GRADUATE MANAGEMENT
ADMISSION COUNCIL

Modeling Measurement Facets and Assessing Generalizability in a Large-Scale Writing Assessment

Xiaohong Gao, Robert L. Brennan, and Fanmin Guo

GMAC® Research Reports • RR-15-01 • August 12, 2015

Abstract

Measurement error and reliability are two important psychometric properties for large-scale assessments. Generalizability theory has often been used to identify sources of error and to estimate score reliability. The complicated nature of sparse matrix data collection designs in some assessments, however, can cause challenges in conducting generalizability analyses. The present study examines potential sources of measurement error associated with large-scale writing assessment scores by modeling multiple measurement components and conducting multistep analyses based on both univariate and multivariate generalizability theory. The study demonstrates how to use multiple generalizability analyses to produce approximate estimates of measurement error and reliability under complex measurement conditions when a single study design cannot capture and disentangle all measurement facets.

Introduction

The ability to identify sources of measurement error and estimate score reliability is fundamental for the proper interpretation of any measurement outcomes, including writing assessment scores (AERA, APA, NCME, 1999). Large-scale writing and other performance assessments often employ complex data collection designs, both to preserve measurement quality and to accommodate practical constraints. These data collection designs often result in a complicated mixture of missing data and partial nesting. The use of multifaceted measurement designs in large-scale assessments, therefore, can pose challenges for estimating sources of measurement error and score reliability.

The magnitudes of estimated error variances and reliability coefficients are associated with (a) how errors are defined and modeled, (b) how data are collected, (c) how assessment scores are generalized, and (d) what measurement decisions will be made. Generalizability theory (G-theory, Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) provides a comprehensive conceptual framework and powerful methodology that enables investigators to conceptualize and disentangle multiple sources of error and to evaluate the impact of changes in measurement procedures on score generalizability. G-theory has been used broadly to evaluate psychometric properties of writing and other performance assessments. In G-theory, an individual's observed score is considered an estimate of his or her universe score. Universe score is analogous to the true score in classical test theory, and is defined as the expected score over all measurement conditions. Therefore, the use of G-theory requires an investigator to identify a *universe of admissible observations* and potential sources of error (*facets*) in a *generalizability study* (or G-study), and also to define the *universe of generalization* for assessment scores associated in a *decision study* (or D-study). A G-study analysis focuses on estimating the magnitudes of variance components in the universe of admissible observations. A D-study estimates measurement error variances (*relative error variance* and *absolute error variance*) and reliability-like coefficients (*generalizability coefficient* and *dependability coefficient*) for a specific measurement procedure (e.g., a specific number of tasks and raters in the universe of generalization). Relative error variance and a generalizability coefficient are appropriate for making *relative decisions*, such as rank ordering of examinees' scores. Absolute error variance and a dependability coefficient are appropriate for making *absolute decisions*, such as whether or not examinees have achieved some level of proficiency.

The complicated nature of the data collection designs in a large-scale assessment, however, can cause two major challenges when conducting generalizability analyses: (a) how to estimate variance (and possibly covariance) components, measurement error, and reliability for a sparse matrix design; and (b) how to treat confounded or hidden facets in the estimation.

In writing and other similar performance assessments, raters and tasks often are the primary sources of potential measurement error. Substantial evidence exists to show that task sampling variability is usually the main source of measurement error. Rater differences, on the other hand, contribute relatively little to score variability for well-designed assessments with well-trained raters (Brennan, Gao, & Colton, 1995; Gao, Shavelson, & Baxter, 1994; Lane, Stone, Ankenmann, & Liu, 1992; Linn & Burton, 1994). In other investigations of task sampling variability, researchers have found several factors that could lead to a large prompt or task sampling variability in writing and other performance assessments. One such factor is domain misspecification associated with an inappropriately broad or narrow specification of a content domain. Another factor is inappropriate treatment of a fixed facet as random. Either situation could lead to an over- or underestimation of error variance and score generalizability (Keller, Clauser, Swanson, Harik, & Clyman; 2000; Shavelson, Gao, & Baxter, 1993). In addition, a large person-by-task (*pt*) interaction (which is present in most studies) may contain a hidden person-by-task-by-occasion (*pto*) interaction effect, especially since examinees typically perform assessment tasks on a single occasion (Cronbach, Linn, Brennan, & Haertel; 1997; Shavelson, Ruiz-Primo, & Wiley, 1999).

Often in large-scale writing assessments, raters are assigned randomly to score essays, which typically leads to sparse data matrices. To estimate measurement error and reliability under a sparse data matrix, one of three procedures typically is used: (a) samples of data are selected to obtain fully crossed designs, (b) average ratings are used to simplify generalizability analyses when rater variability is negligible, or (c) a rating facet is introduced to replace a rater facet.

The phrase “rater facet” refers to a facet in which all raters are essentially treated as being indistinguishable in generalizability analyses. The most common example occurs when raters are all persons who have similar training on the score rubrics. The use of a rating facet instead of a rater facet makes it possible to assess differences associated with ratings when raters are randomly assigned to score examinees and prompts. The two facets, however, may consist of conditions that are potentially different in a substantive manner.

Specifically, the phrase “rating facet” has come to mean a facet that consists of conditions that are potentially different. A common example is a facet in which the ratings may come from a trained person or a computer scoring system. In such a case, variability associated with a rating facet involves some unknown amount of confounding of variability attributable to raters (i.e., persons) and variability attributable to applications of a computer algorithm. Previous research has reported that rater variability is slightly underestimated and generalizability is slightly overestimated with designs that use a rating facet as opposed to a rater facet (Lee & Kantor, 2005; Wang, Zhang, & Li, 2007).

When multiple tasks are nested within fixed content categories or when multiple prompts are nested within fixed essay types, measurement precision can also be evaluated using multivariate generalizability theory (Brennan, 2001; Cronbach et al., 1972; Jarjoura & Brennan, 1983). This approach assesses both the distinct sources of errors associated with each measure (e.g., content domain or essay type) and the correlated errors that exist between the measures. A multivariate generalizability analysis deconstructs the total observed score variance-covariance matrix into separate matrices of variance-covariance components for universe scores and potential error sources. Such analysis also considers universe-score interrelations among the multiple measures such as content categories or essay types.

For example, the variance-covariance matrices for a $p^{\bullet} \times o^{\bullet}$ multivariate G-study design, in which examinees are tested on multiple occasions (*o*) for each of two content categories (v and v') are:

$$\Sigma_p = \begin{bmatrix} \sigma_v^2(p) & \sigma_{vv'}(p) \\ \sigma_{vv'}(p) & \sigma_{v'}^2(p) \end{bmatrix} \qquad (1)$$

$$\Sigma_o = \begin{bmatrix} \sigma_v^2(o) & \sigma_{vv'}(o) \\ \sigma_{vv'}(o) & \sigma_{v'}^2(o) \end{bmatrix} \quad \text{and} \qquad (2)$$

$$\Sigma_{po} = \begin{bmatrix} \sigma_v^2(po) & \sigma_{vv'}(po) \\ \sigma_{vv'}(po) & \sigma_{v'}^2(po) \end{bmatrix}. \qquad (3)$$

Note that “$\bullet$” represents joint (as linked) sampling of content categories across occasions.

In addition, a multivariate generalizability analysis can estimate measurement precision for composite scores across multiple measures. Specifically, the variance of composite universe scores is

$$\sigma_C^2(\tau) = \sum_{v=1}^{n_v} w_v^2 \sigma_v^2(\tau) + \sum_{v \neq v'}\sum w_v w_{v'} \sigma_{vv'}(\tau). \qquad (4)$$

where n_v is the number of measures, and the w_v and $w_{v'}$ are usually proportional weights summing to unity.

The measurement error and generalizability of composite scores depends not only on the universe of generalization and type of decision (absolute or criterion-referenced vs. relative or norm-referenced) but also on the relationship between measures (linked or jointly sampled vs. independently sampled), for example, letting “$\bullet$” represent linked sampling and “$\circ$” represent independent sampling. In a $p^{\bullet} \times i^{\circ}$ multivariate generalizability analysis, the relative error variance of average scores for an independent sampling of prompts (*i*) within each essay type is

$$\sigma_C^2(\delta) = \sum_{v=1}^{n_v} w_v^2 \sigma_v^2(\delta) = \sum_{v=1}^{n_v} w_v^2 \left[\frac{1}{{}_v n'_i} \sigma_v^2(pi) \right]. \qquad (5)$$

where ${}_{v}n'_{i}$ is the number of prompts within each fixed category. The absolute error variance of average scores is

$$\sigma_C^2(\Delta) = \sum_{v=1}^{n_v} w_v^2 \sigma_v^2(\Delta) = \sum_{v=1}^{n_v} w_v^2 \left[\frac{1}{{}_v n'_i}\sigma_v^2(i) + \frac{1}{{}_v n'_i}\sigma_v^2(pi)\right]. \qquad (6)$$

When a facet is jointly sampled, both the error variances and covariances for the facet contribute to total error variances. For example, in the $p^{\bullet} \times o^{\bullet}$ multivariate generalizability analysis, with all examinees assessed on both occasions, the relative error variance of composite scores across occasions is

$$\sigma_C^2(\delta) = \sum_{v=1}^{n_v} w_v^2 \sigma_v^2(\delta) + \sum_{v \neq v'}\sum w_v w_{v'} \sigma_{vv'}(\delta). \qquad (7)$$

The absolute error variance is

$$\sigma_C^2(\Delta) = \sum_{v=1}^{n_v} w_v^2 \sigma_v^2(\Delta) + \sum_{v \neq v'}\sum w_v w_{v'} \sigma_{vv'}(\Delta). \qquad (8)$$

No matter what the nature of the sampling (joint or independent), the generalizability coefficient for composite scores is

$$E\rho^2 = \frac{\sigma_C^2(\tau)}{\sigma_C^2(\tau) + \sigma_C^2(\delta)}, \qquad (9)$$

and the dependability coefficient for composite scores is

$$\Phi = \frac{\sigma_C^2(\tau)}{\sigma_C^2(\tau) + \sigma_C^2(\Delta)}. \qquad (10)$$

The present study examines potential sources of measurement error in a large-scale writing assessment with a multifaceted measurement design. The Graduate Management Admission Test® (GMAT®) Analytical Writing Assessment (AWA) test data were used for this analysis. It should be noted that the Graduate Management Admission Council® (GMAC®) subsequently changed the test design in 2012. In previous GMAT AWA reliability research, Breland, Bridgeman, and Fowles (1999) reported Cronbach's alpha estimates ranging from .66 to .79 for the paper version of the GMAT test administered in 1995. After the test was moved from paper-and-pencil mode to a computer adaptive environment, measurement conditions also changed. Siegert and Guo (2009) conducted generalizability analyses using two G-study designs and compared the results across multiple data samples. Specifically, they used a *p x (r′:t)* design with a rating (*r′*) facet and a *p x t* design with a hidden rater (*r*), and treated both ratings and essay types (*t*) as random. The analysis involved two essay types, Analysis of an Issue (AI) and Analysis of an Argument (AA). They found that average difficulty levels of the two GMAT AWA essay types and average ratings from a scoring engine and human raters were similar. The variance component for the person-by-essay type (*pt*) interaction was notable, however, suggesting that the rank orderings of examinees varied somewhat between the essay types. The average generalizability coefficient was .835 based on these two designs. In addition, Siegert and Guo (2009) computed both inter-rater reliability and test-retest reliability and reported them as .88 and .78, respectively. These values, however, are likely upper-bound reliability estimates given that each design did not model all effects or potential sources of error associated with the GMAT AWA measurement procedure, such as prompts within essay types, occasions, and raters.

The present study examines potential sources of measurement error associated with GMAT AWA scores by modeling multiple measurement components confounded in a sparse data matrix, which is very common in large-scale assessments. The complexity of the measurement procedures in these assessments does not permit the use of a single generalizability study design to disentangle the multiple sources of error that affect reliability. The study thus demonstrates how to use multiple generalizability analyses based on both univariate and multivariate G-theory approaches to model multiple measurement facets and how to provide approximate estimate of measurement error and reliability under complex measurement situations.

Methodology

Data and Instrument

The data for this study were collected during the GMAT AWA administration in testing year 2009–2010. More than 28,000 examinees who took the test for a second time (repeaters) were included in all of the generalizability analyses reported here.

The GMAT AWA assessment was developed to assess analytical reasoning skills through writing. Before the redesign of the test in 2012, examinees were asked to construct two essays: one to explain their viewpoint in responding to an issue (AI), and the other to investigate and critique the reasoning behind an argument (AA). Test takers had 30 minutes to complete each of the essays. Prompts for each essay type were selected randomly from each of their corresponding essay pools.

GMAT AWA essays were rated by a human rater and an automated essay scoring engine (AES) developed by Vantage Learning. The human raters went through a standard training process before starting the ratings.

The AES was calibrated with human-scored essays for each prompt before it was deployed. Occasionally, for quality assurance purposes, examinees' essays were rated by two human raters. Both human raters and the AES rated essays on a scale of 0 to 6 with increments of one. If the two original (unadjusted) ratings for a given essay differed by more than one point, an experienced human rater would provide a third rating for adjudication (i.e., adjusted rating) to replace both of the initial ratings. Prompt scores were calculated as the means of the adjusted ratings on the essays, and the final AWA reported score was the average of the prompt scores, rounded up to 0.5 from 0.25 or to 1.0 from 0.75. Therefore, the GMAT AWA scores could contain measurement errors from multiple sources, including sampling errors from prompts, raters, and occasions. All analyses reported here are based on unadjusted ratings (i.e., no adjudication) from the original computer and human raters.

Generalizability Analyses

Objects of measurement and universe of admissible observations. Since the GMAT AWA measures individual analytical writing proficiencies, examinees (*p*) constitute the objects of measurement. All examinees were randomly administered one prompt from each of the two essay types, AI and AA. Each essay is scored by a human rater and an automated essay scoring engine (AES). Examinees could take the test more than once on different occasions. Therefore, the universe of admissible observations here can be conceptualized as containing at least four facets: *prompts or items* (*i*), *essay types* (*t*), *raters* (*r*) or *ratings* (*r′*), and *occasions* (*o*). In the idealized universe of admissible observations, each of these facets is random, except for essay type, which has two fixed levels (AI and AA).

For the current operational measurement procedure, relationships among these facets and the objects of measurement are very complicated, as specified below:

1. The assessment contains two *essay types* (*t*) and each has different *prompts* (*i*); i.e., (*i:t*);
2. Examinees take both essay types; i.e., (*p* x *t*);
3. Different items are randomly administered to different examinees; i.e., (*i:p*);
4. Each person receives two *ratings* (*r'*) for each prompt; i.e., (*p* x *r′*);
5. Different human *raters* (*r*) score different examinees; i.e., (*r:p*);
6. Most human raters are trained to score only one essay type; i.e., (*r:t*);
7. Raters may score any of the administered prompts within an essay type; i.e., [(*r* x *i):t*];
8. Examinees may take the test on multiple *occasions* (*o*); i.e., (*p* x *o*);
9. The same essay types are administered on each occasion; i.e., (*t* x *o*);
10. Raters often score the same essay type across different occasions; i.e., [(*r:t)* x *o*];
11. Since only one prompt within each essay type is administered to an examinee, prompt and prompt types are confounded; i.e., (*i, t*); and
12. Since only one prompt of each essay type is administered on each occasion, prompt and occasion are also confounded; i.e., (*i, o*).

Due to the complexity of the GMAT AWA measurement procedure, the objects of measurement were not fully crossed with all facets—some facets were nested (e.g., *r:t*), and others were confounded (e.g., *i* and *o*). Also, some facets, such as raters, prompts, and occasions, were random but others were fixed—e.g., only two essay types were administered. Consequently, given available data, the measurement conditions cannot be modeled by a single completely crossed or nested G-study design. The available data from the measurement procedure also precluded obtaining nonconfounded and unbiased estimates of each possible variance component in the universe of admissible observations. Nevertheless, as Brennan (1992, p. 122) pointed out, "from a practical viewpoint it is quite difficult to conduct an extensive G study. However, it is possible, in principle, to estimate variance components for a universe of admissible observations using multiple G studies, any one of which involves only a subset of facets in the full universe."

In the present study, multiple generalizability analyses were conducted that, in total, give a picture of the contributions of facets to error variances and coefficients. Final results are reported in terms of ranges of likely values of standard errors of measurement and coefficients, since it cannot be argued convincingly that any single analysis is clearly preferable to all others. *GENOVA* (Crick & Brennan, 1983) was used for all univariate generalizability analyses and mGENOVA (Brennan, 1999) was used for the multivariate generalizability analyses in the study.

Univariate G-study designs. Since each item belonged to just one specific essay type, a complete design reflecting the universe of admissible observations with all four facets would be: *p* x *r* x *(i:t)* x *o*. Since examinees took only one prompt within each essay type on each occasion, however, facets *i* and *o* were confounded. A reduced design could be *p* x *r* x *t* x *o*. Moreover, as the prompts were from two types of essays, AI and AA, strictly speaking, the *t*-facet was fixed. Due to the fact that (a) raters were not fully crossed with persons and essay types, (b) a single item was used for each essay type and on each occasion, and (c) prompts were randomly assigned to examinees, the following three reduced G-study designs were considered for the univariate generalizability analyses: *r:(p x t) x o*, *p x (r′ : t) x o*, and (*i x r):(p x t)*.

In total, these three designs covered all relevant facets in the universe of admissible observations. Each design involved a subset of facets and disentangled some unique main sampling effects and interactions that the other designs did not (see Appendix A.). The Venn diagrams for the designs are shown in Figures 1 to 4. It should be pointed out, however, that "nesting" (:) in these designs is not strictly true and does not fully characterize the data. It also should be noted that the rater facet (*r*) was used under the first and third designs in which raters were nested within examinees (*p*) and essay types (*t*). But the rating facet (*r′*) was used under the second design since all examinees received the same number of ratings.

Figure 1. Venn Diagram for a p x r x t x o Design

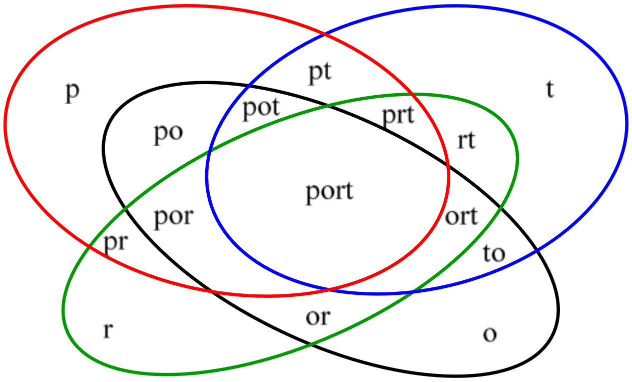


Figure 2. Venn Diagram for a *r:(p x t) x o* Design

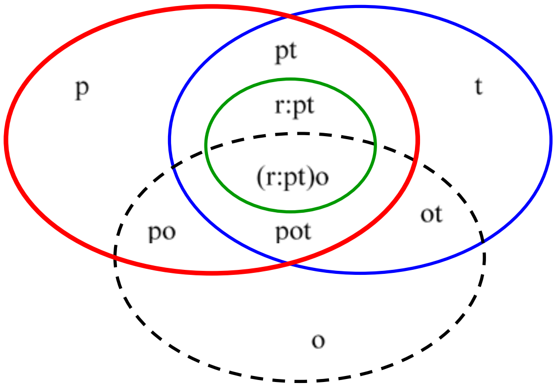


Figure 3. Venn Diagram for a p x (r’: t) x o Design

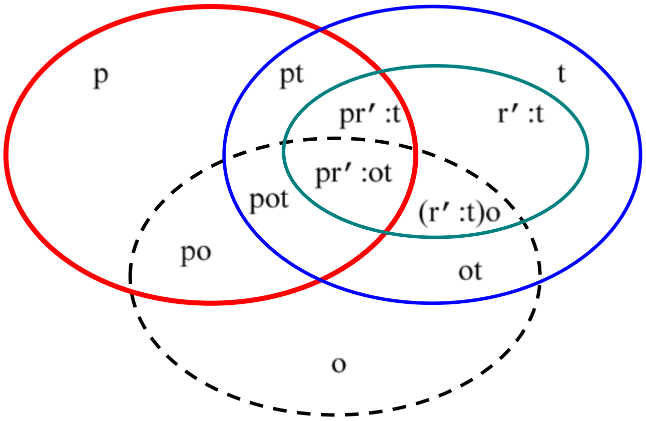
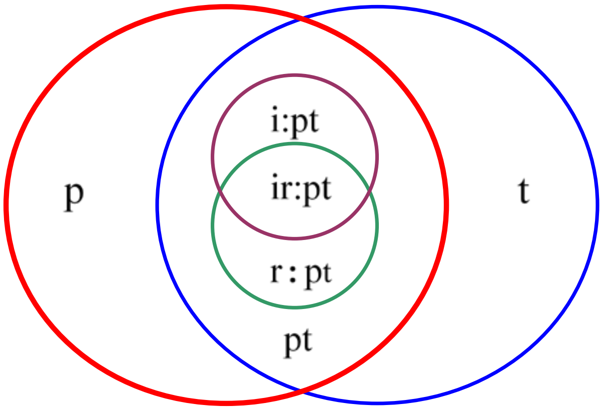


Figure 4. Venn Diagram for a (*i* x *r*)*:(p* x *t)* Design



In addition, several two-facet crossed and nested designs were considered for additional univariate generalizability analyses. In many large-scale performance assessments, examinees take a test only once, respond to only one type of essay prompt, and have their responses scored by either human raters or an automated scoring engine. These two-facet analyses evaluated consistency of the estimated variance components across occasions, between essay types, or between computer and human ratings. Specifically, within each occasion, an *r:(p x t)* design was used to assess consistency of estimated sampling variabilities associated with raters and essay types *between occasions (o)*. A ***p* x *o* x *r′*** design was used to evaluate consistency of estimated sampling variabilities associated with occasions and ratings *between essay types (t)*. The third design, *i:(p x t),* was used to examine consistency of estimated sampling variabilities associated with prompts and essay types *between computer and human ratings* (*r′*). It should be noted that although these three designs together cover all relevant facets in the *p x r x (i:t) x o* design, each individual design involved only two main facets, with the other facets either hidden or confounded.

Multivariate G-study design. Multivariate generalizability analyses permit examination of universe score correlations and correlated errors across different measures, as well as estimation of error variances and generalizability of combined composite scores. In the GMAT AWA, there were two essay types, AI and AA, and they could be viewed as two fixed conditions for a multivariate generalizability analysis. More specifically, examinees had two universe scores, one for AI and the other for AA. A random-effects design also was associated with each essay type.

Specifically, a $\boldsymbol{p}^{\bullet} \times \boldsymbol{o}^{\bullet} \times \boldsymbol{r'}^{\circ}$ multivariate G-study design was used, in which examinees took the same essay types across multiple occasions (*o*) and received the same number of ratings (r') on each occasion. Both examinees and occasions were considered jointly sampled across the essay types ($\bullet$). Therefore, the universe scores and errors associated with occasions were linked across the essay types through covariance components for persons and occasions. Since different raters scored different essay types, however, *r'* was considered independent ($\circ$) between the essay types. Original unadjusted ratings were used in the analysis.

Results

Univariate Generalizability Analyses

Three-facet designs. Univariate generalizability analyses under the ***r:(p x t) x o*** and ***p x (r':t) x o*** designs provide estimates of sampling variability (i.e., *variance components*) associated with raters (*r*) or ratings (*r'*), essay types (*t*), and occasions (*o*), in addition to examinees (*p*) or universe score variance for the GMAT AWA scores. Tables 1 and 2 present the estimated G-study variance components $\hat{\sigma}^2(\alpha)$, D-study error variances $\hat{\sigma}^2(\delta)$ and $\hat{\sigma}^2(\Delta)$, and generalizability coefficients, $E\hat{\rho}^2$ and $\hat{\phi}$.

Note that these tables report the D-study results for essay types as random and fixed. Although it is common practice to treat essay types as random in generalizability analysis, it should be recalled that for the GMAT AWA assessment, essay types are in fact fixed, since each form of the test contains the same two essay types. What makes the analyses challenging to interpret is the fact that there is only one operational prompt for each essay type. This means that prompt and essay type are completely confounded in the operational data, which is often denoted (*i, t*). To make matters even more complicated, prompts are random (since there are different prompts per form) yet essay types are fixed. Thus, there is a sense in which estimates (not parameters) of error variance and coefficients could legitimately be based on analyses in which *t* is random or fixed. Fortunately, as indicated in these two tables and as will be discussed below, this very complicated matter seems to make little difference for GMAT AWA.

Results from the two generalizability analyses indicate that (1) variances associated with an occasion effect, $\hat{\sigma}^2(o)$, and essay type effect, $\hat{\sigma}^2(t)$, were very small and close to zero, respectively, suggesting that mean differences were very small between the two occasions and the two essay types; (2) interactions between occasions and essay types, $\hat{\sigma}^2(ot)$, also were very small; (3) there was some variation between the computer and human ratings confounded by interactions between ratings and essay types, since $\hat{\sigma}^2(r':t)$ is not negligible; (4) interactions between examinees and measurement facets existed, such as $\hat{\sigma}^2(po)$, $\hat{\sigma}^2(pt)$, $\hat{\sigma}^2(pot)$, $\hat{\sigma}^2(r:pt)$, and $\hat{\sigma}^2(pr':t)$, indicating that the rank orderings of the examinees might vary across occasions, between essay types, and between raters or ratings; and (5) the largest source of error came from the residuals, $\hat{\sigma}^2(r:pot)$ or $\hat{\sigma}^2(pr':ot)$.

It should be pointed out these variance component estimates involve complex confounding since just one prompt (*i*) was used within each essay type and on each occasion. Thus, any effects associated with occasions or essay types contain error from prompt sampling. In addition, the results show that estimates of error variances and generalizability coefficients were not very different for norm-referenced and domain-referenced measurement decisions since the interactions between examinees and measurement facets were the major sources of error. Due to nesting, however, some of these interactions were confounded with main effects (see Appendix A). Also noteworthy is the similarity of the total estimated variances for the two designs. Most of the individual estimates were similar except for the variance component $\hat{\sigma}^2(pt)$.

Figure 5 shows the percentages of total variance contributed by each component across the two designs. The results indicate that the use of a rater or rating facet has little impact on the estimated variance components, although the ***p x (r': t) x o*** generalizability analysis provides more distinguishable variance component estimates than the ***r:(p x t) x o*** analysis due to different levels of nesting (see Appendix A). Estimates of error variances and generalizability coefficients also were similar between the ***R:(p x T) x O*** design and the ***p x (R':T) x O*** design. Under the current measurement conditions with two ratings (or raters) per essay, two essay types, and a single test administration, the estimated standard error of measurement (both δ and Δ types) was about 0.40 and the estimated generalizability coefficient was about 0.76 for norm-referenced decisions with *T* as random or fixed.

Table 1. Estimated Variance Components, Error Variance, and Generalizability Coefficients Under the *r:(p x t) x o* Design

Sources	$\hat{\sigma}^2(\alpha)$	Percentage	n'_o	n'_t	n'_r	$\hat{\sigma}^2(\delta)$	$\hat{\sigma}^2(\Delta)$	$E\hat{\rho}^2$	$\hat{\phi}$
r:(p x t) x o G-Study			*R:(p x T) x O* Random-Effects D-studies						
p	0.502	56.34%	1	1	1	0.386	0.389	0.565	0.563
r:pt	0.062	6.94%	1	1	2	0.260	0.263	0.659	0.656
t	0.000	0.00%	1	2	1	0.221	0.223	0.694	0.692
o	0.002	0.25%	**1**	**2**	**2**	**0.158**	**0.160**	**0.761**	**0.758**
pt	0.002	0.24%	2	1	1	0.225	0.227	0.690	0.689
po	0.056	6.23%	2	1	2	0.146	0.148	0.774	0.772
ot	0.001	0.06%	2	2	1	0.126	0.128	0.799	0.797
pot	0.076	8.52%	2	2	2	0.087	0.088	0.852	0.850
r:pot	0.191	21.41%	*R:(p x T) x O* D-Studies With *T* Fixed						
Total	**0.891**		1	2	1	0.220	0.222	0.696	0.694
			1	**2**	**2**	**0.157**	**0.159**	**0.763**	**0.760**
			2	2	1	0.125	0.127	0.801	0.799
			2	2	2	0.086	0.087	0.854	0.852

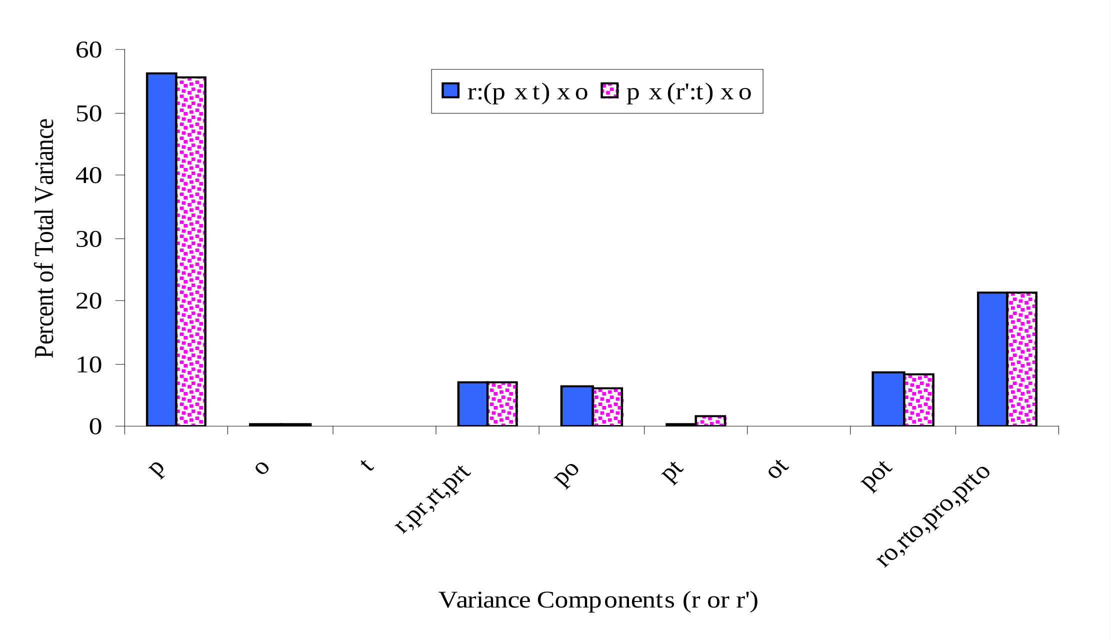
Note. Bold values are the estimates under the operational conditions.

Table 2. Estimated Variance Components, Error Variance, and Generalizability Coefficients Under the *p x (r′ :t) x o* Design

Sources	$\hat{\sigma}^2(\alpha)$	Percentage	n'_o	n'_t	n'_r	$\hat{\sigma}^2(\delta)$	$\hat{\sigma}^2(\Delta)$	$E\hat{\rho}^2$	$\hat{\phi}$
p x (r′:t) x o G-Study			p x (R′:T) x O D-Study						
p	0.502	55.61%	1	1	1	0.374	0.401	0.573	0.556
r′:t	0.023	2.59%	1	1	2	0.260	0.274	0.659	0.647
t	0.000	0.00%	1	2	1	0.215	0.229	0.700	0.687
o	0.002	0.25%	**1**	**2**	**2**	**0.158**	**0.166**	**0.761**	**0.751**
p r′:t	0.038	4.26%	2	1	1	0.213	0.238	0.702	0.678
pt	0.014	1.53%	2	1	2	0.146	0.160	0.774	0.759
po	0.056	6.15%	2	2	1	0.121	0.134	0.806	0.790
r′:ot	0.000	0.00%	2	2	2	0.087	0.094	0.852	0.842
ot	0.001	0.06%	*p x (R′:T) x O* D-Study With *T* Fixed						
ot	0.076	8.41%	1	2	1	0.208	0.222	0.710	0.696
p r′:ot	0.191	21.13%	1	2	2	**0.151**	**0.159**	**0.771**	**0.762**
Total	**0.903**		2	2	1	0.114	0.127	0.817	0.801
			2	2	2	0.080	0.087	0.864	0.854

Note. Bold values are the estimates under the operational conditions.

Figure 5. Percentage of Total Variance Contributed by Each Effect From the *r:(p* x *t)* x *o* and *p* x *(r′ :t)* x *o* Designs (*r* = Raters and *r′* = Ratings)



The *(i* x *r):(p* x *t)* univariate generalizability analysis considered essay types (*t*) and raters (*r*), in addition to the *i*-facet (prompts), which were considered instead of the *o*-facet (these two facets are confounded in the data collection). **Table 3** presents the estimated variance components associated with these facets as well as error variances and generalizability coefficients associated with different measurement procedures. The results indicate that the estimated variance components $\hat{\sigma}^2(i:pt)$ and $\hat{\sigma}^2(ir:pt)$ were the major sources of error, while $\hat{\sigma}^2(t)$ $\hat{\sigma}^2(pt)$ were near zero, suggesting that prompt sampling (within essay times) is more variable than essay type. Another finding showed the existence of rater effects and/or interactions, given that $\hat{\sigma}^2(r:pt)$ was not negligible. The estimated relative standard error of measurement was 0.36 and the estimated generalizability coefficient was 0.80 for relative decisions with one prompt within each of the two essay types and two raters.

It should be noted that the *i:pt* component in the generalizability analysis represents the confounding of the *pi*, *it*, and *pit* interactions in addition to the *i* effect. On the other hand, since prompts and occasions were confounded in all three designs, any estimated variance components associated with prompts (*i*) could also contain occasions (*o*) variance or vice versa. Actually, the magnitude of the estimated variance component for *i:pt* was similar to the sum of the estimated variance components for *o*, *po*, *ot*, and *pot* from the *r:(p* x *t)* x *o* and *p* x *(r′: t)* x *o* generalizability analyses.

Moreover, it is noteworthy that the total estimated variance and the universe score variance were slightly higher for the *(i* x *r):(p* x *t)* generalizability analysis than for the *r:(p* x *t)* x *o* or *p* x *(r′: t)* x *o* generalizability analyses when the *i* or *o* facet was considered, but the corresponding variance components associated with error variances were similar. **Figure 6** presents the percentages of total variance contributed by each component across the *(i* x *r):(p* x *t)* and *r:(p* x *t)* x *o* designs with *i* and *o* used interchangeably in the graph since they are confounded. Specifically, the variance components from *(i* x *r):(p* x *t)* are presented in **Figure 6** and the corresponding variance components from *r:(p* x *t)* x *o* are collapsed to produce corresponding variance components between the two designs (see Appendix A).

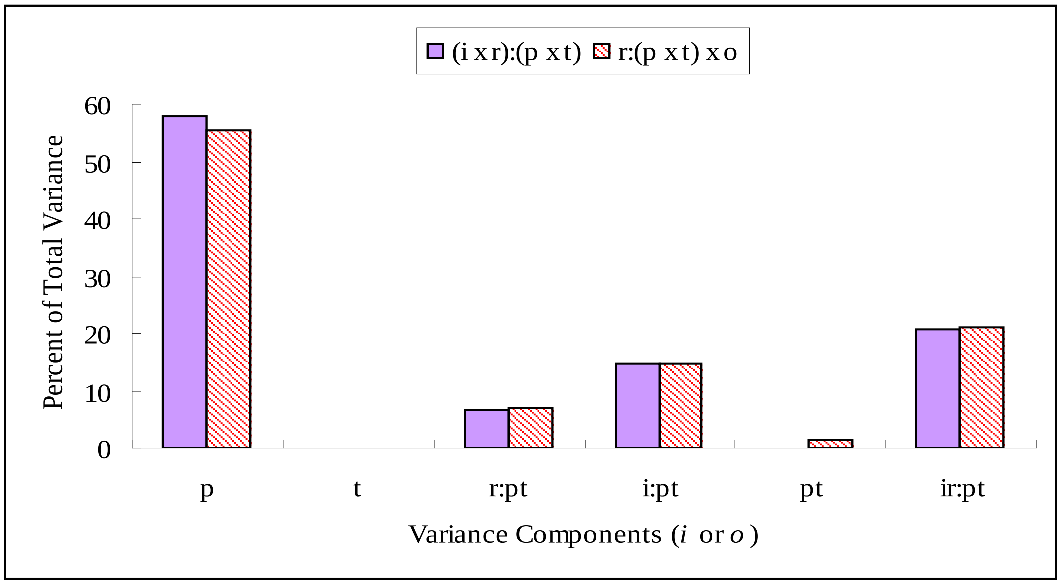
Two-facet designs. To evaluate consistency of the variance component estimates across occasions, essay types, and rating types, several two-facet generalizability analyses were performed. The results of separate *r:(p* x *t)* generalizability analyses for Occasion 1 and Occasion 2 are reported in **Table 4**. The pattern of the variance component estimates was similar for both occasions. The largest error source came from the residual $\hat{\sigma}^2(r:pt)$, which included *r*, *pr*, *rt*, and *prt* within each occasion. A small estimated $\hat{\sigma}^2(t)$ and a relatively large $\hat{\sigma}^2(pt)$ suggests that the average ratings for the two essay types were similar, although rank orderings of examinees could differ between them. Since there was only one prompt per type, however, variability among prompts could also contribute to the *pt* interaction. In addition, effects associated with occasion were hidden (e.g., *po*, *por*, and *pot*).

Table 3. Estimated Variance Components, Error Variance, and Generalizability Coefficients Under the *(i x r):(p x t)* Design

Sources	$\hat{\sigma}^2(\alpha)$	Percentage	n'_t	$n'_{i:pt}$	$n'_{r:pt}$	$\hat{\sigma}^2(\delta)$	$\hat{\sigma}^2(\Delta)$	$E\hat{\rho}^2$	$\hat{\phi}$
(i x r):(p x t) G-Study			*(I x R):(p x T) D-Study*						
p	0.530	57.79%	1	1	1	0.387	0.387	0.572	0.572
t	0.000	0.02%	1	1	2	0.261	0.261	0.665	0.664
i:pt	0.134	14.64%	1	2	1	0.224	0.224	0.697	0.697
r:pt	0.062	6.74%	1	2	2	0.146	0.146	0.780	0.780
pt	0.000	0.00%	2	1	1	0.193	0.193	0.727	0.727
ir:pt	0.191	20.81%	**2**	**1**	**2**	**0.130**	**0.130**	**0.799**	**0.798**
			2	2	1	0.112	0.112	0.822	0.821
			2	2	2	0.073	0.073	0.876	0.876
Total	**0.917**								

Note. Bold values are the estimates under the operational conditions.

Figure 6. Percentage of Total Variance Contributed by Each Effect From the *(i x r):(p x t)* and *r:(p x t) x o* Designs (*i* = Prompts and *o* = Occasions)



As seen previously, the estimates for error variances and generalizability coefficients were very similar for both norm-referenced and domain-referenced decisions as well as between the two occasions. Table 5 reports the estimated variance components, error variances, and generalizability coefficients for each essay type (AI and AA) under a ***p* x *o* x *r′*** design. The results indicate that the major sources of error came from *po* and *por´*. Since only one prompt was used on each occasion, it is possible prompt sampling might have contributed to these interactions. These values may also be considered as estimates for the ***p* x *i* x *r′*** design with one prompt and two ratings for each essay type since the *i*-facet and *o*-facet are confounded here. In addition, interactions between essay types and occasions were hidden. It should also be noted that the universe score variance $\hat{\sigma}^2(p)$ and dependability coefficient for AI were slightly higher than those for AA. Since only one prompt was used on each occasion, however, prompt sampling also might have contributed to these interactions. The estimated standard errors of measurement were 0.50 for AI and 0.51 for AA and dependability coefficients were 0.68 for AI and 0.65 for AA.

Table 4. Estimated Variance Components, Error Variance, and Generalizability Coefficients Under the ***r:(p x t)*** Design for Occasion 1 and Occasion 2

r:(p x t) G-Study					
Sources	$\hat{\sigma}^2(\alpha)$	**Percentage**	**Sources**	$\hat{\sigma}^2(\alpha)$	**Percentage**
	Occasion 1			***Occasion 2***	
p	0.571	63.01%	*p*	0.544	62.46%
r:pt	0.251	27.73%	*r:pt*	0.254	29.14%
t	0.001	0.10%	*t*	0.000	0.00%
pt	0.083	9.16%	*pt*	0.073	8.40%
Total	**0.906**		Total	**0.871**	
R:(p x T) D-Study					
n'_t	$n'_{r:pt}$	$\hat{\sigma}^2(\delta)$	$\hat{\sigma}^2(\Delta)$	$E\hat{\rho}^2$	$\hat{\phi}$
Occasion 1					
1	1	0.334%	0.335	0.631	0.630%
2	1	0.167%	0.168	0.774	0.773%
1	2	0.209%	0.209	0.732	0.732%
2	**2**	**0.104%**	**0.105**	**0.846**	**0.845%**
Occasion 2					
1	1	0.327%	0.327	0.625	0.625%
2	1	0.164%	0.164	0.769	0.769%
1	2	0.200%	0.200	0.731	0.731%
2	**2**	**0.100%**	**0.100**	**0.845**	**0.845%**

Note. Bold values are the estimates under the operational conditions.

Table 5. Estimated Variance Components, Error Variance, and Generalizability Coefficients Under the ***p x o x r'*** Design for Each Essay Type

***p x o x r'* G-Study**					
Sources	$\hat{\sigma}^2(\alpha)$	**Percentage**	**Sources**	$\hat{\sigma}^2(\alpha)$	**Percentage**
	T1 (AI)			***T2 (AA)***	
p	0.528	58.81%	*p*	0.503	55.5%
r'	0.015	1.69%	*r'*	0.032	3.47%
o	0.001	0.12%	*o*	0.005	0.50%
p r'	0.039	4.38%	*p r'*	0.038	4.14%
po	0.134	14.96%	*po*	0.128	14.17%
or'	0.000	0.01%	*or'*	0.000	0.00%
por'	0.180	20.03%	*p r'*	0.201	22.22%
Total	**0.898**		Total	**0.907**	
***p x O x R'* D-Study**					
n'_o	$n'_{r'}$	$\hat{\sigma}^2(\delta)$	$\hat{\sigma}^2(\Delta)$	$E\hat{\rho}^2$	$\hat{\phi}$

Table 5. Estimated Variance Components, Error Variance, and Generalizability Coefficients Under the *p* x *o* x *r′* Design for Each Essay Type

(Cont'd)			*T1 (AI)*		
1	1	0.354%	0.370	0.599	0.588%
1	**2**	**0.244%**	**0.253**	**0.684**	**0.676%**
2	1	0.197%	0.212	0.729	0.713%
2	2	0.132%	0.140	0.800	0.791%
			T2 (AA)		
1	1	0.368%	0.404	0.578	0.555%
1	**2**	**0.248%**	**0.268**	**0.670**	**0.652%**
2	1	0.203%	0.236	0.713	0.680%
2	2	0.133%	0.151	0.790	0.769%

Note. Bold values are the estimates under the operational conditions.

Sampling variabilities were examined for each rating type (i.e., computer vs. human) using the ***i*:(*p* x *t*)** design. The results presented in Table 6 indicate that the largest source of error was *i:pt*, which included *i*, *pi*, *it*, and *pit*, whereas the *pt* interaction was small within each type of rating. The results also suggest that the sampling variability was slightly larger for the human ratings than for the scoring engine.

Table 6. Estimated Variance Components, Error Variance, and Generalizability Coefficients Under the *i*:(*p* x *t*) Design for Computer and Human Ratings

		i:(p x t) G-Study			
Sources	$\hat{\sigma}^2(\alpha)$	**Percentage**	**Sources**	$\hat{\sigma}^2(\alpha)$	**Percentage**
	Computer			***Human***	
p	0.549	63.82%	*p*	0.541	62.83%
t	0.000	0.02%	*t*	0.002	0.18%
i:pt	0.306	35.54%	*i:pt*	0.344	39.99%
pt	0.005	0.61%	*pt*	0.011	1.23%
Total	**0.861**		Total	**0.897**	
		I:(p x T) D-Study			
n'_t	$n'_{i:pt}$	$\hat{\sigma}^2(\delta)$	$\hat{\sigma}^2(\Delta)$	$E\hat{\rho}^2$	$\hat{\phi}$
		Computer			
1	1	0.311%	0.311	0.638	0.638%
1	2	0.158%	0.158	0.776	0.776%
2	**1**	**0.156%**	**0.156**	**0.779**	**0.779%**
2	2	0.079%	0.079	0.874	0.874%
		Human			
1	1	0.355%	0.356	0.604	0.603%
1	2	0.183%	0.184	0.748	0.746%
2	**1**	**0.177%**	**0.178**	**0.753**	**0.752%**
2	2	0.091%	0.092	0.856	0.854%

Note. Bold values are the estimates under the operational conditions.

Table 7 provides a range of (a) estimated universe score variances, (b) total G-study variances, (c) sums of G-study variances contributing to relative error variances in the D-studies, (d) relative error estimates, and (e) estimated generalizability coefficients based on the three-facet and two-facet designs. Both the lowest and highest estimates are bolded in the table and the averages are presented in italics. Due to various degrees of nesting and confounding in these designs, not only can the estimates of G-study variance components and universe score variances be different, but the estimated standard errors also can vary due to different *n′* terms in the denominators in the computations. Consequently, the estimated generalizability coefficients among these analyses also vary.

Although no single design can disentangle all the variance components in the full ***p x r x (i:t) x o*** design, these estimates approximate universe score variance, measurement error, and reliability for the GMAT AWA scores and provide an overview of their measurement precision. Further, the estimates from the three-facet designs should be more reasonable than those from the two-facet designs because (a) the universe of admissible observations and universe of generalization in the three-facet design are closer to those in the full design; and (b) each of the three-facet designs involves hidden or confounded effects that distort estimates.

Table 7. Estimated Variances and Generalizability Coefficients From Different Generalizability Analyses

Design	$\hat{\sigma}^2(p)$	Total Variance[a]	Error Variance[b]	$\hat{\sigma}(\delta)$	$E\hat{\rho}^2$
Univariate Analyses					
R:(p x T) x O	**0.502**	0.891	0.386	0.400	0.760
p x (R′:T) x O	**0.502**	0.903	0.375	0.400	0.760
(I x R):(p x T)	0.530	**0.917**	0.387	0.360	0.800
R:(p x T) for Occasion 1	**0.571**	0.906	0.334	**0.320**	**0.850**
R:(p x T) for Occasion 2	0.544	0.871	0.327	**0.320**	0.840
p x O x R′ for AI	0.528	0.898	0.354	0.490	0.680
p x O x R′ for AA	0.503	0.907	0.368	**0.500**	**0.670**
I:(p x T) for Computer Scores	0.549	**0.861**	**0.311**	0.390	0.780
I:(p x T) for Human Raters	0.541	0.897	0.355	0.420	0.750
Average	*0.530*	*0.895*	*0.355*	*0.400*	*0.770*
Multivariate Analysis					
$p^{\bullet} \times O^{\bullet} \times R'^{\circ}$				0.39	0.77

[a]Total G-study variance; [b]Sum of G-study variance components contributing to relative error variance.

Note: $\hat{\sigma}^2(p)$ = estimated universe score variance; $\hat{\sigma}(\delta)$ = estimated relative standard error of measurement, and $E\hat{\rho}^2$ = estimated generalizability coefficient with n'_o = 1, n'_r = 2, n'_t = 2, and $n'_{i:t}$ = 1.

Multivariate Generalizability Analyses

The $\mathbf{p}^{\bullet} \times \mathbf{o}^{\bullet} \times \mathbf{r}'^{\circ}$ *design*. Under the multivariate generalizability theory framework, there are multiple universe scores for the objects of measurement, one for each fixed condition or measure (e.g., essay type AI or AA). Also, measurement facets can be either jointly sampled (·) or independent (°) across the fixed conditions. In the $p^{\bullet} \times o^{\bullet} \times r'^{\circ}$ design, the *o*-facet is jointly sampled ($o^{\bullet}$) across the essay types, while the *r′*-facet is independent (r'°) since different human raters scored different essay types. Thus, in addition to the variance components for the main effects and interactions under each essay type, there are covariance components associated with persons, occasions, and the person-by-occasion interactions between the essay types. Both estimated variances and covariances contribute to the universe score variance, the error variances, and generalizability coefficients for the average (composite) GMAT AWA scores.

Table 8 reports estimated G-study and D-study variances (bolded diagonals) and covariances (lower off-diagonals); universe score correlations (italic upper off-diagonals); estimated error variances and covariances for each measure or essay type; and estimated universe score variance, error variances, and generalizability coefficients for average or composite (c) scores.

Table 8. Estimated Variance and Covariance Components, Composite Score Error Variances, and Generalizability Coefficients Under the $p^{\bullet} \times o^{\bullet} \times r'^{\circ}$ Multivariate Design

G-Study	T1 - AI	T2 - AA	D-Study	T1 - AI	T2 - AA
p	**0.528**	*0.973*	*p*	**0.528**	*0.973*
	0.502	**0.503**		0.502	**0.503**
o	**0.001**		*O (n = 1)*	0.001	
	0.002	**0.005**		0.002	0.005
r'	0.015		*R' (n = 2)*	0.008	
		0.032			0.016
po	**0.134**		*pO*	0.134	
	0.056	**0.128**		0.056	0.128
p r'	**0.039**		*pR'*	0.020	
		0.038			0.019
o r'	**0.000**		*OR'*	0.000	
		0.000			0.000
po r'	**0.180**		*pOR'*	0.090	
		0.201			0.101
Universe Score and Error Variance			**Composite Score Indices**		
***o* = 1, *r* = 2**	**T1 - AI**	**T2 - AA**		***o* = 1, *r* = 2**	***o* = 2, *r* = 2**
$\hat{\sigma}^2(p)$	**0.528**	*0.973*	$\hat{\sigma}_C^2(p)$	0.50891	0.50891
	0.502	**0.503**	$\hat{\sigma}_C^2(\delta)$	0.15079	0.08020
$\hat{\sigma}^2(\delta)$	**0.244**	0.226	$\hat{\sigma}_C^2(\Delta)$	0.15916	0.08731
	0.056	**0.248**	$E_C\hat{\rho}^2$	**0.77143**	0.86386
$\hat{\sigma}^2(\Delta)$	**0.253**	0.222	$\hat{\phi}_C$	**0.76177**	0.85357
	0.058	**0.268**			

Note. Bold values are the estimates under the operational conditions.

The results indicate that (1) the universe scores for the two essay types were highly correlated (.97). In other words, examinees who scored high on AI would also generally receive high scores on AA, and vice versa; (2) correlated measurement error was small with a δ-type correlation coefficient of about .23 across the essay types; and (3) the estimated standard error of measurement was about 0.39 and the estimated generalizability coefficient was about 0.77 for the composite scores under a single occasion. These results are similar to those obtained for the three-facet univariate designs.

Discussion

One critical psychometric property of any good assessment is its score reliability. Assessment scores should be consistent and generalizable over the measurement conditions that will be used to make decisions. To evaluate and understand measurement error and score reliability, it is important to consider and to model various measurement components associated with the assessment scores. The complex nature of large-scale performance assessments, however, in which some measurement conditions are confounded in a sparse matrix, does not permit the use of a single typical G-study design to disentangle the multiple sources of measurement error that affect score reliability.

The results of this study using the GMAT AWA data suggest that it may be feasible to use multiple generalizability analyses—based on both univariate and multivariate generalizability theory approaches—to model multiple measurement facets and to approximately estimate measurement error and reliability of test scores under complex measurement situations. By comparing results from a variety of generalizability analyses, we hope not only to present a better picture of measurement error and reliability of the GMAT AWA scores, but also to provide a framework for evaluating measurement precision for other large-scale assessments with complicated measurement procedures, such as the anticipated assessments associated with the Common Core State Standards.

The generalizability analysis results from the present study lead to the following conclusions. First, the GMAT AWA has the ability to differentiate examinees' analytical reasoning skills through writing. Second, the two essay types (AI and AA) measure similar writing skills. Third, average scores across occasions are not very different. Fourth, the variance components associated with the interactions between the person-effect and measurement facets are generally notable and they are the major sources of measurement error. But main effects for facets (and their own interactions) tend to be rather small.

Therefore, the measurement error estimates for relative and absolute decisions are similar, as are the generalizability and dependability coefficients under each study design. Most estimated standard errors of measurement are between 0.30 and 0.40 and generalizability coefficients range from 0.75 to 0.80 (see Table 7). For the multivariate analyses, the estimated correlated errors are negligible and the estimated measurement error and generalizability for relative decisions are similar to those from the univariate generalizability analyses.

It should be pointed out that the generalizability study designs under the GMAT AWA test administration do not permit disentangling occasion variance and prompt variance, which cannot be differentiated from essay type variance since only a single prompt is administered for each essay type on each occasion. The three-facet designs used in the study are the most comprehensive designs given the nature of the data, but they can only capture a subset of the measurement facets in the universe of generalization.

To achieve more precise estimates of different sources of measurement error, a more comprehensive design is needed to help resolve important confounding issues among the facets. In particular, the most crucial confounding issues can be resolved only with a design (or designs) in which there are at least two prompts for each essay type, and in which there are at least two ratings from human raters. With such a design (or designs), it would be possible to estimate how much of the total measurement error variance is attributable to prompts (nonconfounded with essay types), and how much of the total measurement error variance is attributable to raters. In that case, we would be able to estimate the magnitudes of measurement error associated with prompt and rater samplings and their impact on writing assessment score reliability. If practical constraints should prevent such a comprehensive design, then it wouldn't make much sense to provide a single estimate of absolute error variance, relative error variance, a generalizability coefficient, and/or a dependability coefficient. Instead, providing a range of estimates should more faithfully reflect the uncertainty in any one estimate. Doing so may be uncomfortable, but it reflects the uncertainty that inevitably arises when data do not conform to the structure of the intended universe of admissible observations and universe of generalization.

Authors

Xiaohong Gao, Senior Director, Measurement and Research, ACT, Inc.; Robert L. Brennan, Co-Director, Center for Advanced Studies in Measurement and Assessment (CASMA), University of Iowa; and Fanmin Guo, Vice President of Psychometric Research, 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 authors and do not necessarily reflect those of GMAC.

Acknowledgements

The authors wish to thank Paula Bruggeman, Writer/Editor, Manager, Research Services Department, GMAC, for editorial review.

References

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

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.

Brennan, R. L. (1992). *Elements of generalizability theory* (Rev. ed.). Iowa City, IA: ACT, Inc.

Brennan, R. L. (1999). *Manual for mGENOVA*. (Iowa Testing Programs Occasional Papers, No. 47). Iowa City, IA: University of Iowa.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of work keys listening and writing tests. *Educational and Psychological Measurement, 55*(2), 157–176.

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A GENeralized Analysis Of VAriance System.* (ACT Technical Bulletin, No. 43). Iowa City, IA: The American College Testing Program.

Cronbach, L. J., Gleser, G. C., Nanda, H. I., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*(3), 373–399.

Gao, X., & Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7*(2), 323–343.

Jarjoura, D., & Brennan, R. L. (1983). Multivariate generalizability models for tests developed from tables of specifications. In L. J. Fyans, Jr. (Ed.). *Generalizability theory: Inferences and practical applications: New directions for testing and measurement*. San Francisco: Jossey-Bass.

Keller, L., Clauser, B., Swanson, D., Harik, P., & Clyman, S. (2000, April). *A multivariate generalizability analysis to examine the effects of content stratification on precision of scores for a performance assessment.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1992, April). *Empirical evidence for the reliability and validity of performance assessments*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Lee, Y., & Kantor, R. (2005). *Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemed*. ETS Monograph Series, MS-3 (ISSN No. 1556–9021). Princeton, NJ: Educational Testing Service.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice, 13*(1), 5–8.

Shavelson, R. J., Gao, X., & Baxter, G. P. (1993). On the content validity of performance assessments: Centrality of domain specification. In M. Birenbaum & F. J. R. C. Dochy (Eds.), *Alternative in assessment of achievements, learning processes and prior knowledge* (pp.131–141). Boston: Kluwer.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement*, *36*, 61–71.

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

Wang, L., Zhang, Y., & Li S. (2007, April). *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.

Appendix A. Variance Components Associated with Different G-Study Designs

Table A. 1. Variance Components Associated With Different G-Study Designs

Crossed	i Confounded	i Confounded	i Confounded	o Confounded	Each O	Each T	Each R
p x r x (i:t) x o	p x r x t x o	r:(p x t) x o	p x (r′:t) x o	(i x r):(p x t)	r:(p x t)	p x r′ x o	i:(p x t)
p	p	p	p	p	p, po	p, pt	p, pr
r	**r**	**r:pt (r, pr, rt, prt)**	**r′:t (r′, r′t)**	**r:pt (r, pr, rt, prt)**	r:pt, or, por, ort, port	r, r′t	
t	t	t	t	t	t, ot		t, rt
o	o	o	o			o, ot	
i:t (i, ti)							i:pt, ir, pir, iri, pirt
pr	**pr**		**pr′:t (pr′, pr′t)**			pr′, pr′t	
pt	pt	pt	pt	pt	pt, pot		pt, prt
po	po	po	po			po, pot	
pi:t (pi, pit)				i:pt (i, pi, ti, pti) *o, po, to, pto			
rt	**rt**						
or	or		*r′:ot (or′, or′t)*			or′, or′t	
ot	ot	ot	ot				
prt	**prt**						
por	por					por′, por′t	
pot	pot	pot	pot				
ort	ort						
ir:t (ir, irt)							
port	port						
pir:t (pir, pirt)		*r:pot (or, por, ort, port)*	*pr′:ot (por′,por′t)*	*ir:pt (ir, irt, pir, pirt);*or, ort, por, port*			
i:ot (io, iot)							
pi:to (pio, piot)							
ri:ot (ior, iort)							
pir:ot (pior, piort)							

Note. Individual symbol styles (bold, italic, or font) represent corresponding variance components among the designs; the shaded components are inestimable in the study; prompt (i) and occasion (o) are confounded; and hidden variance components are embossed under the two-facet designs.