEME = Africa, Eastern Europe, and Middle East.

Table 4. Simple Correlations

	K	N	Min	Max	Median	Weighted Mean (SD)	Standardized Within School	2008 Meta-Analysis
Total	28	5968	0.03	0.68	0.37	0.38 (0.10)	0.38	0.34
Verbal	28	5945	0.04	0.76	0.30	0.31 (0.10)	0.31	0.26
Quant	28	5945	-0.28	0.48	0.23	0.24 (0.13)	0.25	0.25
AWA	28	5893	-0.10	0.28	0.13	0.13 (0.09)	0.13	0.17
IR	28	3476	0.05	0.51	0.26	0.27 (0.09)	0.27	—
UGPA	27	5004	0.11	0.50	0.31	0.32 (0.08)	0.32	0.25

Using the combined data after standardizing within school, simple correlations were calculated for each of the admissions variables by group, and the results are presented in **Table 5**. Using Fisher's z transformation, simple correlations were compared for each focal group versus the reference group for that category. The reference groups were the male, White, English-speaking, UCAP, and ≤ 25 groups for the gender, race/ethnicity, language, citizenship, and age categories, respectively. Because statistical significance is a function of sample size and the sample sizes differ considerably by group, the results of significance tests thus should be interpreted with caution.

Generally, the values show predictive validity evidence across groups for the admissions variables. The data suggested no differential validity by gender. There were two statistically significant results for the race/ethnicity comparisons, with the Asian American group showing higher validity for IR scores, and the African American group showing lower validity for UGPA, when each was compared with the respective values for the White group.

For the language category, significant differences existed across all the variables except for Verbal scores. Although some of the previous research showed lower validity across the admissions variables for those whose best language was not English (Mattern et al., 2008), the results for the current study were mixed. When comparing the validity values for the two language groups, Total, Quant, and UGPA had lower values and AWA and IR had higher validity values for the Other language group.

The citizenship comparisons seemed to augment the language category findings. The finding of lower validity values for Total and Quant, which was observed for the Other language group, was also observed for the ESA group compared with the UCAP group, but was not observed for any other citizenship focal group. Three of the citizenship focal groups—CSA, Mexico, Caribbean, and Latin America (MCLA), and Africa, Eastern Europe, and Middle East (AEEME)— showed lower validity values for UGPA compared with the UCAP reference group. Within the citizenship category, validity values for IR were highest for the MCLA group. The findings for age seemed to be consistent with previous research (Hecht et al., 1989)—for the older groups, validity values were higher for Total and Quant but lower for UGPA.

Differences in validity values among groups suggest there may be differential validity, though the results do not necessarily suggest a difference in the fairness of the use of scores. Compared with the reference groups, many of the focal group comparisons showed higher validity values. Among the lower validity value findings, some of the group differences may be explained by looking closely at the data. For instance, one of the lowest simple correlations is the Quant score for the ESA citizenship group. A look at the summary data shows that this group had the highest Quant average with one of the lowest standard deviations. Data showed that more than two-thirds of the group were within a standard error of the maximum score.

Correlations are a measure of variance explained, but one cannot measure differences in outcomes when there are no differences in the predictor. Another low correlation was observed with the UGPA predictor for the African American race/ethnicity group. Although data suggested a low average UGPA for this group, the largest concern in interpreting this finding was the small sample size available for this group.

Table 5. Simple Correlations by Group

Group	Total	Verbal	Quant	AWA	IR	UGPA
All	0.38	0.30	0.25	0.13	0.27	0.32
Gender						
Female	0.36	0.29	0.23	0.10	0.26	0.33
Male	0.38	0.31	0.24	0.13	0.26	0.32
Race/Ethnicity						
White (non-Hispanic)	0.38	0.25	0.34	0.05	0.18	0.32
Asian American	0.44	0.33	0.32	0.08	0.33*	0.40
African American	0.51	0.41	0.36	0.05	0.18	0.08*
Other	0.38	0.21	0.33	0.03	0.10	0.29
Language						
English	0.44	0.32	0.36	0.04	0.22	0.35
Other	0.31**	0.29	0.19**	0.15**	0.29*	0.24**
Citizenship						
UCAP	0.43	0.32	0.34	0.10	0.26	0.36
ESA	0.25**	0.29	0.04**	0.11	0.26	0.30
CSA	0.37	0.30	0.27	0.15	0.20	0.21*
WE	0.48	0.35	0.41	0.20	0.34	0.31
MCLA	0.46	0.28	0.40	0.06	0.41*	0.19*
AEEME	0.41	0.28	0.31	0.20	0.27	0.13*
Age						
≤ 25	0.30	0.30	0.15	0.10	0.24	0.41
26–30	0.37	0.27	0.25*	0.08	0.26	0.29**
≥ 31	0.42*	0.36	0.26*	0.17	0.26	0.25**

*$p < 0.05$, **$p < 0.01$.
Note: UCAP = United States, Canada, Australia and Pacific Islands; ESA = East and Southeast Asia; CSA = Central and South Asia; WE = Western Europe; MCLA = Mexico, Caribbean, and Latin America; AEEME = Africa, Eastern Europe, and Middle East.

Analyzing differential prediction allows examination of the interplay of admissions factors. Multiple regression was conducted for the full data set and for each of the groups using UGPA, Verbal, Quant, AWA and IR as predictors. Results, summarized in **Table 6,** again provided evidence for predictive validity across groups. The lowest multiple correlation was observed for the ESA citizenship group, which was not surprising based on the previous examination of the range-restricted Quant scores. Regression equations for each group were compared with the full sample equation by testing the intercepts and regression coefficients for statistically significant differences. Several groups showed differences in intercepts versus the full sample. **Table 3**, which displayed the standardized mean values for each group, showed that most of the intercept differences were consistent with initial group differences in GGPA.

Table 6. Regression Results by Group

Group	N	R	Intercept	Verbal	Quant	AWA	IR	UGPA	Std Res
All	2966	0.51	0.01	0.20	0.20	0.01	0.12	0.27	0.00
Gender									
Female	1029	0.50	-0.09**	0.17	0.18	0.01	0.13	0.28	-0.11
Male	1848	0.51	0.07*	0.22	0.21	0.02	0.11	0.28	0.07
Race/Ethnicity									
White (non-Hispanic)	936	0.51	0.20**	0.15	0.30**	-0.03	0.05*	0.32	0.14
Asian American	166	0.62*	-0.08	0.15	0.31	0.00	0.13	0.42*	-0.14
African American	89	0.49	-0.54**	0.25	0.12	0.01	0.06	-0.04**	-0.27
Other	136	0.44	-0.11	0.10	0.21	-0.01	-0.03	0.30	-0.16
Language									
English	1315	0.56*	0.10**	0.19	0.27*	-0.04	0.07	0.32	0.05
Other	809	0.43**	-0.03	0.18	0.12*	0.03	0.16	0.21	-0.09
Citizenship									
UCAP	1843	0.56*	0.07*	0.17	0.27**	-0.01	0.08	0.32*	0.02
ESA	406	0.42*	-0.07	0.19	0.02**	0.02	0.18	0.25	-0.23
CSA	169	0.45	-0.21*	0.11	0.33	0.04	0.13	0.25	-0.16
WE	105	0.55	0.36**	0.22	0.26	0.04	0.15	0.17	0.38
MCLA	126	0.52	0.20*	0.15	0.30	0.01	0.24	0.10**	0.24
AEEME	88	0.39	0.15	0.19	0.18	0.08	0.08	0.08*	0.12
Age									
≤ 25	552	0.54	0.02	0.21	0.11**	0.03	0.10	0.38	0.03
26-30	1296	0.49	0.02	0.18	0.24	0.00	0.12	0.26	0.00
≥ 31	385	0.47	0.01	0.26	0.18	-0.02	0.15	0.21	0.01

*$p < 0.05$, **$p < 0.01$.

Note: UCAP = United States, Canada, Australia and Pacific Islands; ESA = East and Southeast Asia; CSA = Central and South Asia; WE = Western Europe; MCLA = Mexico, Caribbean, and Latin America; AEEME = Africa, Eastern Europe, and Middle East.

Differences in coefficients also were relatively consistent with differences in simple correlations. The biggest differences appeared to be among the citizenship category. For instance, the CSA group had a low intercept relative to the group's average GGPA. The two Asian groups had similar Quant scores, but the impact of Quant on prediction differed. Examining the data within groups in more detail showed some background differences, particularly in undergraduate major: More than two-thirds of the students in the CSA group majored in Engineering, whereas most of the ESA group majored in Business and related fields. This informal review of the data suggested factors missing from the analyses may be affecting the prediction.

Another way to assess differential prediction is by using a single prediction equation and looking at differences in residuals across groups. Results are found in the last column of **Table 6**. In interpreting the findings, it is again imperative to consider the average values for each group. For instance, a group that has lower than average grades and/or predictors, as shown in **Table 3**, will be more likely to show overprediction: Negative residuals show that the predicted grades are higher than the observed grades. The largest average residuals were about a third of a standard deviation, which would be about 0.1 on a 4.0 grading scale, based on the observed GGPA distributions for the programs in the current study. The findings for the groups with the largest residuals were consistent, both in magnitude and direction, with previous results for race/ethnicity groups (Sireci & Talento-Miller, 2006) and citizenship groups (Talento-Miller, 2008).

Discussion

Combining data across several programs that differed in location, curriculum, and student demographic profiles provided support for the validity of GMAT scores and UGPA for admission to graduate study in business. Although there were some differences seen among groups for both validity and prediction, results suggest the GMAT exam scores are fair for all groups. Even though some groups tended to have lower average test scores, using a common prediction equation led to higher predicted grades on average than what was received. Some differences, such as those for the ESA group, were partially explained by the data, though other differences, such as those for the CSA group, were less apparent.

Future research, particularly on citizenship groups, is warranted to determine the full extent of the interactions among scores, previous educational experience (including major and grades), and subsequent program performance. Collecting additional data would help to increase the sample sizes of groups and possibly allow disaggregation of some of the groups to gain more insight into differences. Although additional meta-analyses on existing tests may seem unnecessary, inevitable changes to the test or to the programs they support necessitate constant vigilance to gather evidence to support the use of scores. The current study shows that the GMAT scores and UGPA are appropriate and fair for admissions decisions to graduate business programs around the world.

Author

Eileen Talento-Miller, Senior Psychometrician, Psychometric Research Department, GMAC.

Contact Information

For questions or comments regarding study findings, methodology or data, please contact Fanmin Guo, Vice President, Psychometric Research, GMAC, at research@gmac.com.

The views and opinions expressed in this article are those of the author and do not necessarily reflect those of GMAC.

Acknowledgements

The authors wish to thank Paula Bruggeman, Research Publications Manager & Editor, GMAC, for editorial review.

References

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124.

Crooks, S., & Heuvelmans, A. (1999). *The GMAT as a predictor of academic performance on eight European MBA programs*. A Study for the Graduate Management Admission Council (GMAC) by the National Institute for Educational Measurement (CITO). Arnhem, Netherlands: CITO.

Graduate Management Admission Council. (2013). *Profile of graduate management admission test candidates, 2008–09 to 2012–13*. Reston, VA: Graduate Management Admission Council.

Halpern, D. (2000). Validity, fairness, and group differences: Tough questions for selection testing. *Psychology, Public Policy, and Law, 6*, 56–62.

Hecht, L., Manning, W., Swinton, S., & Braun, H. (1989). *Assessing older applicants for admission to graduate study in management: The role of GMAT scores and undergraduate grades*. Los Angeles, CA: Graduate Management Admission Council.

Koys, D. (2005). The validity of the Graduate Management Admissions Test® for non-U.S. students. *Journal of Education for Business, 80(4)*, 236–239.

Kuncel, N., Credé, M., & Thomas, L. (2007). A meta-analysis of the predictive validity of the Graduate Management Admission Test (GMAT) and undergraduate grade point average (UGPA) for graduate student academic performance. *Academy of Management Learning and Education, 6*, 51–68.

Kuncel, N., & Hezlett, S. (2007). Standardized tests predict graduate students' success. *Science, 315*, 1080–1081.

Mattern, K., Patterson, B., Shaw, E., Kobrin, J., & Barbuti, S. (2008). *Differential validity and prediction of the SAT®*. (College Board Research Report No. 2008–4). New York: The College Board.

Oh, I., Schmidt, F., Shaffer, J., & Le, H. (2008). The Graduate Management Admission Test (GMAT) is even more valid than we thought: A new development in meta-analysis and its implications for the validity of the GMAT. *Academy of Management Learning and Education, 7*, 563–570.

Qian, H., Trang, K., & Kingston, N. (2016, April). *A meta-analysis of the predictive validity of the Graduate Management Admission Test*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Siegert, K. (2008). Executive education: Predicting student success in executive MBA programs. *Journal of Education for Business, 83*, 221–226.

Sireci, S., & Talento-Miller, E. (2006). Evaluating the predictive validity of Graduate Management Admission Test scores. *Educational and Psychological Measurement, 66*, 305–317.

Stolzenberg, R., & Relles, D. (1991). Foreign student academic performance in U.S. graduate schools: Insights from American MBA programs. *Social Science Research,* 20, 74–92.

Talento-Miller, E. (2008). Generalizability of GMAT validity to programs outside the U.S. *International Journal of Testing, 8*, 127–142.

Talento-Miller, E. (2009). *Validity study of non-MBA programs*. (Research Report Series RR-09-12*)*. McLean, VA: Graduate Management Admission Council.

Talento-Miller, E., Guo, F., & Siegert, K. (2008, July). *When are grades no longer valid? A look at the effect of time on the usefulness of previous grades.* Paper presented at the International Test Commission Conference, Liverpool, United Kingdom.

Talento-Miller, E., & Rudner, L. (2008). The validity of Graduate Management Admission Test scores: A summary of studies conducted from 1997 to 2004. *Educational and Psychological Measurement, 68*, 129–138.

Wright, R., & Bachrach, D. (2003). Testing for bias against female test takers of the Graduate Management Admissions Test and potential impact on admissions to graduate programs in business. *Journal of Education for Business, 78*, 324–328.

Young, J., & Kobrin, J. (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis*. (Research Report No. 2001–6). New York: College Entrance Examination Board.

Appendix: Citizenship Groups by World Region

United States, Canada, Australia & Pacific Islands (UCAP)

Australia
Canada
New Zealand
United States

East & Southeast Asia (ESA)

China
Hong Kong
Indonesia
Japan
Malaysia
Mongolia
Myanmar
Philippines
Singapore
South Korea
Sri Lanka
Taiwan
Thailand
Vietnam

Central & South Asia (CSA)

Bangladesh
India
Kazakhstan
Kyrgyzstan
Pakistan
Uzbekistan

Western Europe (WE)

Austria
Belgium
Cyprus
Denmark
Finland
France
Germany
Greece
Ireland
Italy
Netherlands
Norway
Portugal
Spain
Switzerland
United Kingdom

Mexico, Caribbean, & Latin America (MCLA)

Argentina
Bahamas
Bermuda
Bolivia
Brazil
Chile
Colombia
Costa Rica
Ecuador
Jamaica
Mexico
Peru
Venezuela

Africa, Eastern Europe, & Middle East (AEEME)

Algeria
Armenia
Azerbaijan
Bahrain
Belarus
Benin
Bosnia and Herzegovina
Bulgaria
Czech Republic
Egypt
Georgia
Ghana
Hungary
Iran
Israel
Ivory Coast
Jordan
Kenya
Lebanon
Lithuania
Mauritius
Moldova
Morocco
Niger
Nigeria
Poland
Romania
Russian Federation
Serbia/Montenegro
Slovakia
South Africa
Turkey
Ukraine
Zimbabwe