GRADUATE MANAGEMENT ADMISSION COUNCIL

Evaluating Analytical Writing for Admission to Graduate Business Programs

Eileen Talento-Miller, Kara O. Siegert, Hillary Taliaferro

GMAC® Research Reports • RR-11-03 • March 15, 2011

Abstract

The assessment of writing proficiency has been a part of admissions requirements for undergraduate, graduate, and professional programs for years. Research suggests the predictive validity of assessments may be limited for undergraduate admissions, and cites the need for studies in graduate schools. The current study addresses predictive validity of the writing portion of the Graduate Management Admission Test®, used for graduate business programs worldwide, by meta-analytically combining results across studies. Results suggest that average incremental validity, over other test scores and previous grades, is similar to values observed in other studies. Evidence suggests that differences related to program type and diversity within programs relative to language and citizenship moderate the validity estimates for the writing scores.

Introduction

The relatively recent additions of writing sections to the standardized tests used most frequently to make undergraduate admissions decisions in the United States draw attention to other writing assessments being used in postgraduate admissions. Writing components have been included in testing programs for admission to graduate and professional education for many years. The available research on the validity of using writing samples or writing section scores for admissions decisions is limited, however. Recent validity research on the new writing tests for the SAT and the ACT seem to suggest that although the scores are related to success, the incremental validity over what is already being explained by the traditional section scores and previous grades is very small. The current study examines predictive validity for the writing portion of the Graduate Management Admission Test (GMAT®) using meta-analysis to determine what factors contribute to the usefulness of scores in graduate business programs.

Breland, Bridgeman, and Fowles (1999) conducted a review of writing assessments used in admissions that addressed the challenges of assessing a writing construct or evaluating applicants through assigned essays. The article specifically addressed the need for predictive validity studies in graduate school, where the evaluation of writing is not as direct as may be the case in an undergraduate program. Years later, the validity evidence of writing assessments is still limited. Hojat et al. (2000) reported on the predictive validity of the essay section of the Medical College Admission Test (MCAT), although the unique scoring method employed with the section does not allow for evaluation through correlations and regression, as is typically the case with validity studies. The evidence suggested no effect for the criterion of grade point average (GPA) in medical school. For the Graduate Record Examination (GRE), an article from the Educational Testing Service (ETS, 2007) provided evidence supporting the construct validity of the writing assessment, but did not provide correlational analyses with graduate GPA. A meta-analysis of validity studies for the GMAT exam suggested that the correlations of scores from the Analytical Writing Assessment (AWA) section of the test were lower than those of the other test sections. The study, however, did not specify the incremental increase in validity that can be attributed to the inclusion of the section as part

of the admissions decision (Talento-Miller & Rudner, 2008).

Similar to the MCAT and GMAT studies, research on the new assessments for undergraduate admissions suggest the contribution of a writing section could be limited. The technical report for the ACT Writing Test indicates that the multiple correlation predicting grades is increased by 0.02 when the writing score is added to the ACT English score (ACT, 2009). Validity studies conducted using the new writing section of the SAT revealed similar results. Adding the writing section to the other predictor variables increased the multiple correlations by 0.01 when predicting first-year average. The SAT study also revealed that the correlations of writing scores with grades in English Composition courses were lower than the correlations with first-year GPA. In addition, the correlations of the essay portion of the writing test with the outcome variables were lower than those of the multiple-choice portion of the writing assessment (Norris, Oppler, Kuang, Day, & Adams, 2006).

Other studies have suggested that multiple-choice tests are more valid in predicting performance than an essay test (e.g. Michael & Shaffer, 1979). Although essay tests would arguably exhibit more face validity in measuring writing, multiple-choice tests can administer more questions and draw from a greater sampling of the desired domain, leading to more reliability. Challenges in assessing the reliability of essay tests include the administration of different essay prompts to examinees, which are then rated by various evaluators (Siegert & Guo, 2009). Comparisons of validity between sections should include consideration of differential reliability to ensure the evaluations are made between the abilities desired as opposed to the measures used.

Other considerations may affect the usefulness of a writing assessment for admissions. For instance, an early study of the GMAT AWA suggested that including the section in admissions decisions could affect the gender composition of admitted students leading to the inclusion of more female students (Bridgeman & McHale, 1996). In surveying users of the AWA, Owens (2006) found that programs with higher concentrations of non-native English applicants tended to find the section more useful than the other programs. The meta-analysis of GMAT validity studies showed that AWA scores had higher correlations with GPA for programs located outside the United States compared to programs within the United States (Talento-Miller & Rudner, 2008). In addition, an earlier study suggested differences in predicting graduate program grades based on whether the student's primary language was English (Hendel & Doyle, 1978). Studying whether the predictive validity of AWA scores differs by group would be necessary in order to determine the overall effect of their inclusion in admissions decisions.

Using the information collected from numerous programs, this study examined the validity of the GMAT AWA scores for predicting mid-program grades in graduate business programs. Appropriate adjustments for reliability and restriction of range were made to the correlations according to methods described by Hunter and Schmidt (1990). Overall and incremental validity were evaluated. Differences were examined by demographic groups to determine whether or not the measure is more useful with some groups compared to others. In addition, descriptive information was gathered to determine whether program characteristics could be identified that moderate the usefulness of AWA scores.

Methodology

The AWA section of the GMAT exam has been in use since 1994. Examinees are asked to write two essays based on prompts—one that requires an analysis of an issue and the other an analysis of an argument. The prompts are randomly selected by the computer-based testing software from a list that is available prior to testing. Because the entire exam is computer administered, the essays are written on the computer and require basic word processing skills. Thirty minutes are allowed for completing each essay. The essays are scored holistically by two raters using a scale that ranges from zero to six, with adjudication by a third rater if necessary. At least one rater is human, with computer scoring available as the other rater. The use of computer scoring as one of the raters is well established (Rudner, Garcia, & Welch, 2005) and the inter-rater reliability is relatively high (Siegert & Guo, 2009). Final scores are calculated as the average of the ratings on the two essays rounded to the nearest half

point. In addition to AWA scores, the essays themselves are made available to schools to aid in their admissions decisions.

This study used data submitted to the Graduate Management Admission Council (GMAC) as part of its Validity Study Service (VSS). Through the VSS, GMAC collects data on GMAT scores and undergraduate GPA of enrolled students as predictors, and mid-program grades as the criterion variable. More than 300 studies have been conducted since 1997, though not all of the datasets include information on AWA scores. Studies were removed from the sample if their sample size was less than 100 cases. The dataset used for this study included 195 validity studies that included information on AWA scores.

Predictive validity was calculated using correlations and multiple regressions with adjustments for restriction of range and attenuation based on Hunter and Schmidt (1990). Restriction of range corrections allow for the generalization from the sample of admitted students to the population of applicants. The restriction of range corrections used scores sent to the school as the hypothesized distribution of applicants for that program. For the corrections for attenuation, reliability values were obtained from previous research. Reliability estimates for the AWA measure were based on the research by Siegert and Guo (2009). For the reliability of the criterion of mid-program GPA, this study followed the example of meta-analyses by Kuncel, Hezlett, and Ones (2001) and Kuncel, Credé, and Thomas (2007) in using the average of values for GPA from previous studies. Incremental validity will be determined as the difference in the multiple correlations after entering the other GMAT scores and undergraduate GPA.

In order to determine whether validity differs among groups, data was summarized across VSS studies that included these variables as standard. Validity studies conducted since 2005 included gender and citizenship as standard grouping variables. Because citizenship groupings are coded as domestic versus nondomestic, comparisons included only US-based programs. Data was available from 27 studies for the gender comparison and from 17 studies for the citizenship comparison. Within each study, correlations were calculated separately by group. Validity values for AWA were summarized across programs and compared by group.

Previous data suggested that program characteristics might affect estimates of validity. Using the current dataset, validity was summarized for different program types and for US versus non-US locations. To determine further whether there are specific program characteristics that moderate the degree of predictive validity observed, data was gathered on the current applicant pool for the programs, which consisted of all examinees sending scores to the program within the most recent testing year. Several applicant characteristics were examined as possible moderating variables: percentage of female applicants, percentage of non-US citizens, percentage of non-native English students, average work experience, and average GMAT scores.

To examine the potential impact of these characteristics on validity, VSS results of the programs were categorized based on the adjusted simple correlations of AWA scores with mid-program grades. The low AWA validity category consisted of programs with correlations less than 0.1 including any observed negative relationships. The high validity category was defined as programs with correlations greater than 0.4. Programs within these two AWA validity categories were compared based on percentage of female applicants, percentage of non-US citizens, percentage of non-native English students, average work experience, and average GMAT scores to determine whether program characteristics are related to AWA predictive validity.

Results

For the 195 studies with adequate sample sizes that included AWA score as a predictor, the average adjusted correlation with mid-program grades is approximately 0.2. The incremental validity values were derived from several different combinations including either GMAT Total score or the combination of Verbal and Quantitative score, with or without undergraduate GPA in the combination. Because of multiple possible combinations, each VSS study could have multiple values for incremental validity resulting in a total of 668 comparisons. The median value suggests the multiple correlation

increases by only 0.01 when AWA scores are added to the predictive power already accounted for by different GMAT scores with or without UGPA included in the equation. The mean and standard deviation suggest that the increase could be quite a bit more for a few programs. Table 1 summarizes the results.

Table 2 shows the average validity values for recent studies that included separate analyses by gender and citizenship groups. Results for all groups from these recent studies are slightly higher than the overall average, which includes studies dating back to 1996. In general, average validity values were higher for the male and non-US citizen groups. The effect was more noticeable when comparing US to non-US citizens with a Cohen's *d* effect size of 0.358 (using the full sample standard deviation of 0.165); however, the number of non-US citizens was limited relative to the other comparisons.

Differences were then compared by program type and location. In Table 3, the highest validity values are observed for executive MBA (EMBA) and full-time program types and programs in non-US locations. The mean values for full-time and EMBA programs were about half a standard deviation higher than the part-time and other program types. The difference was smaller when comparing program location means, and the median values were relatively close. As with the non-US citizenship comparison within programs, the number of cases available for study was limited in relation to the multitude of data available for other comparisons.

Table 1: Summary of AWA Validity and Incremental Validity for VSS Studies

	K	Mean (SD)	25th percentile	Median	75th percentile
All	195	0.222 (0.165)	0.128	0.228	0.332
Incremental validity	668	0.025 (0.043)	0.001	0.010	0.029

Table 2: AWA Validity for Demographic Groups within VSS Studies

	K	N	Mean (SD)	Median
All	*195*	*38,061*	*0.222 (0.165)*	*0.228*
Male	27	6,636	0.266 (0.154)	0.250
Female	27	3,450	0.232 (0.159)	0.223
US citizens	17	6,378	0.272 (0.113)	0.250
Non-US citizens	17	1,106	0.331 (0.181)	0.352

Table 3: Differences by Program Type and Location

	K	N	Mean (SD)	Median
All	*195*	*38,061*	*0.222 (0.165)*	*0.228*
Full-time	94	15,702	0.259 (0.127)	0.274
Part-time	18	5,538	0.178 (0.147)	0.157
Executive MBA	11	1,288	0.266 (0.144)	0.268
Other	29	4,147	0.173 (0.203)	0.169
US programs	180	35,072	0.219 (0.162)	0.227
Non-US programs	15	2,989	0.268 (0.187)	0.233

Consistent with the other findings, data about programs with the extreme validity values suggest that AWA scores may be more effective for programs with high concentrations of non-US citizens or non-native English speakers. Table 4 shows that while most of the applicant characteristics are similar for low- and high-validity programs, the main difference appears to be with the percentage of non-US citizens and the percentage of non-native English speakers. An area that may be relevant for further study is the apparent difference between the two programs in terms of average years of work experience, where the programs with high validity tend to attract applicants with less work experience than those with low validity.

Discussion

The use of writing assessments for the selection of applicants for admission into graduate study has not been widely evaluated (Breland, Bridgeman, & Fowles, 1999). The current study examined the predictive power of one writing assessment, the GMAT AWA, using data collected from 195 validity studies. Overall, the AWA was able to predict 5 percent of the variability in mid-program graduate management grades. Similar to previous research examining writing assessments used for admissions selection into undergraduate programs, incremental validity results for this graduate-level writing assessment were small, averaging 0.01.

The AWA, nevertheless, appeared to be a better predictor of performance for non-US programs and citizens and non-native English speakers. Similar results were reported in previous research conducted by Owens (2006), Talento-Miller and Rudner (2008), and Hendel and Doyle (1978). Moreover, for the GMAT writing assessment, predictive validity was higher for executive and full-time MBA programs. Perhaps performance in these programs is more influenced by writing and research paper assignments than it is in part-time programs. Additional research should attempt to replicate the EMBA findings using a larger sample of executive programs and students. In addition, a qualitative examination of the difference in program requirements may also indicate why predictive validity is higher for executive and full-time programs.

Overall, the AWA writing assessment can be useful in predicting graduate management success, particularly for full-time and executive programs and for programs with a high concentration of non-US citizens and non-native English speakers. This is not surprising given that English language barriers, both written and spoken, are likely one of the largest obstacles standing in the way of success for these students. As such, the GMAT AWA may be a useful tool in selecting non-US and non-native English-speaking applicants who are more likely to succeed in graduate management programs.

Table 4: Mean (SD) of Applicants to Programs with Low and High AWA Validity

	Low (< 0.10) K = 29	High (> 0.40) K = 20
GMAT Total	545 (53)	547 (50)
GMAT Quant	36 (5)	37 (5)
GMAT Verbal	29 (3)	28 (3)
GMAT AWA	4.6 (0.4)	4.5 (0.3)
Work experience	4.2 (1.5)	3.5 (2.1)
% Female	36.2% (5.9)	37.2% (9.9)
% Non-US citizen	36.0% (22.8)	51.4% (30.9)
% Non-native English	38.4% (20.8)	50.2% (26.8)

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

ACT. (2009). *ACT writing test technical report*. Iowa City, IA: Author.

Breland, H., Bridgeman, B., & Fowles, M. (1999). *Writing assessment in admission to higher education: A review and framework* (College Board Research Report 99-3). New York: College Board.

Bridgeman, B., & McHale, F. (1996). *Gender and ethnic group differences on the GMAT Analytical Writing Assessment* (ETS Research Report 96-2). Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2007). *A comprehensive review of published GRE validity data*. Princeton, NJ: Author.

Hendel, D., & Doyle, K. (1978). Predicting success for graduate study in business for English-speaking and non-English-speaking students. *Educational and Psychological Measurement, 38,* 411–414.

Hojat, M., Erdmann, J., Veloski, J., Nasca, T., Callahan, C., Julian, E., & Peck, J. (2000). A validity study of the writing sample section of the Medical College Admission Test. *Academic Medicine, 75*, 525–527.

Hunter, J., & Schmidt, F. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage Publications.

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

Kuncel, N., Hezlett, S., & Ones, D. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127*, 162–181.

Michael, W., & Shaffer, P. (1979). A comparison of the validity of the Test of Standard Written English (TWSE) and of the California State University and Colleges English Placement Test (CSUC-EPT) in the prediction of grades in a basic English composition course and of overall freshman-year grade point average. *Educational and Psychological Measurement, 39*, 131–145.

Norris, D., Oppler, S., Kuang, D., Day, R., & Adams, K. (2006). *The College Board SAT Writing validation study: An assessment of predictive and incremental validity* (College Board Research Report 06-02). New York: College Board.

Owens, K. (2006). *Use of the GMAT Analytical Writing Assessment: Past and present* (GMAC Research Report 07-01). McLean, VA: Graduate Management Admission Council.

Rudner, L., Garcia, V., & Welch, C. (2005). *An evaluation of IntelliMetric Essay Scoring System using responses to GMAT AWA prompts* (GMAC Research Report 05-08). McLean, VA: Graduate Management Admission Council.

Siegert, K., & Guo, F. (2009). *Assessing the reliability of GMAT Analytical Writing Assessment* (GMAC Research Report 09-02). McLean, VA: Graduate Management Admission Council.

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.