Assessing the Reliability of GMAT® Analytical Writing Assessment

Kara O. Siegert and Fanmin Guo

Abstract

A variety of educational constituencies are increasingly using assessments that evaluate students' analytical writing abilities. The complexity of these assessments makes it challenging to evaluate the reliability of task ratings. Many performance assessments do not lend themselves to standard generalizability (G) theory designs or to inter-rater reliability estimation procedures. This paper used a series of methods, including G theory, inter-rater reliability, and test-retest reliability, to determine the consistency of scores for one writing assessment, the Graduate Management Admission Test® (GMAT®) Analytical Writing Assessment (AWA). Results are compared and suggestions for determining the reliability for nonconventional writing assessment designs are discussed.

Purpose

The assessment and measurement field touts the importance of evaluating the reliability of test scores and the validity of inferences made from these scores. Without this critical information, test publishers would have very little evidence that the assessments they produce will meet the needs of those buying their products. The test evaluation process can become much more challenging, however, when the assessments involve writing prompts, rubrics, and essay raters, rather than multiple-choice questions and automated, objective scoring of responses.

The purpose of this study was to evaluate the reliability of an analytical writing assessment using a multifaceted approach. Specifically, data from the Analytical Writing Assessment (AWA), one component of the Graduate Management Admission Test (GMAT), were examined using generalizability (G) theory, inter-rater reliability, and test-retest reliability estimates.

The assignment of essay prompts to test takers and the scoring of essays by AWA raters do not follow typical G theory designs. As a result, this study uses two G study designs applied to multiple data sets to define a reliability coefficient range. Additionally, given that the AWA section requires test takers to compose two essays, each of which is then scored by two different raters; standard estimates of inter-rater reliability do not apply. This study presents and applies a method to determine reliability across essays and raters. We calculated estimates of test-retest reliability using data collected from examinees who took the GMAT exam on two separate occasions (repeaters). We hope that by comparing results from a variety of reliability methods, other writing assessment programs with nonconventional designs can apply these techniques to evaluate the dependability and accuracy of their own assessments.

Theoretical Framework

One critical component of any good assessment—regardless of its complexity or simplicity—is that it must be reliable. Scores for a given examinee must be consistent over different occasions in order to make valid inferences. To evaluate the reliability of writing assessments, it is especially important to consider the various components that influence a test taker's score. The evaluation of reliability, however, is often challenging, especially with large-scale, high-stakes writing assessments such as the AWA, which is delivered through computers.

GMAT® AWA

Description and scale

In 1994, the AWA was added to the GMAT exam with the intent that it would assist in admission selection and diagnosis of writing deficiencies among applicants to graduate management education programs (Graduate Management Admission Council® [GMAC®], n.d.). During the AWA, examinees are asked to construct essays based on two prompts, Analysis of an Issue (AI) and Analysis of an Argument (AA). The AI task presents examinees with an issue and asks them to explain their viewpoint. Examinees must provide a sound reasoning for their response that is based on personal experiences or relevant examples. The AA task requires examinees to investigate and critique the reasoning behind an argument. The entire

AWA section lasts 60 minutes. Examinees have 30 minutes' time to complete each of the essays.

Prompts for the AI and AA tasks can span a number of different topic or subject areas including, but not limited to, business-related scenarios. From the large pool of essay prompts, the computer program selects two, one AI and one AA, for each examinee. The selection of essay prompts is random, resulting in a sparse matrix with a great deal of missing data. Some subsets of examinees see the same AI and AA essay prompts, while others see two different prompts. (See Table 1 for an example.) The AWA score is calculated as the mean of the two essay scores, which are the means of the two raters on each essay. The AWA score scale is 0 (*unscorable*) to 6 (*outstanding*) with increments of 0.5.

Table 1. AWA Task and Prompts Format										
	Tasks									
	AI Essay Prompts					AA Essay Prompts				
Test takers	AI1	AI2	AI3	AI4	AIn	AA1	AA2	AA3	AA4	AAn
1	X					X				
2		X								X
3				X		X				
4		X								X
5					X					X
n				X					X	

Although all examinees are administered both an AI and an AA task (i.e., persons [p] are crossed with task types [t] [AI and AA]), each examinee only sees one prompt for each task type. The assumption of (p x t) in the data is not satisfied because prompts are unique to both the AI and the AA tasks. As a result, prompts are confounded within the AI and AA tasks. This makes it challenging to determine if task effects are the result of prompt or essay type differences.

Raters and scoring

Complexity increases even more when we examine the AWA scoring. As previously mentioned, each examinee-written essay, AI and AA, is holistically scored by two different raters. At least one of the raters for each essay is human; the second rating may

be provided by an automated essay scoring engine. Human raters primarily include college and university faculty members. Standardized procedures are used to train them as readers for either the AI or the AA section. If the two ratings for a given essay differ by more than one point, a third rater, who is human, provides an additional rating, which is used for adjudication. Finally, the four ratings, two for each essay, are averaged to provide a holistic AWA score based on responses to both prompts (GMAC®, n.d.).

Similar to the random assignment of essay prompts to examinees, the assignment of raters to examinee-written essays is also somewhat random. Each rater (r) scores only essays written for either the AI or AA section, so two different raters are nested within each of the two task types, r : t. Within the AA or AI task,

however, raters may score any of the administered essay prompts and are randomly assigned to do so. As a result, the pair of raters scoring AI or AA essays may vary across examinees, as shown in Table 2. Given the complicated nature of the data, evaluation of reliability is a complex process for the GMAT AWA.

	Table 2. AWA Task, Prompts, and Rater Format									
	Tasks									
	AI Essay Prompts					AA Essay Prompts				
Test takers	AI1	AI2	AI3	AI4	AIn	AA1	AA2	AA3	AA4	AAn
1	R_1R_2					$R_{11}R_{12}$				
2		R_3R_2								$R_{14}R_{15}$
3				R_4R_5			$R_{13}R_{12}$			
4			R_1R_3					$R_{11}R_{15}$		
5					R_5R_3					$R_{15}R_{13}$
n				R_4R_2					$R_{14}R_{13}$	

Previous AWA Reliability Research

Breland, Bridgeman, and Fowles (1999) provided a comprehensive summary of information on the reliability and validity of test scores from several large-scale admission test writing assessments. Based on their personal communications with researchers, Breland et al. (1999) reported reliability research on the GMAT AWA. Specifically, Cronbach's estimates of reliability ranged from .66 to .79 for three test administrations during 1995, while G coefficients were reported between .54 and .76. Split-half estimates of reliability for a one-month period in 1994 resulted in estimates between .51 and .77. Unfortunately, the methodology used to obtain these estimates was not described and is based on the paper version of the GMAT exam. Since the prompts to examinees were not randomly assigned and the number of prompts used for paper-based administrations was limited, it is difficult to determine whether these results would generalize to the computer-adaptive version of the exam. In addition, this research provided little information about the methodology used to obtain these estimates. An investigation of other research can provide alternative reliability methods that are applicable to the current study.

G Theory Research

Two main G study designs have been suggested to manage problems posed by complex data, specifically with regard to sparse rater data. Lee and Kantor (2005) used a method that would overlook rater differences across examinees. Instead, this design uses ratings (r') rather than raters as a facet. For example, since all examinees receive the same number of ratings, this means all examinees have equal data that can be treated as a random facet for the study design. As a result, missing data due to different rater assignments was no longer an issue. The study design allows for estimation of variance components and assessment of impact for ratings.

Wang, Zhang, and Li (2007) compared estimates of rater variance and G coefficients utilizing designs that included either a rating or rater facet. Their results showed slight differences in the percent of variance accounted for by raters depending on how the facet was modeled (i.e., raters vs. ratings). These differences were not statistically significant, however. Specifically, G coefficients for rating study designs were slightly, but not meaningfully, higher than those produced from rater designs. One could conclude that rating designs may slightly underestimate the impact of raters on performance and overestimate reliability, although neither discrepancy was large enough to be meaningful.

Gao, Wang, and Brennan (2000) described another alternative to complex data situations. Their study demonstrated the results of treating the rater facet as hidden within a G study design. Similar to Lee and Kantor (2005), they argued that if the rater facet contributes little to the variance of examinee scores, ratings can be averaged across raters to provide one overall rating for each task. In the resulting G study design, the rater effect would then be hidden or not included as a facet.

Haertel (2006) also discussed the use of hidden facets with respect to G theory analyses. It was noted that hidden facets can be especially useful when the data collection method leads to difficulty in interpreting results because of confounded variables. Specifically, occasions are often treated as a hidden or unmeasured facet in G study designs due to limitations in collecting data over multiple instances. This does not diminish the potential impact that occasion has on performance, however. Both Gao et al. (2000) and Haertel (2006) also note that the use of hidden facet designs may lead to difficulty interpreting results.

Complications such as missing data and confounded variables make it challenging to evaluate the consistency of AWA scores. To determine the range of possible G coefficients for the AWA, given the complex nature of administration and scoring of the section, two G study designs were implemented and the results compared across multiple samples of data. In addition, estimates of inter-rater reliability provide information on the consistency of different raters. This is important given that raters, in the traditional sense, were not included as a facet in the G study designs. Finally, test-retest reliability—a more traditional method of evaluating consistency across testing occasions—was calculated to provide verification of the G study results and complete this multifaceted approach.

Method

G Study

Similar to other large-scale writing assessments, the administration and scoring methods for the AWA do not follow a fully crossed design (p x r x t), where all persons (p) take all tasks (t) and all tasks and persons are scored by all raters (r). Also, the randomness with which raters and prompts are assigned, as well as volume of the prompts and rater pools, makes it challenging to include these variables as facets in a traditional crossed G study design. Although challenging from an analysis perspective, the random assignment of raters and prompts satisfies the major G theory assumption that conditions of each facet are randomly sampled. While samples of data that match the p x r x t design could be selected from the examinee population, their size would be quite small— fewer than 20 examinees. Such small sample sizes might not generalize to the larger population or replicate in future research.

As a result, the current study used both the rating facet (Lee and Kantor, 2005) and hidden rater (Gao et al., 2000) methods to calculate a range of reliability coefficients from multiple samples of essays. The rating facet will be applied to a two-facet p x (r': t) design, where ratings are nested within tasks and all persons complete both AI and AA tasks and receive two ratings. Both ratings and tasks will be treated as random facets. This design is similar to the existing structure of AWA, where raters score only one of the two tasks, and all examinees take all tasks. All raters, however, do not score all examinees. Thus, the use of a rating rather than rater facet will be useful in performing the analysis. As mentioned in the previous research (Wang et al., 2007), by including the rating, rather than rater, facet variance components may slightly underestimate rater variance. Because the rating facet is nested within tasks, the variance component for r' will be confounded with variance also accounted for by tasks.

For the hidden rater design, a one-facet p x t design, where all persons take both tasks, will be modeled and raters will be hidden. Tasks will be treated as a random facet and scores on each task will be averaged across raters in this design. Although the simplicity of this design makes it easy to replicate, the impact of raters or ratings on performance cannot be determined. If the rating facet contributes little to examinee performance, the results from the two designs should be similar. Otherwise, the p x t design should produce the higher G coefficients. Both designs will likely produce an upper-bound G coefficient given that neither design strictly represents the current data collection and scoring procedures.

Data

Nine samples of data were extracted from a larger population of GMAT examinees. The larger population consisted of 148, 210 tests that were completed between January 2006 and March 2007. Each of the extracted samples contained at least 115 examinees who were administered the same AI essay prompt and the same AA essay prompt. Additionally, these samples included only those examinees who received one machine-score rating for both essays and whose scores did not require adjudication. This resulted in a total sample of 125,291 examinees from which nine samples were drawn.

The selection of multiple samples allows for comparisons between the estimated variance components and G coefficients across different essay prompts and study designs (i.e., p x t and p x (r': t)). Haertel (2006) recommended the use of multiple data sets to ensure precision among G study variance component estimates. A total of seven different AI and seven different AA prompts were examined across the nine samples. A total of 15 AI and 13 AA raters were used across the nine samples. Because there was some overlap in the AI and AA essay prompts and raters across samples, the subsets are not completely independent of one another. A general understanding of precision can be gauged, however, by examining variance and standard error estimates across the multiple samples. To further evaluate the impact of rater differences not modeled in the G study designs, we calculated estimates of inter-rater reliability.

Inter-Rater Reliability

This study serves two purposes. One is to present a method of estimating the scoring reliability for the unique test configuration of the AWA. The other is to provide an estimate of the inter-rater error. If error is minimal, hiding the rater effect or using a rating effect in the G study might be justifiable.

Since each examinee writes two essays during the AWA session, the estimation of inter-rater reliability needs to consider the errors of two pairs of raters. Livingston (2004) proposed a solution for calculating inter-rater reliability when more than one performance assessment task is administered; however, only a random proportion of examinees are scored by two raters. The others are scored by only one rater. This method can easily be adapted to current data situations by setting the proportion of double-rated tasks to 1 and ignoring the calculations for the single-rated tasks. For simplicity, we will use abbreviations: AWA for the final total score, AA for the Argument essay score, AI for the Issue essay score, 1st for Rater One, and 2nd for Rater Two. We will also use rel for reliability coefficient, VES for variance of error of scoring, and Var for variance. Here are the calculations adapted from Livingston's formulas:

$$rel_{AWA} = 1 - \frac{VES_{AWA}}{Var_{AWA}},$$

where

$$VES_{AWA} = 0.5(VES_{AA}) + 0.5(VES_{AI}),$$

$$VES_{AA} = \frac{1}{4}(1 - rel_{AA,1st,2nd})(Var_{AA,1st} + Var_{AA,2nd}), \text{ and}$$

$$VES_{AI} = \frac{1}{4}(1 - rel_{AI,1st,2nd})(Var_{AI,1st} + Var_{AI,2nd}).$$

Using the same testing period selected for the G study analyses, a total of 143,859 examinees with valid scores were identified and included in this study.

Test-Retest Reliability

In order to estimate the test-retest reliability of the AWA, we identified 11,593 repeaters in the data set and calculated the correlations between their first and second AWA scores. Although this was not a perfect data set for this purpose, the estimated reliability will provide an approximation of test-retest reliability.

Results

G Study

The descriptive statistics for the nine samples can be found in Table 3. The statistics summarize the data across all raters who scored essays within a given sample. For all samples and both AI and AA tasks presented in Table 3, Rating 1 represents the average computer score, while Rating 2 represents the average human score. The two ratings for each task were also averaged to provide a mean rating for the AI and AA

for each sample. These AI and AA averages can be compared across the samples to determine whether the ratings vary across the different prompts. The average AI and AA ratings across the nine samples, 4.59 and 4.58, respectively, were not significantly different from one another, $t(7) = 0.91$, $p > .05$.

	Table 3. Descriptive Statistics for AI and AA Samples											
	AI						AA					
	Rating 1		Rating 2		Average		Rating 1		Rating 2		Average	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
1	4.71	0.98	4.71	0.90	4.71	0.90	4.53	0.94	4.44	0.91	4.49	0.86
2	4.78	0.93	4.59	0.79	4.68	0.82	4.56	0.90	4.57	0.87	4.57	0.83
3	4.78	1.01	4.69	0.89	4.73	0.91	4.81	1.00	4.63	0.89	4.72	0.90
4	4.60	0.89	4.53	0.97	4.56	0.89	4.63	0.98	4.50	0.87	4.56	0.97
5	4.47	0.93	4.40	1.02	4.44	0.93	4.60	1.04	4.45	1.01	4.53	0.98
6	4.53	0.97	4.49	1.01	4.51	0.94	4.58	1.01	4.55	0.97	4.56	0.95
7	4.68	1.00	4.63	1.05	4.66	0.99	4.54	1.08	4.53	0.95	4.53	0.97
8	4.73	0.96	4.73	0.98	4.73	0.93	4.71	1.01	4.58	1.03	4.65	0.98
9	4.36	1.06	4.29	0.83	4.32	0.90	4.80	0.92	4.54	0.94	4.67	0.89

Two-facet p x (r': t) design

In this design, the rating facet was nested within tasks, meaning that there were multiple, different ratings for each task. In addition, all persons took all tasks (AI and AA) and received scores on all ratings. As such, the G coefficients for this design are likely to be upper-bound estimates of actual AWA reliability given that the current data collection procedures use different raters rather than ratings.

Table 4 provides the variance component and standard error estimates for each of the nine extracted samples, as well as averages across all samples. The rating main effect is confounded with the interaction between ratings and tasks, thus a variance component for the main effect of the rating facet is not available. To determine if there were wide variations in raters, we provide estimates of inter-rater reliability in a following section. By examining the average variance component for the rating and task nesting (r: t), it does not appear that ratings varied greatly based on the type of task (i.e., AI vs. AA). Similarly, when we examined the task main effect, we found negligible differences across the AI and AA sections. The AI and AA tasks appear to be of equal difficulty.

Table 4. Variance and Standard Error Estimates Using p x (r': t) Design										
	p		t		r:t		pt		pr:t	
	$\hat{\sigma}^2$	SE	$\hat{\sigma}^2$	SE	$\hat{\sigma}^2$	SE	$\hat{\sigma}^2$	SE	$\hat{\sigma}^2$	SE
Sample 1 (*n* = 122)	.569	.087	.023	.021	.001	.001	.122	.028	.178	.016
Sample 2 (n=119)	.407	.072	.001	.007	.008	.007	.178	.035	.177	.016
Sample 3 (n=118)	.645	.095	.000	.004	.009	.008	.084	.023	.171	.016
Sample 4 (n=118)	.509	.075	.000	.002	.003	.003	.173	.035	.175	.016
Sample 5 (n=117)	.620	.102	.000	.004	.005	.005	.213	.040	.172	.016
Sample 6 (n=116)	.676	.103	.000	.001	.000	.000	.125	.030	.184	.017
Sample 7 (n=116)	.725	.111	.006	.006	.000	.001	.154	.031	.160	.015
Sample 8 (n=115)	.674	.105	.000	.003	.003	.003	.154	.032	.160	.015
Sample 9 (n=115)	.598	.093	.050	.050	.017	.013	.124	.028	.161	.015
Average	0.552	0.094	0.009	0.011	0.005	0.005	0.147	0.031	0.171	0.016

The variance component for the p x t interaction indicated that, on average, the relative standing of examinees varied somewhat based on task type. Approximately 17% of the variance in AWA performance was related to the interaction between persons and task type. Thus, some examinees found the AI task to be more difficult while others found the AA task to be more challenging. The r x p interaction was confounded with the interaction between the object of measurement, the rating and task facets, and error. This combination, referred to as the residual, accounted for approximately 19% of the variability in examinee scores. Finally, the object of measurement accounted for the greatest amount of variance, an average of 62%. Thus, there was a great deal of variability in examinee performance on the AWA.

The G coefficients and estimates of relative error for the p x (r': t) design can be found in Table 5. The G coefficients across the nine samples ranged from .753 to .884, with an average value of .835. There also were slight variations in relative error across the different samples, with a range of .085 to .149 and an average estimate of .116. These findings indicate that accurate generalizations about examinee analytical writing ability can be made based on performance on the AWA. Moreover, performance was mostly attributable to differences in examinee ability with slight variations due to the examinee and task type interaction.

Table 5. G Coefficient and Relative Error Estimates Using p x (r': t) Design										
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Average
$E\hat{\rho}^2$	.844	.753	.884	.796	.806	.862	.861	.851	.854	0.835
$\hat{\sigma}^2 (\delta)$	.105	.133	.085	.130	.149	.108	.117	.117	.102	0.116

One-facet p x t design

In this design, the influence of raters was hidden or not modeled in the G study design. Based on the previous one-facet results reported for this study, it does not appear that rating differences influenced performance. Thus, there is some evidence to support the hiding of the rater facet; however the inter-rater reliability results will provide additional verification. Although prompts were selected to mimic a fully crossed design for this study, during actual AWA administrations all examinees do not see all prompts.

Table 6 provides the variance component and standard error estimates for the hidden rater design. Again, the task facet accounts for little variability, approximately 1%, in examinee performance and does not appear to differ in terms of difficulty. The major contributor to AWA performance was systematic individual differences among examinees, accounting for 71% of the variance in scores. The remaining unaccounted for variance, approximately 28%, was due to unmeasured error and the interaction between examinees and tasks. Overall, the variance components for the object of measurement and task facet were similar across the two different designs.

Table 6. Variance and Standard Error Estimates Using p x t Design						
	p		**t**		**pt**	
	$\hat{\sigma}^2$	**SE**	$\hat{\sigma}^2$	**SE**	$\hat{\sigma}^2$	**SE**
Sample 1 (n=122)	.569	.087	.024	.021	.211	.027
Sample 2 (n=119)	.407	.072	.005	.006	.266	.034
Sample 3 (n=118)	.645	.095	.000	.000	.170	.022
Sample 4 (n=118)	.509	.085	.000	.000	.261	.033
Sample 5 (n=117)	.620	.102	.001	.003	.299	.039
Sample 6 (n=116)	.676	.103	.000	.001	.217	.028
Sample 7 (n=116)	.725	.111	.005	.006	.234	.031
Sample 8 (n=115)	.674	.105	.001	.003	.234	.031
Sample 9 (n=115)	.598	.093	.059	.049	.204	.027
Average	.603	.095	.011	.010	.233	.030

The G coefficients and relative error estimates for the p x t design are presented in Table 7. Again the results for this design were comparable to those found with the p x (r': t) design. The G coefficient estimates ranged from .753 to .884, with an average estimate of .835. Standard error estimates averaged across all nine samples were also identical in both designs at .116. When we compared both designs, we found that main effects for ratings and tasks have little impact on overall performance on the AWA. We can infer that accurate generalizations can be made using either of the two AWA data collection designs presented in this study. Although the G coefficients reported previously in this study were relatively high, these values are likely upper-bound estimates of reliability given the differences that exist between the researched and actual observed designs.

Table 7. G Coefficient and Relative Error Estimates Using p x t Design										
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Average
$E\hat{\rho}^2$	.844	.753	.884	.796	.806	.862	.861	.852	.854	.835
$\hat{\sigma}^2 (\delta)$	.105	.133	.085	.130	.149	.108	.117	.117	.102	.116

Inter-Rater Reliability

We identified a total of 143,859 test takers who had four valid rater scores, two for each essay, and an AWA total score. The means, standard deviations, and variances of the four rater scores and the AWA total score are displayed in Table 8. The observed correlation coefficients between the two raters were 0.806 for the AA essay and 0.826 for the AI essay. We used these as the estimates for the rater reliability coefficients ($rel_{AA, 1st, 2nd}$ and $rel_{AI, 1st, 2nd}$) together with the variances of the rater scores in the calculations of the overall inter-rater reliability. The estimated inter-rater reliability for the AWA final score is 0.88. This suggests that the rater effect is very small.

Table 8. Means, Standard Deviations, and Variances of Rater and Total Scores			
	M	SD	Var
AA_{1st}	4.53	1.002	1.003
AA_{2nd}	4.46	0.954	0.910
AI_{1st}	4.56	0.992	0.983
AI_{2nd}	4.52	0.969	0.939
AWA	4.63	0.877	0.769
Note: N = 143,859			

Test-Retest Reliability

We identified and used a total of 11,593 repeaters with valid AWA scores in this study. The mean time between their first and second tests was 81 days with a standard deviation of 64. It seemed that their first and second AWA scores were very similar with similar means and standard deviations (See Table 9). The means were 4.49 for the first test and 4.56 for the second test. The standard deviations were also very similar, 0.88 for the first test and 0.85 for the second test.

Table 9. Means and Standard Deviations of the First and the Second AWA Scores			
	N	M	SD
Score 1	11,593	4.49	0.88
Score 2	11,593	4.56	0.85
All examinees	143,859	4.63	0.88

The average gain was 0.06 with a standard deviation of 0.58 on a scale of 0 to 6. Figure 1 presents the frequency distribution of the gain scores. Approximately 37 % of the repeaters received identical scores on the two tests; nearly 46 % received two scores that differed by 0.5 points; and nearly 13.5 % had two scores with a one-point difference. The estimated correlation is 0.78 between the first and second scores. By definition, that 0.78 is also our estimated test-retest reliability coefficient with this sample.

Although the repeater data are not the best data for estimating test-retest reliability, the observed reliability coefficient might be very reasonable because

1. The gain from the first to the second test is very small.

2. The means and standard deviations of the first and second scores are very similar.

3. The distribution characteristics of the two scores of the repeaters are very close to those for the entire group (See Table 9).

Figure 1. Distribution of Gains from Test One to Test Two
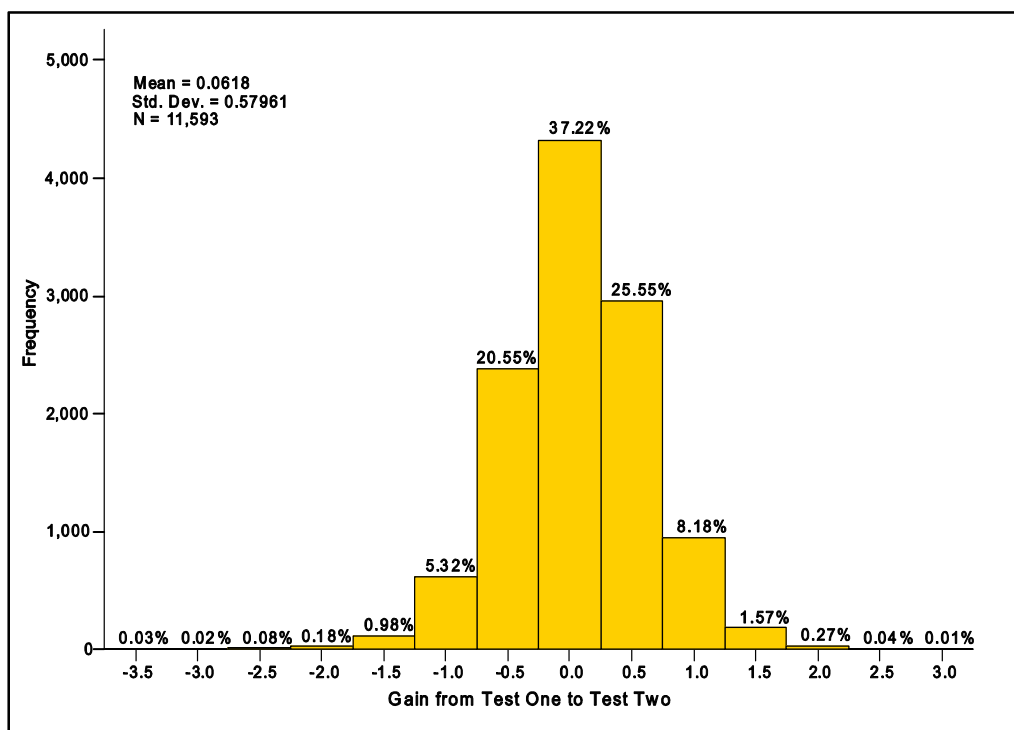
Conclusion

We hope this study can be used as a framework for evaluating reliability for other large-scale writing assessments. Although much testing research focuses on providing evidence of validity, validation is not possible without attaining reliability. This study helps to fill in a gap in the reliability research for one writing assessment—the AWA. Although the AWA is a component of an admission test unique to graduate management education, the complex manner in which the assessment is presented and scored is comparable to other large-scale writing assessments. This study provides a multifaceted approach to evaluating reliability that can be applied to other writing assessments.

While the results of the two G study designs examined here were promising, they may overestimate the reliability of the AWA. Previous research has shown that rating and hidden rater designs are useful methods for evaluating reliability with messy, large-scale performance assessments (Gao et al., 2000; Haertel, 2006; Lee and Kantor, 2005; Wang et al., 2007). This research, however, has also shown that the G design selected can affect the interpretation and accuracy of results.

To determine the impact of hiding or not modeling the rater facet, we calculated an estimate of inter-rater reliability. The inter-rater reliability was as high as 0.88. We may infer that the error variance due to raters is very small. This supports the use of two G study designs reported previously. The test-retest reliability was also high (0.78), although it is lower than the average G coefficients reported in the two G studies.

We believe the use of the automated essay scoring (AES) process has had a positive impact on the reliability coefficients reported in these studies. The AES is first programmed for each prompt, using essays that have been scored by experienced raters. Then, the AES system applies the rules it has learned when scoring subsequent essays. If these rules are correct, the computer tends to implement consistently. As such, we are not surprised that our results are an improvement over the reliability coefficients reported from other comparable tests.

Limitations and Future Research

As with all research, the current study has some limitations that we should mention. First, occasion was not included as a facet in this study. The computer-adaptive format of the GMAT exam allows for on-demand testing. Within certain parameters, examinees can choose from a variety of test dates, times, and locations to suit their testing needs. As a result, testing occasion could influence examinee performance on the AWA section. There could be a differential impact for examinees testing in the morning versus the afternoon or for weekday compared to weekend test occasions. As with the rater effect in the one-facet model, the occasion facet was hidden in this study. Future research could include this facet in G study designs to determine the influence of occasion on performance.

Also, because prompts and task type (i.e., AI and AA) are confounded with one another in the AWA, the impact of each cannot be separated from the overall task impact. Thus, estimated variance components for the task facet represent the impact of both the two prompts and two task types. The use of data from a variety of prompt pairings allows for comparisons across different AI and AA prompts. Due to limitations in sample size, however, the nine samples are not completely independent regarding prompts and raters. Some samples use one of the same prompts and/or some of the same raters. Future research with a larger sample of test takers could be used to examine subsets where all levels of facets are unique.

Finally, the repeater data used in the test-retest reliability estimation is not an optimized design. It needs to be replicated with a representative sample, although the statistics show that the current sample is good enough for the purposes of this study. Future research should address these limitations.

Contact Information

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

Acknowledgements

The views and opinions expressed in this paper are those of the authors and do not necessarily reflect those of the Graduate Management Admission Council.

References

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). *Writing assessment in admission to higher education: Review and framework.* (Report No. 96-12R). New York: College Entrance Examination Board.

Gao, X., Wang, L., & Brennan, R. L. (2000, April). *Using a hidden rater facet in generalizability analyses of a performance assessment: An empirical evaluation.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Graduate Management Admission Council® (n.d). *The GMAT® analytical writing assessment: An introduction.* [Brochure]. Santa Monica, CA: Author.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). (pp. 65-110). Westport, CT: Praeger.

Lee, Y., & Kantor, R. (2005). Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes. *ETS® Monograph Series, MS-3*(ISSN No. 1556-9021).

Livingston, S. (2004). An interesting problem in the estimation of scoring reliability. *Journal of Educational and Behavioral Statistics*, 29(3), 333–341.

Wang, L., Zhang, Y., & Li S. (2007). *Evaluating the effects of excluding the rater facet in a special generalizability application.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.