Center for Advanced Studies in
Measurement and Assessment

CASMA Research Report

Number 4

A Multivariate Generalizability Analysis
of the Multistate Bar Examination[*]

Ping Yin[†]

January 2004

for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: ping-yin@uiowa.edu).

The Multistate Bar Examination (MBE) is developed by the National Conference of Bar Examiners (NCBE) and is widely used in the U.S. The MBE is a 200-item multiple-choice test that is administered in most U.S. jurisdictions. The MBE has six sub-tests: Constitutional Law (CNL), Contracts (CTR), Criminal Law and Procedure (CRM), Evidence (EVD), Real Property (RLP), and Torts (TOR). There are 34 questions each in Contracts and Torts, and 33 questions each in Constitutional Law, Criminal Law and Procedure, Evidence, and Real Property. The MBE is administered twice a year: February and July.

The structure of the MBE fits perfectly into the "table of specifications" model treated in Brennan (2001a, pp. 268–277), that was originally discussed by Jarjoura and Brennan (1982). In Brennan (2001a), a multivariate generalizability theory framework is recommended for treating tests developed according to a "table of specifications." Specifically, in the multivariate generalizability theory model, each sub-test represents one content area, or one fixed category (v), and there is a different set of items or questions (i) associated with each category. In the terminology of generalizability theory, the multivariate design is $p^{\bullet} \times i^{\circ}$, where p stands for persons or examinees. That is, examinees respond to all questions, and questions are nested within fixed categories. The superscript filled circle for p (a "linked" facet) indicates that persons have scores on all six sub-tests. The superscript empty circle for i indicates that items are nested within sub-tests, that is, each item belongs to only one sub-test. The design is also unbalanced because different numbers of items are associated with different sub-tests.

Generalizability theory liberalizes classical test theory by permitting analyses that explicitly incorporate multiple sources of error in the measurement model. This is usually accomplished using analysis of variance procedures that focus on estimating variance components (univariate generalizability theory) and/or covariance components (multivariate generalizability theory) (Brennan, 2001a; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). The conceptual framework of generalizability theory includes a *universe of admissible observations* and a G (generalizability) study, as well as *universes of generalization* and D (decision) studies defined by the investigator. In multivariate G theory, each object of measurement has multiple universe scores (fixed levels), each of which is associated with a random-effects variance components design, and there are covariance components associated with pairs of fixed levels for linked facets.

The main purpose of this study is to examine the MBE from a multivariate generalizability theory perspective. Specifically, using MBE data collected over three years (three from February administrations, and three from July administrations), a multivariate generalizability analysis is conducted for each set of data. Various results are "averaged" over February administrations, over July administrations, and over all administrations.

Statistics of interest include variance and covariance component estimates for both G and D studies, error variances, standard errors of measurement, generalizability coefficients, and dependability coefficients. These statistics can be obtained from each of the six administrations of the MBE. Means and standard deviations of these statistics can also be obtained. Standard deviations

1

Table 1: Descriptive Statistics of the MBE Scores

Score	Data	N	Mean	SD	Min[a]	Max[a]	Skew	Kurt
Raw	Feb1	19673	120.81	17.16	49	181	-0.05	3.10
	Feb2	20288	127.98	18.31	38	186	-0.19	3.11
	Feb3	19399	120.15	17.33	38	179	-0.10	3.12
	July1	43791	126.90	16.69	39	180	-0.25	3.04
	July2	44637	128.72	19.46	12	182	-0.33	2.95
	July3	45732	130.27	18.33	35	183	-0.37	3.05
Scale	Feb1	19673	136.85	14.83	75	189	-0.05	3.10
	Feb2	20288	136.55	15.49	60	186	-0.19	3.11
	Feb3	19399	135.26	15.15	63	187	-0.10	3.12
	July1	43791	141.97	15.90	58	193	-0.25	3.04
	July2	44637	142.84	15.33	51	185	-0.33	2.95
	July3	45732	141.17	15.72	59	186	-0.37	3.05

[a] Scale scores are rounded.

are empirical estimates of the standard errors of these statistics without making normality assumptions.

Data

Data used in the study were from administrations of the MBE from three years. Because the MBE is administered in both February and July, three years' worth of data yields scores from six administrations. The three data sets for February administrations and the three data sets for July administrations also provide useful information about different characteristics of February- and July-tested examinees. (There are usually more examinees who test in July, and they tend to score higher.)

A new form is used for each administration of the MBE. Equating is conducted for each form so that scores for each new form are on the same scale as scores for previous forms. The relationship between equated raw scores and scale scores is linear. The principal analyses reported in the paper involves raw scores, with each item scored as either "0"(incorrect) or "1" (correct). Statistics about scale scores are also included whenever possible.

Table 1 reports descriptive statistics for raw and scale total scores obtained from each of the six administrations of the MBE. Number of examinees (N), mean, standard deviation (SD), minimum (Min), maximum (Max), skewness (Skew), and kurtosis (Kurt) are reported in the table.

From Table 1, it can be observed that more examinees (more than twice as many) took the July administrations of the MBE than the February administrations. For both raw and scale scores, means for the July administrations

are higher than those for the February administrations. The difference is about 5 or 6 in the scale score metric. Standard deviations based on scale scores for February and July administrations are similar in magnitude. Scores for the July administrations are also slightly more negatively skewed than the February administrations.

For the raw scores, within the three February and July administrations, means are sometimes quite different. For example, the mean score for "Feb2" is much higher than means for the other two February administrations. By contrast, for the scale scores, the February means are similar, indicating that the "Feb2" test form was easy relative to the other forms, but this difference was controlled by equating. However, there are clear differences in mean scale scores between February and July administrations (means for July administrations are higher).

Method

G Study

The multivariate G study $p^{\bullet} \times i^{\circ}$ design is used here to characterize sub-structures of the MBE. Letting v_1 to v_6 stand for the six sub-tests, p stands for persons or examinees, and i stands for items or questions, a linear model can be used to describe the observed scores for each sub-test.

For example, for v_1:

$$X_{piv_1} = \mu_{v_1} + \nu_{pv_1} + \nu_{iv_1} + \nu_{piv_1} \tag{1}$$

where "ν" terms in Equation 1 represent effects for v_1, and μ_{v_1} is the grand mean for the universe score for v_1[1]. An examinee's observed score (X_{piv_1}) for Constitutional Law can be decomposed into effects due to persons (ν_{pv_1}), items (ν_{iv_1}), person by item interaction or residual (ν_{piv_1}), and a grand mean for the category (μ_{v_1}).

Similar equations can be used to describe observed scores for v_2 to v_6. That is, there is a univariate $p \times i$ design associated with each level of v. The superscript filled circle for p ($\bullet$) indicates that persons have scores on all six sub-tests. The superscript empty circle for i ($\circ$) indicates that items are nested within sub-tests; that is, each item belongs to only one sub-test. The fixed levels of v are linked in the sense that the same group of examinees respond to all items for all levels.

The multivariate $p^{\bullet} \times i^{\circ}$ design is also unbalanced because different numbers of items are associated with different sub-tests: there are 34 questions each in Contracts and Torts, and 33 questions each in the other four sub-tests. To estimate variance components for each sub-test, the ANOVA procedure was used, which is appropriate because the univariate $p \times i$ design for each level of v is balanced. To estimate covariance components across sub-tests, the CP

---

[1]Note the distinction between ν and v.

(sum-of-cross-products) procedures is used. For a detailed description of these procedures, see Brennan (2001a). The computer program mGENOVA was used for conducting the analysis (Brennan, 2001b).

Variance and covariance components for the $p^\bullet \times i^\circ$ design are presented here using matrix conventions:

$$
\mathbf{\Sigma}_p = \left[\begin{array}{cccccc}
\sigma_1^2(p) & \sigma_{12}(p) & \sigma_{13}(p) & \sigma_{14}(p) & \sigma_{15}(p) & \sigma_{16}(p) \\
\sigma_{21}(p) & \sigma_2^2(p) & \sigma_{23}(p) & \sigma_{24}(p) & \sigma_{25}(p) & \sigma_{26}(p) \\
\sigma_{31}(p) & \sigma_{32}(p) & \sigma_3^2(p) & \sigma_{34}(p) & \sigma_{35}(p) & \sigma_{36}(p) \\
\sigma_{41}(p) & \sigma_{42}(p) & \sigma_{43}(p) & \sigma_4^2(p) & \sigma_{45}(p) & \sigma_{46}(p) \\
\sigma_{51}(p) & \sigma_{52}(p) & \sigma_{53}(p) & \sigma_{54}(p) & \sigma_5^2(p) & \sigma_{56}(p) \\
\sigma_{61}(p) & \sigma_{62}(p) & \sigma_{63}(p) & \sigma_{64}(p) & \sigma_{65}(p) & \sigma_6^2(p)
\end{array} \right] \tag{2}
$$

$$
\mathbf{\Sigma}_i = \left[\begin{array}{cccccc}
\sigma_1^2(i) & & & & & \\
& \sigma_2^2(i) & & & & \\
& & \sigma_3^2(i) & & & \\
& & & \sigma_4^2(i) & & \\
& & & & \sigma_5^2(i) & \\
& & & & & \sigma_6^2(i)
\end{array} \right] \tag{3}
$$

$$
\mathbf{\Sigma}_{pi} = \left[\begin{array}{cccccc}
\sigma_1^2(pi) & & & & & \\
& \sigma_2^2(pi) & & & & \\
& & \sigma_3^2(pi) & & & \\
& & & \sigma_4^2(pi) & & \\
& & & & \sigma_5^2(pi) & \\
& & & & & \sigma_6^2(pi)
\end{array} \right] \tag{4}
$$

In these matrices, each column represents a sub-test or fixed category in the following order: CNL, CTR, CRM, EVD, RLP, and TOR. Elements on the diagonal are estimated variance components, and elements on the off-diagonal are estimated covariance components. Because persons are "linked," covariance components can be estimated between categories. Because items and person-by-item interaction terms are not "linked," covariance components between categories are zero, and, therefore, not listed.

Disattenuated correlations for universe scores between categories can also be estimated. For example, the disattenuated correlation coefficient between two categories (v and v') is:

$$
\rho_{vv'}(p) = \frac{\sigma_{vv'}(p)}{\sqrt{\sigma_v^2(p)\ \sigma_{v'}^2(p)}}. \tag{5}
$$

For the MBE data, a high disattenuated correlation coefficient for the universe scores between two sub-tests indicates that the content areas of the two sub-tests are closely related.

D Study

For a universe of generalization that consists of randomly parallel forms of the MBE, the multivariate D study is $p^{\bullet} \times I^{\circ}$, where the uppercase I is used to indicate that the focus of the D study is on average scores over items.

It is relatively simple to estimate variance and covariance components for the D study based on the G study results: $\boldsymbol{\Sigma}_p$ is unchanged, and $\boldsymbol{\Sigma}_I$ and $\boldsymbol{\Sigma}_{pI}$ can be obtained by dividing the diagonal elements in $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Sigma}_{pi}$ by the number of D study items (n'_i) within a sub-test.

Relative and absolute error variance-covariance matrices can then be obtained:

$$\boldsymbol{\Sigma}_\delta = \boldsymbol{\Sigma}_{pI} \tag{6}$$

$$\boldsymbol{\Sigma}_\Delta = \boldsymbol{\Sigma}_I + \boldsymbol{\Sigma}_{pI}, \tag{7}$$

which are diagonal matrices for the $p^{\bullet} \times I^{\circ}$ design. The entries in each of the six cells are the relative-error (δ) or absolute-error (Δ) variances for the six sub-tests.

In generalizability theory, relative error variance is associated with norm-referenced interpretations (e.g., rank ordering examinees), and absolute error variance is associated with criterion-referenced interpretations (comparing an examinee's observed mean score to his or her universe score). The square roots of relative and absolute error variance are the relative and absolute standard errors of measurement (SEMs).

Composite Score

A composite score based on all six sub-tests for an examinee can also be examined. A composite universe score is a weighted average of the universe scores for each level of v (v_1 to v_6), where the weights are proportional to the numbers of items in each sub-test, a weight vector can be defined as:

$$\mathbf{w}' = \left[\begin{array}{cccccc} \frac{n'_{i1}}{n'_+} & \frac{n'_{i2}}{n'_+} & \frac{n'_{i3}}{n'_+} & \frac{n'_{i4}}{n'_+} & \frac{n'_{i5}}{n'_+} & \frac{n'_{i6}}{n'_+} \end{array} \right], \tag{8}$$

where n'_{i1} to n'_{i6} are number of items for each sub-test, and n'_+ is the total number of items over all sub-tests.

Using the weight vector $\mathbf{w}'$, the variance for the composite universe score (in the mean score metric) is:

$$\sigma^2_C(p) = \mathbf{w}' \boldsymbol{\Sigma}_p \mathbf{w}, \tag{9}$$

where C stands for composite. Similarly, estimated relative and absolute error variances for the composite score (in the mean score metric) are:

$$\sigma^2_C(\delta) = \mathbf{w}' \boldsymbol{\Sigma}_\delta \mathbf{w} \tag{10}$$

$$\sigma^2_C(\Delta) = \mathbf{w}' \boldsymbol{\Sigma}_\Delta \mathbf{w} \tag{11}$$

Square roots of relative and absolute error variances for the composite are the relative and absolute standard errors of measurement (SEMs) for the composite.

A generalizability coefficient($\boldsymbol{E}\rho_C^2$) for the composite score is:

$$\boldsymbol{E}\rho_C^2 \quad = \quad \frac{\sigma_C^2(p)}{\sigma_C^2(p) + \sigma_C^2(\delta)}. \tag{12}$$

A generalizability coefficient is often used for making norm-referenced interpretations. If an investigator wants to quantify the extent to which the composite universe score μ_p (subscript C is omitted here) is large or small relative to a certain cut score or standard λ, then interest focuses on $\mu_p - \lambda$, and a dependability coefficient is more appropriate (Kane, 1996). In the dependability index, the "squared standard tolerance" (ST) is used instead of the universe score variance (in generalizability coefficient). A dependability index relative to λ is defined as:

$$\begin{aligned} \Phi(\lambda) \quad &= \quad \frac{ST^2}{ST^2 + \sigma^2(\Delta)}, \\ &= \quad \frac{\boldsymbol{E}(\mu_p - \lambda)^2}{\boldsymbol{E}(\mu_p - \lambda)^2 + \sigma^2(\Delta)}. \end{aligned} \tag{13}$$

where $\sigma^2(\Delta)$ is the absolute error variance.

Results

G Study Results

Tables 2 and 3 report G study results for the six administrations of the MBE. Note that the results in these tables are based on raw scores only. In these tables, for the universe score matrix (p), estimated variance components are on the diagonal (bold), estimated covariance components are on the lower off-diagonal, and estimated disattenuated correlation coefficients are on the upper off-diagonal (italics). For the i and pi matrices, only estimated variance components are reported.

For the six administrations, estimated variance and covariance components for p are relatively small (which is not unusual for results reported on a 0, 1 scale) and similar in magnitude across different sub-tests. Estimated disattenuated correlation coefficients for p between different sub-tests are relatively high (between 0.80 to 0.99), which indicate that universe scores for sub-tests are highly correlated.

Estimated variance components for i are about three to four times larger than those for p, which indicates that there is relatively large variability in average difficulty for items. Estimated variance components for pi are the largest (about five times larger than variance component estimates for i), which indicate that relative difficulties of items tend to vary across different persons.

Means of estimated G study results for all six administrations, the three February administrations, and the three July administrations are reported in

Table 2: Estimated G Study Results for February Administrations of the MBE

		CNL	CTR	CRM	RLP	TOR	EVD
Feb1							
	$\widehat{\boldsymbol{\Sigma}}_p$	**0.009**	*0.843*	*0.878*	*0.854*	*0.861*	*0.880*
		0.007	**0.007**	*0.907*	*0.942*	*0.994*	*0.905*
		0.006	0.005	**0.005**	*0.837*	*0.932*	*0.906*
		0.008	0.008	0.006	**0.009**	*0.882*	*0.868*
		0.005	0.006	0.004	0.006	**0.005**	*0.936*
		0.008	0.007	0.006	0.008	0.006	**0.009**
	$\widehat{\boldsymbol{\Sigma}}_i$	0.023	0.043	0.038	0.043	0.043	0.052
	$\widehat{\boldsymbol{\Sigma}}_{pi}$	0.182	0.196	0.191	0.195	0.192	0.186
Feb2							
	$\widehat{\boldsymbol{\Sigma}}_p$	**0.008**	*0.843*	*0.904*	*0.853*	*0.862*	*0.877*
		0.007	**0.008**	*0.943*	*0.912*	*0.947*	*0.913*
		0.006	0.006	**0.005**	*0.909*	*0.956*	*0.967*
		0.008	0.009	0.008	**0.013**	*0.867*	*0.881*
		0.006	0.007	0.006	0.008	**0.006**	*0.914*
		0.008	0.008	0.007	0.010	0.007	**0.010**
	$\widehat{\boldsymbol{\Sigma}}_i$	0.035	0.040	0.038	0.023	0.032	0.041
	$\widehat{\boldsymbol{\Sigma}}_{pi}$	0.156	0.185	0.192	0.204	0.191	0.192
Feb3							
	$\widehat{\boldsymbol{\Sigma}}_p$	**0.009**	*0.839*	*0.872*	*0.837*	*0.812*	*0.881*
		0.007	**0.007**	*0.893*	*0.896*	*0.910*	*0.897*
		0.006	0.005	**0.005**	*0.868*	*0.930*	*0.944*
		0.007	0.006	0.005	**0.008**	*0.817*	*0.862*
		0.006	0.006	0.005	0.006	**0.007**	*0.893*
		0.008	0.007	0.006	0.007	0.007	**0.009**
	$\widehat{\boldsymbol{\Sigma}}_i$	0.032	0.026	0.051	0.038	0.034	0.040
	$\widehat{\boldsymbol{\Sigma}}_{pi}$	0.193	0.189	0.196	0.196	0.189	0.199

Table 3: Estimated G Study Results for July Administrations of the MBE

		CNL	CTR	CRM	RLP	TOR	EVD
July1							
	$\widehat{\boldsymbol{\Sigma}}_p$	**0.007**	*0.805*	*0.885*	*0.878*	*0.876*	*0.892*
		0.005	**0.005**	*0.803*	*0.959*	*0.917*	*0.837*
		0.006	0.005	**0.007**	*0.849*	*0.919*	*0.910*
		0.007	0.006	0.006	**0.008**	*0.890*	*0.886*
		0.005	0.005	0.006	0.006	**0.005**	*0.911*
		0.007	0.005	0.007	0.007	0.006	**0.008**
	$\widehat{\boldsymbol{\Sigma}}_i$	0.022	0.033	0.033	0.034	0.044	0.038
	$\widehat{\boldsymbol{\Sigma}}_{pi}$	0.176	0.193	0.187	0.198	0.191	0.198
July2							
	$\widehat{\boldsymbol{\Sigma}}_p$	**0.011**	*0.890*	*0.924*	*0.896*	*0.895*	*0.928*
		0.008	**0.007**	*0.916*	*0.931*	*0.947*	*0.909*
		0.009	0.007	**0.009**	*0.908*	*0.923*	*0.956*
		0.011	0.009	0.010	**0.014**	*0.874*	*0.906*
		0.007	0.006	0.007	0.008	**0.006**	*0.914*
		0.010	0.008	0.009	0.011	0.007	**0.011**
	$\widehat{\boldsymbol{\Sigma}}_i$	0.029	0.032	0.029	0.030	0.033	0.041
	$\widehat{\boldsymbol{\Sigma}}_{pi}$	0.173	0.193	0.190	0.201	0.191	0.178
July3							
	$\widehat{\boldsymbol{\Sigma}}_p$	**0.009**	*0.890*	*0.904*	*0.875*	*0.878*	*0.921*
		0.008	**0.009**	*0.934*	*0.960*	*0.947*	*0.938*
		0.006	0.006	**0.005**	*0.884*	*0.968*	*0.955*
		0.009	0.009	0.006	**0.011**	*0.917*	*0.890*
		0.006	0.006	0.005	0.007	**0.005**	*0.934*
		0.009	0.009	0.007	0.009	0.007	**0.010**
	$\widehat{\boldsymbol{\Sigma}}_i$	0.030	0.029	0.048	0.033	0.039	0.033
	$\widehat{\boldsymbol{\Sigma}}_{pi}$	0.149	0.188	0.189	0.193	0.180	0.194

Table 4: Means of Estimated G Study Results for All Datasets, February, and July Administrations of the MBE

		CNL	CTR	CRM	RLP	TOR	EVD
All							
	$\widehat{\mathbf{\Sigma}}_p$	**0.009**	*0.852*	*0.894*	*0.866*	*0.864*	*0.896*
		0.007	**0.007**	*0.899*	*0.933*	*0.944*	*0.900*
		0.006	0.006	**0.006**	*0.876*	*0.938*	*0.940*
		0.008	0.008	0.007	**0.010**	*0.875*	*0.882*
		0.006	0.006	0.005	0.007	**0.006**	*0.917*
		0.008	0.007	0.007	0.009	0.007	**0.010**
	$\widehat{\mathbf{\Sigma}}_i$	0.029	0.034	0.039	0.034	0.038	0.041
	$\widehat{\mathbf{\Sigma}}_{pi}$	0.171	0.191	0.191	0.198	0.189	0.191
February							
	$\widehat{\mathbf{\Sigma}}_p$	**0.008**	*0.842*	*0.884*	*0.848*	*0.845*	*0.879*
		0.007	**0.007**	*0.914*	*0.916*	*0.951*	*0.905*
		0.006	0.005	**0.005**	*0.871*	*0.939*	*0.939*
		0.008	0.008	0.006	**0.010**	*0.855*	*0.870*
		0.006	0.006	0.005	0.006	**0.006**	*0.914*
		0.008	0.007	0.006	0.008	0.007	**0.009**
	$\widehat{\mathbf{\Sigma}}_i$	0.030	0.036	0.042	0.035	0.036	0.044
	$\widehat{\mathbf{\Sigma}}_{pi}$	0.177	0.190	0.193	0.198	0.190	0.192
July							
	$\widehat{\mathbf{\Sigma}}_p$	**0.009**	*0.861*	*0.904*	*0.883*	*0.883*	*0.914*
		0.007	**0.007**	*0.884*	*0.950*	*0.937*	*0.895*
		0.007	0.006	**0.007**	*0.880*	*0.937*	*0.940*
		0.009	0.008	0.008	**0.011**	*0.894*	*0.894*
		0.006	0.006	0.006	0.007	**0.006**	*0.920*
		0.009	0.007	0.008	0.009	0.007	**0.010**
	$\widehat{\mathbf{\Sigma}}_i$	0.027	0.031	0.037	0.032	0.039	0.037
	$\widehat{\mathbf{\Sigma}}_{pi}$	0.166	0.191	0.189	0.197	0.187	0.190

Table 5: Means of Estimated G Study Results for Differences between February and July Administrations of the MBE

	CNL	CTR	CRM	RLP	TOR	EVD
February-July						
$\widehat{\boldsymbol{\Sigma}}_p$	**-0.001**					
	-0.000	**0.001**				
	-0.001	-0.000	**-0.002**			
	-0.001	-0.000	-0.001	**-0.001**		
	-0.000	0.001	-0.001	-0.000	**0.000**	
	-0.001	0.000	-0.001	-0.001	0.000	**-0.000**
$\widehat{\boldsymbol{\Sigma}}_i$	0.003	0.005	0.006	0.003	-0.002	0.007
$\widehat{\boldsymbol{\Sigma}}_{pi}$	0.011	-0.002	0.005	0.002	0.003	0.002

Table 4. Table 5 reports differences of variance and covariance component estimates between averages for February and July administrations.

From Table 4, among the six sub-tests, on average, the disattenuated correlations between CTR and RLP, between CTR and TOR, between CRM and TOR, and between CRM and EVD are relatively high (above 0.9), which indicates that the Contracts subscore is particularly closely related to Real Property and Torts, and the Criminal Law and Procedure subscore is closely related to Torts and Evidence. From Tables 4 and 5, it can be observed also that there is almost no difference in $\widehat{\boldsymbol{\Sigma}}_p$ from different administrations. Differences in variance component estimates for i and pi are also relatively small. Clearly, G study results for February and July administrations of the MBE are very similar.

Standard deviations[2] of estimated G study results for all six administrations, three February administrations, and three July administrations are reported in Table 6[3]. Treating different administrations of the MBE as replications of the measurement procedure, these standard deviations are direct estimates of the standard errors of the estimated variance components and covariance components for a single form.

Standard deviations of estimated variance and covariance components for p are small, which indicates that estimates of variance and covariance components are relatively stable across forms. Standard deviations of $\hat{\sigma}^2(i)$ are

---

[2]The standard deviation are estimated standard errors for a single administration. Another statistic that could be considered is the standard error of the mean, which provides information about standard errors for an average administration (e.g., average February or July administration). The standard error of the mean is obtained by dividing the standard deviations by the square root of the number of administrations over which the mean is computed. Standard errors of the mean are necessarily smaller than the standard deviations reported in Table 6. That is, results based on means are more stable than results based on a single administration.

[3]These standard deviations were obtained by using a denominator of $n - 1$, with n being the number of administrations.

Table 6: Standard Deviations of Estimated G Study Results for All Datasets, February, and July Administrations of the MBE

		CNL	CTR	CRM	RLP	TOR	EVD
All							
	$\widehat{\boldsymbol{\Sigma}}_p$	**0.001**					
		0.001	**0.001**				
		0.001	0.001	**0.002**			
		0.002	0.001	0.002	**0.002**		
		0.001	0.001	0.001	0.001	**0.001**	
		0.001	0.001	0.001	0.001	0.001	**0.001**
	$\widehat{\boldsymbol{\Sigma}}_i$	0.005	0.007	0.008	0.007	0.005	0.006
	$\widehat{\boldsymbol{\Sigma}}_{pi}$	0.016	0.004	0.003	0.004	0.004	0.008
February							
	$\widehat{\boldsymbol{\Sigma}}_p$	**0.000**[a]					
		0.000	**0.000**				
		0.000	0.001	**0.000**			
		0.001	0.001	0.001	**0.002**		
		0.000	0.000	0.001	0.001	**0.001**	
		0.000	0.001	0.001	0.001	0.001	**0.001**
	$\widehat{\boldsymbol{\Sigma}}_i$	0.006	0.009	0.007	0.011	0.006	0.006
	$\widehat{\boldsymbol{\Sigma}}_{pi}$	0.019	0.006	0.002	0.005	0.002	0.007
July							
	$\widehat{\boldsymbol{\Sigma}}_p$	**0.002**					
		0.002	**0.002**				
		0.002	0.001	**0.002**			
		0.002	0.002	0.002	**0.003**		
		0.001	0.001	0.001	0.001	**0.000**	
		0.002	0.002	0.001	0.002	0.001	**0.001**
	$\widehat{\boldsymbol{\Sigma}}_i$	0.005	0.002	0.010	0.002	0.005	0.004
	$\widehat{\boldsymbol{\Sigma}}_{pi}$	0.015	0.003	0.002	0.004	0.006	0.011

[a] 0.000 in table means <0.0005.

relatively small, which indicates that variability of the variances of item difficulties across forms is relatively small. Standard deviations for $\hat{\sigma}^2(pi)$ are quite small except for Constitutional Law (CNL), which indicates that variability of interaction variances for most sub-tests is relatively small for different forms. G study results indicate that variance component estimates for February and July administrations are very stable, which suggests that the table of specifications for the MBE is well-defined and faithfully followed.

D Study Results

The convention in generalizability theory is to report D study results in terms of a mean score metric. For the composite score, the weight for each sub-test is defined by:

$$\mathbf{w}' = \left[\begin{array}{cccccc} \frac{33}{200} & \frac{34}{200} & \frac{33}{200} & \frac{33}{200} & \frac{34}{200} & \frac{33}{200} \end{array} \right], \tag{14}$$

where columns 1 to 6 represent CNL, CTR, CRM, RLP, TOR, and EVD, respectively. Equations 9, 10, 11, 12, and 13 were used to obtain D study results for the composite score with weights defined by Equation 14.

Sometimes it is useful to express results in terms of the total score metric, rather in the mean score metric. To express D study results in the total score metric, the simplest method is to multiply the "mean" score variances by the square of the number of observations in the D study. For example, the universe score variance and relative error variance in terms of the total score metric are:

$$\hat{\sigma}_T^2(p) = (n'_+)^2 \, \hat{\sigma}_M^2(p) \tag{15}$$
$$\hat{\sigma}_T^2(\delta) = (n'_+)^2 \, \hat{\sigma}_M^2(\delta), \tag{16}$$

where T stands for "total score metric", and M stands for "mean score metric". Subscript C is omitted in Equations 15 and 16. Note that the generalizability coefficient remains the same regardless of whether variances are expressed in the mean or total score metric.

As mentioned earlier, equating is conducted for each form of the MBE so that scores for each new form are on the same scale as scores from previous forms. Scale scores, therefore, are of great interest in practice. For the MBE, conversions from raw (composite score in total score metric) to composite scale scores are linear, and D study results for the scale scores can be expressed by a linear transformation. Specifically:

$$\hat{\sigma}_S^2(p) = b^2 \, \hat{\sigma}_R^2(p) \tag{17}$$
$$\hat{\sigma}_S^2(\delta) = b^2 \, \hat{\sigma}_R^2(\delta) \tag{18}$$

where b stands for the slope of the linear conversion from raw score to scale score, S stands for composite scale score, and R stands for composite raw score (both S and R are in the total score metric).

Table 7 reports universe score standard deviations[4], relative SEMs, and gen-

---

[4]Universe score standard deviations are reported instead of the commonly-reported universe score variances because the the universe score standard deviation is on the same scale as the relative SEM.

Table 7: D Study Results for six Administrations of the MBE

	Raw							Scale
	CNL	CTR	CRM	RLP	TOR	EVD	Comp.	Comp.
Feb1								
$\hat{\sigma}(p)$	3.05	2.84	2.30	3.21	2.31	3.07	16.01	13.84
$\hat{\sigma}(\delta)$	2.45	2.58	2.51	2.54	2.55	2.48	6.16	5.33
$\boldsymbol{E}\hat{\rho}^2$	0.61	0.55	0.45	0.62	0.45	0.61	0.87	0.87
Feb2								
$\hat{\sigma}(p)$	2.93	2.96	2.44	3.70	2.69	3.31	17.26	14.60
$\hat{\sigma}(\delta)$	2.27	2.51	2.52	2.59	2.54	2.51	6.10	5.16
$\boldsymbol{E}\hat{\rho}^2$	0.63	0.58	0.48	0.67	0.53	0.63	0.89	0.89
Feb3								
$\hat{\sigma}(p)$	3.10	2.81	2.24	2.90	2.84	3.20	16.17	14.14
$\hat{\sigma}(\delta)$	2.53	2.53	2.54	2.55	2.53	2.56	6.23	5.45
$\boldsymbol{E}\hat{\rho}^2$	0.60	0.55	0.44	0.56	0.56	0.61	0.87	0.87
July1								
$\hat{\sigma}(p)$	2.77	2.30	2.78	2.97	2.48	3.03	15.50	14.78
$\hat{\sigma}(\delta)$	2.41	2.56	2.48	2.55	2.55	2.56	6.16	5.87
$\boldsymbol{E}\hat{\rho}^2$	0.57	0.45	0.56	0.58	0.49	0.58	0.86	0.86
July2								
$\hat{\sigma}(p)$	3.47	2.75	3.09	3.86	2.60	3.39	18.47	14.55
$\hat{\sigma}(\delta)$	2.39	2.56	2.50	2.57	2.55	2.42	6.13	4.83
$\boldsymbol{E}\hat{\rho}^2$	0.68	0.54	0.60	0.69	0.51	0.66	0.90	0.90
July3								
$\hat{\sigma}(p)$	3.17	3.17	2.28	3.43	2.50	3.38	17.31	14.85
$\hat{\sigma}(\delta)$	2.22	2.53	2.50	2.52	2.47	2.53	6.03	5.18
$\boldsymbol{E}\hat{\rho}^2$	0.67	0.61	0.45	0.65	0.51	0.64	0.89	0.89

Table 8: Means and Standard Deviations of D Study Results

	Raw							Scale
	CNL	CTR	CRM	RLP	TOR	EVD	Comp.	Comp.
All								
Mean								
$\hat{\sigma}(p)$	3.08	2.81	2.52	3.34	2.57	3.23	16.79	14.46
$\hat{\sigma}(\delta)$	2.38	2.55	2.51	2.55	2.53	2.51	6.14	5.30
$\boldsymbol{E}\hat{\rho}^2$	0.63	0.55	0.50	0.63	0.51	0.62	0.88	0.88
SD								
$\hat{\sigma}(p)$	0.24	0.29	0.34	0.39	0.18	0.15	1.09	0.39
$\hat{\sigma}(\delta)$	0.12	0.03	0.02	0.03	0.03	0.05	0.07	0.35
$\boldsymbol{E}\hat{\rho}^2$	0.04	0.06	0.07	0.05	0.04	0.03	0.01	0.01
February								
Mean								
$\hat{\sigma}(p)$	3.03	2.87	2.32	3.27	2.61	3.19	16.48	14.19
$\hat{\sigma}(\delta)$	2.41	2.54	2.53	2.56	2.54	2.52	6.16	5.31
$\boldsymbol{E}\hat{\rho}^2$	0.61	0.56	0.46	0.62	0.51	0.62	0.88	0.88
SD								
$\hat{\sigma}(p)$	0.09	0.08	0.10	0.40	0.27	0.12	0.68	0.39
$\hat{\sigma}(\delta)$	0.13	0.04	0.02	0.03	0.01	0.04	0.06	0.14
$\boldsymbol{E}\hat{\rho}^2$	0.01	0.02	0.02	0.05	0.06	0.02	0.01	0.01
July								
Mean								
$\hat{\sigma}(p)$	3.14	2.74	2.72	3.42	2.53	3.27	17.10	14.73
$\hat{\sigma}(\delta)$	2.34	2.55	2.49	2.55	2.52	2.50	6.11	5.29
$\boldsymbol{E}\hat{\rho}^2$	0.64	0.53	0.54	0.64	0.50	0.63	0.89	0.89
SD								
$\hat{\sigma}(p)$	0.35	0.43	0.41	0.44	0.07	0.20	1.49	0.16
$\hat{\sigma}(\delta)$	0.11	0.02	0.01	0.03	0.04	0.07	0.07	0.53
$\boldsymbol{E}\hat{\rho}^2$	0.06	0.08	0.08	0.06	0.01	0.04	0.02	0.02

eralizability coefficients for both raw and scale scores in the total score metric for the six administrations. For the six sub-tests, the universe score standard deviations were obtained by multiplying the "mean" universe score standard deviations by the corresponding D study sample size for items within each sub-test, and relative SEMs were obtained by multiplying the "mean" relative SEMs by the corresponding D study sample size for items within each sub-test. In Table 7, D study sample sizes for items within each sub-test are the same as those in the G study. Equations 17 and 18 were used to compute D study statistics for the scale scores. Table 8 reports means and standard deviations of D study statistics for the six administrations, the three February administrations, and the three July administrations.

From Table 7, generalizability coefficients[5] are moderate for sub-tests (between 0.4 and 0.6), and relatively large for the composite (0.86 to 0.9). Among the six sub-tests, universe score standard deviations and generalizability coefficients for CNL, RLP, and EVD are larger than those for CTR, CRM, and TOR. Relative SEMs for different sub-tests are similar in magnitude.

For the composite score, generalizability coefficients for the six administrations are similar in magnitude (between 0.86 and 0.9). Because the slopes of the conversion functions from raw to scale scores are less than 1, the square roots of the universe score standard deviations and relative SEMs/error variances for the composite scale scores are smaller than those for the raw scores.

Table 8 reports means and standard deviations of G study results based on all six administrations, the three February administrations, and the three July administrations. It can be observed that for the composite score, means of the average universe score standard deviations for February administrations are slightly smaller than those for July administrations. Means of relative SEMs are similar for all six administrations, February administrations, and July administrations, which indicates that relative error SEMs are quite stable across different administrations.

For the six sub-tests, the average relative error SEMs are very similar in magnitude. Universe score standard deviations for CRM and TOR are relatively small, and consequently, coefficients for Criminal Law and Procedure and Torts are relatively small as well. Universe score standard deviations for RLP and EVD are relatively large, and generalizability coefficients for Real property and Evidence are relatively large as well. As expected, for the composite score, relative SEMs are smaller, and generalizability coefficients are larger than those for the sub-tests.

To obtain an estimate of the dependability coefficient with respect to cut score λ for the composite scale scores, the following equation was used:

$$\hat{\Phi}_S(\lambda) \quad = \quad \frac{\hat{\sigma}_S^2(p) + (\overline{X}_S - \lambda)^2 - \hat{\sigma}^2(\overline{X}_S)}{\hat{\sigma}_S^2(p) + (\overline{X}_S - \lambda)^2 - \hat{\sigma}^2(\overline{X}_S) + \hat{\sigma}_S^2(\delta)}, \qquad (19)$$

where $\hat{\sigma}_S^2(p)$ is the estimated universe composite scale score variance, $\overline{X}_S$ is the

---

[5]Values of generalizability coefficients and Cronbach's α agree up to the third decimal places.

Dependability Coefficients
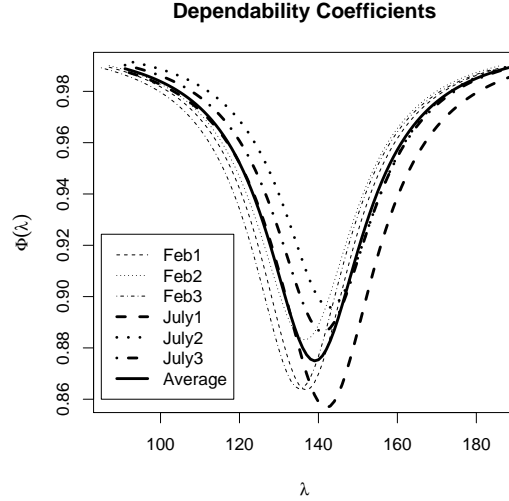
Figure 1: Plot of $\Phi(\lambda)$

observed mean composite scale score, and $\hat{\sigma}^2(\overline{X}_S)$ is the estimated error variance of using the sample mean scale score $\overline{X}_S$ for estimating the population mean scale score μ_S. A simple and direct estimate of $\sigma^2(\overline{X}_S)$ is the observed variance of the observed mean composite scale scores from the six administrations.

In Equation 19, the numerator is an estimate of $\boldsymbol{E}(\mu_p - \lambda)^2$ (see Equation 13) for scale scores, and the denominator uses $\hat{\sigma}_S^2(\delta)$ as an estimate of error variance for scale scores. Here $\hat{\sigma}_S^2(\delta)$ is used rather than $\hat{\sigma}_S^2(\Delta)$ (as might be expected from Equation 13) because the scale score metric involves an equating adjustment for the relative difficulty of the forms. That is, if equating worked perfectly, there would be no difference in the equated mean scores, and therefore, absolute error variance would be identical to relative error variance.[6]

Figure 1 shows dependability coefficients for the six administrations as a function of λ (where λ represents a possible passing score), and for the average of the six administrations. From these figures, it can be observed that the shape of the curves for the dependability coefficients resembles an up-side-down "normal" curve. As values of λ move away from the mean $(\overline{X}_S)$, dependability coefficients increase. When $\overline{X}_S = \lambda$, the dependability coefficient is at the minimum. It can also be observed from these figures that the observed mean composite scale scores for the three February administrations are slightly lower than those for the three July administrations. The difference is about 5 or 6 in the scale score metric. Overall, the dependability coefficients are relatively high (above 0.85).

---

[6]This is a bit of an exaggeration because the equating of the MBE adjusts for the overall difficulty of forms, not the difficulty of each of the sub-tests.

Discussion

One focus of this study was to examine sub-test structures of the MBE using multivariate generalizability theory. There are six sub-tests in the MBE, and there are 33 or 34 questions associated with each sub-test. Using data collected from six administrations of the MBE (three February and three July administrations), a multivariate $p^\bullet \times i^\circ$ analysis (with six sub-tests being fixed categories) was conducted for each of the six datasets.

Data used in the study were from six administrations of the MBE over three years (three February administrations and three July administrations). Mean scale scores for July administrations are higher than the mean scale scores for February administrations (about 5 or 6 points higher in the scale score metric).

G study results indicate that variance and covariance component estimates for February and July administrations are similar and relatively stable for different administrations. Universe scores for the six sub-tests are also highly correlated, especially between Contracts and Real Property, Contracts and Torts, Criminal Law and Procedure and Torts, and Criminal Law and Procedure and Evidence. Variance component estimates for sub-tests for the February and July administrations are very stable, which indicates that the table of specifications for the MBE is well-defined and faithfully followed, and various forms constructed based on the table of specifications are quite "parallel" to each other.

In the D study, a composite score was formed by applying a weight vector to scores for each sub-test, where the weight is determined by the number of items within each sub-test in the D study. D study results were obtained for both raw and scale composite scores in the total score metric. Universe score standard deviations, relative SEMs, generalizability coefficients, and dependability coefficients for the composite score were obtained for each data set.

For the six administrations, generalizability coefficients for the composite scores are relatively high (between 0.86 and 0.90). Generalizability coefficients for the average February and the average July administrations are almost the same. Among the six subtests, generalizability coefficients for RLP, CNL, and EVD seem to be relatively high in magnitude, and generalizability coefficients for CRM and TOR seem to be somewhat lower. Relative SEMs for composite scale scores are also smaller than those for composite raw scores. The reason is that the transformation from equated raw composite scores to scale scores is linear, and the slope of the linear transformation is less than 1.

The dependability coefficients for the composite scale scores are relatively high as well (above 0.85). When the cut score is the same as the observed mean composite scale score, the dependability coefficient is at the minimum of about 0.86.

In this study, raw scores were used in the G study analyses. As mentioned earlier, MBE forms are equated so that the resulting scale scores are on a common scale. Because the transformation function from equated raw scores to scale

17

scores for the composite is linear, generalizability coefficients are unchanged, and it is also relatively easy to convert raw-score universe score variances and relative and absolute error variances/SEMs to scale-score universe score variances and relative and absolute error variances/SEMs.

Reference

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

Brennan, R. L. (2001b). *Manual for mGENOVA.* Iowa City, IA: Iowa Testing Programs, University of Iowa.

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

Jarjoura, D., & Brennan, R. L. (1982). A variance components model for measurement procedures associated with a table of specifications. *Applied Psychological Measurement, 6,* 161–171.

Kane, M. T. (1996). The precision of measurements. *Applied Measurement in Education, 9*, 355–379.