Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 50

Nominal Weights in Multivariate Generalizability Theory[1]

Robert L. Brennan[2]

December 2016

[1]The research reported here was supported, in part, by a contract with the College Board. It should not be inferred, however, that the College Board supports the methodology discussed here. This report is a substantially modified version of CASMA Technical Note No. 4 that was originally published in February 2009.

[2]Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Founding Director, Center for Advanced Studies in Measurement and Assessment (CASMA). He is also a consultant to the College Board.

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: https://education.uiowa.edu/centers/casma

Contents

Multivariate generalizability (G) theory (Brennan, 2001a) provides a powerful and elegant way to: (i) differentiate between random and fixed characteristics of the structure of test forms in conceptualizing a universe of generalization; and (ii) estimate relevant variance and covariance components, as well as quantities such as error variances and reliability-like coefficients for composites of the user's choosing. A fixed facet represents conditions that are present in all forms of a test; a random facet represents conditions that vary over forms of a test. In this paper, to keep matters simple, we will consider a single fixed facet (item-type) with two levels (multiple-choice and free-response items) and a single random facet (items within item-type).

A crucial factor in defining composites is specifying nominal weights for the levels of the fixed facet that reflect the user's intent. Doing so, however, is sometimes complicated for at least two reasons. First, mGENOVA (Brennan, 2001b) is often used to estimate results, but mGENOVA does so for the mean-score metric (as does almost all of the literature on generalizability theory), whereas some users prefer to work with a total score metric. Also, a user's desired composite may be rather complicated, rendering it challenging to define weights for the fixed parts that are consistent with the user's intent. The purpose of this paper is to provide a framework for considering these matters.

Suppose a test consists of n_m multiple-choice (MC) items each of which is scored $[0, 1, \ldots, s_m]$ (almost always $s_m = 1$) and n_f free-response (FR) items scored $[0, 1, \ldots, s_f]$. Let X_m represent MC scores in the total-score metric $[0,1, \ldots, S_m = n_m s_m]$ and X_f represent FR scores in the total-score metric $[0,1, \ldots, S_f = n_f s_f]$. For ease of interpretation, unless otherwise noted, it is assumed that all s and S scores are non-negative integers. Consequently, the s and S constants have two interpretations: (1) highest score; and (2) number of scores plus 1. (Section 4.2 considers some other cases.) Note that it is *not* assumed here that the composite necessarily consists of integer scores.

We begin by considering four sets of nominal weights for forming a composite of MC and FR scores. Usually nominal weights are defined such that they sum to one. Here there is no such restriction. Ultimately, in this paper, all four sets of nominal weights will be expressed subject to the following

> *Constraint*: MC items contribute r times as much to the composite score range as FR items.

This is a rather common constraint, but it is not dictated by multivariate G theory. To simplify the presentation, we begin by considering weighting scenarios that do *not* involve the constraint.

1 Weighting Scenarios

Four weighting scenarios (without considering the constraint) are described next. These scenarios differ in terms of: (i) whether the observed scores for the parts are expressed in the mean-score or total-score metric; and (ii) whether the composite (C) is expressed in a $[0,S_c]$ metric or a $[0,1]$ metric.

(a) Observed Total Scores for Parts with $C \in [0, S_c]$. Suppose an investigator wishes to construct a composite, C, that ranges from 0 to S_c. We will call this the [0, S_c] composite-score metric. For this metric, with nominal weights of w_m and w_f,

$$w_m X_m + w_f X_f = C. \tag{1}$$

(b) Observed Mean Scores for Parts with $C \in [0, S_c]$. Letting the random variables for the mean-score metric be $\overline{X}_m$ and $\overline{X}_f$, and letting the corresponding nominal weights be v_m and v_f, respectively,

$$v_m \overline{X}_m + v_f \overline{X}_f = C. \tag{2}$$

(c) Observed Total Scores for Parts with $C \in [0, 1]$. Using total scores for the parts, the nominal weights w'_m and w'_f are used for transforming the composite [0,S_c] metric to a composite [0,1] metric:

$$w'_m X_m + w'_f X_f = 1. \tag{3}$$

(d) Observed Mean Scores for Parts with $C \in [0, 1]$. Using mean scores for the parts, the nominal weights v'_m and v'_f are used for transforming the composite [0,S_c] metric to a composite [0,1] metric:

$$v'_m \overline{X}_m + v'_f \overline{X}_f = 1. \tag{4}$$

These are the weights that typically would be used in mGENOVA to obtain results for the composite [0,1] metric.

2 Weighting Scenarios with the Constraint

The constraint states that that MC items contribute r times as many (fractional) score points as FR items. Let $u_m = r/(r+1)$ be the proportion of the composite score range S_c associated with MC items and $u_f = 1/(r+1)$ be the proportion associated with FR items. The constraint means that

$$u_m + u_f = 1, \tag{5}$$

where u_m and u_f are called relative weights by Powers and Brennan (2009). Equation 5 can be applied to all four Cases.

2.1 Composite for $[0, S_c]$ Metric

For Case (a) we wish to find w_m and w_f such that

$$w_m S_m + w_f S_f = S_c, \tag{6}$$

given the constraint in Equation 5. Dividing both sides of Equation 6 by S_c gives

$$\frac{w_m S_m}{S_c} + \frac{w_f S_f}{S_c} = 1.$$

The two terms to the left of the equal sign are u_m and u_f, respectively. It follows that

$$w_m = \left(\frac{u_m}{S_m}\right) S_c, \tag{7}$$

and

$$w_f = \left(\frac{u_f}{S_f}\right) S_c. \tag{8}$$

It is important to note that the above results assume that the D study statistics to be used in obtaining composite results are for the *total*-score metric, which is consistent with the use of S_m and S_f in Equation 6. However, total-score metric D study statistics are rarely employed in G theory. Rather, the usual convention is to use mean-score metric D study statistics, as is the case for mGENOVA. An obvious question to ask, then, is, "What weights should we use when composite results will be computed using mean-score metric D study statistics?"

The answer employs the same logic as in the previous section. We merely replace S_m and S_f (total-score metric ranges) with s_m and s_f (mean-score metric ranges). Specifically, to obtain the weights for Case (b) Equations 6–8 are replaced by

$$v_m s_m + v_f s_f = S_c, \tag{9}$$

$$v_m = \left(\frac{u_m}{s_m}\right) S_c, \tag{10}$$

and

$$v_f = \left(\frac{u_f}{s_f}\right) S_c. \tag{11}$$

2.2 Composite for [0,1] Metric

When weighting issues of the type considered here are encountered, in the author's experience the composite is usually expressed in a total score metric, as discussed in Section 2.1. Sometimes, however, multiple tests are under consideration, and there is some interest in comparing composite statistics for the various tests. This is easily accomplished if S_c is the same for all tests, but if the S_c values are different, comparisons are difficult because the composite scales are different. One route around this problem is to use a common value of S_c for all tests *solely* for the purpose of comparing statistics such as estimated error variances. An obvious choice is to put all composites on a [0,1] scale, which is the topic of this section. To do so, all that is required is to set $S_c = 1$ in the equations in Section 2.1.

To obtain the weights for Case (c) Equations 6–8 are replaced by

$$w'_m S_m + w'_f S_f = 1, \tag{12}$$

$$w'_m = \frac{u_m}{S_m}, \tag{13}$$

and

$$w'_f = \frac{u_f}{S_f}. \tag{14}$$

To obtain the weights for Case (d) Equations 9–11 are replaced by

$$v'_m s_m + v'_f s_f = 1, \tag{15}$$

$$v'_m = \frac{u_m}{s_m}, \tag{16}$$

and

$$v'_f = \frac{u_f}{s_f}. \tag{17}$$

3 Example

Table 1 provides an example based on 100 MC items, each of which is scored [0,1], and four FR items, each of which is scored [0,1, ..., 10].[3] The first quarter of the table provides estimates of the G study variance and covariance components. The left side provides results in the mean score metric, which is typical in G theory and is used in mGENOVA. The right side provides results in the total score metric (designated with an appended "+"), where "total score" refers to total score over items. So, for example, $\hat{\sigma}^2(i+) = n^2\,\hat{\sigma}^2(i)$ because total scores are a linear transformation of mean scores. For FR items, it follows that $\hat{\sigma}^2(i+) = 4^2\,\hat{\sigma}^2(i) = 16 \times .70213 = 11.21968$. In the total-score metric, the covariance component for MC and FR items is $100 \times 4 \times .29081 = 116.32400$.

Note that $\widehat{\boldsymbol{\Sigma}}_p$ is a 2×2 matrix with a covariance in the off-diagonal positions. Technically, there are two other matrices, $\widehat{\boldsymbol{\Sigma}}_i$ and $\widehat{\boldsymbol{\Sigma}}_{pi}$, but since covariances for these two matrices are all zero, only the estimated variance components are provided to save space.

The left side of the second quarter of Table 1 provides the usual D study statistics for MC and FR in the mean-score metric. It is important to differentiate $\hat{\sigma}^2(I) = \hat{\sigma}^2(i)/n$ from $\hat{\sigma}^2(i+) = n^2\,\hat{\sigma}^2(i)$. The former is the variance of the distribution of mean scores for items; the later is the variance of a linear transformation of items scores, as discussed above.

The right side of the second quarter of Table 1 provides D study statistics for MC and FR in the total-score metric. They can be obtained from the mean-score statistics in the following manner:

[3]This example has some similarities with a real data example, but it is not real data per se.

Table 1: Example: Weights for $[0, S_c = 150]$ and [0,1] Composite Metrics with $r = 1.5$

	MC	FR		MC	FR
n	100	4		100	4
s	1	10		1	10
S	100	40		100	40
	G study mean-score components			G study total-score components	
$\widehat{\boldsymbol{\Sigma}}_p$	.02782	.29081	$\widehat{\boldsymbol{\Sigma}}_p+$	278.20000	116.32400
	.29081	3.29714		116.32400	52.75424
$\hat{\sigma}^2(i)$	.03388	.70123	$\hat{\sigma}^2(i+)$	338.80000	11.21968
$\hat{\sigma}^2(pi)$	.18434	3.44762	$\hat{\sigma}^2(pi+)$	1843.40000	55.16192
	Mean-score metric D study statistics			Total-score metric D study statistics	
$\widehat{\boldsymbol{\Sigma}}_p$	.02782	.29081	$\widehat{\boldsymbol{\Sigma}}_p+$	278.20000	116.32400
	.29081	3.29714		116.32400	52.75424
$\hat{\sigma}^2(I)$	.00034	.17531	$\hat{\sigma}^2(I+)$	3.38800	2.80492
$\hat{\sigma}^2(pI)$	.00184	.86191	$\hat{\sigma}^2(pI+)$	18.43400	13.79048
$\hat{\sigma}^2(\delta)$	.00184	.86191	$\hat{\sigma}^2(\delta+)$	18.43400	13.79048
$\hat{\sigma}^2(\Delta)$	.00218	1.03721	$\hat{\sigma}^2(\Delta+)$	21.82200	16.59540
$\boldsymbol{E}\hat{\rho}^2$	.93786	.79276	$\boldsymbol{E}\hat{\rho}^2+$	.93786	.79276
$\widehat{\Phi}$	.92727	.76070	$\widehat{\Phi}+$	.92727	.76070
	(b) Composite for [0,150] metric using mean-score metric D study statistics			(a) Composite for [0,150] metric using total-score metric D study statistics	
v	90.00000	6.00000	w	.90000	1.50000
$\hat{\sigma}_c^2(p+)$	658.11384		$\hat{\sigma}_c^2(p+)$	658.11384	
$\hat{\sigma}_c^2(\delta+)$	45.96012		$\hat{\sigma}_c^2(\delta+)$	45.96012	
$\hat{\sigma}_c^2(\Delta+)$	55.01547		$\hat{\sigma}_c^2(\Delta+)$	55.01547	
$\boldsymbol{E}\hat{\rho}_c^2+$	.93472		$\boldsymbol{E}\hat{\rho}_c^2+$	.93472	
$\widehat{\Phi}_c+$	.92285		$\widehat{\Phi}_c+$	.92285	
	(d) Composite for [0,1] metric using mean-score metric D study statistics			(c) Composite for [0,1] metric using total-score metric D study statistics	
v'	.60000	.04000	w'	.00600	.01000
$\hat{\sigma}_c^2(p*)$	.02925		$\hat{\sigma}_c^2(p*)$	.02925	
$\hat{\sigma}_c^2(\delta*)$	.00204		$\hat{\sigma}_c^2(\delta*)$	.00204	
$\hat{\sigma}_c^2(\Delta*)$	.00245		$\hat{\sigma}_c^2(\Delta*)$	.00245	
$\boldsymbol{E}\hat{\rho}_c^2*$	.93472		$\boldsymbol{E}\hat{\rho}_c^2*$	.93472	
$\widehat{\Phi}_c*$	.92285		$\widehat{\Phi}_c*$	.92285	

- multiply MC variance components on the left side of the second quarter of Table 1 by $n_m^2 = 100^2 = 10,000$;
- multiply FR variance components on the left side by $n_f^2 = 4^2 = 16$;
- multiply the person covariance component on the left side by $n_m n_f = 100 \times 4 = 400$; and
- obtain the total-score metric error variances and coefficients from the total score metric D study variance and covariance components using the usual rules (see, for example, Brennan, 2001a, p. 109).

The bottom half of Table 1 provides composite score results associated with the four weighting schemes, (a)–(d), using $r = 1.5$ as the constraint (see page1). The third quarter of Table 1 provides results for the composite metric $[0,S_c = 150]$, and the bottom quarter provides results for the composite metric [0,1]. The [0,1] composite statistics are identified with an appended "$*$". Results labelled (b) and (d) on the left are obtained using mean-score D study statistics; results labelled (a) and (c) on the right are obtained using total-score D study statistics. The next few paragraphs provide a more detailed discussion of computations.

3.1 Case (a)

$$w_m = \left(\frac{.6}{100}\right)150 = .9 \quad \text{and} \quad w_f = \left(\frac{.4}{40}\right)150 = 1.5.$$

Using the total-score metric D study statistics,

$$\begin{aligned}
\hat\sigma_c^2(p+) &= .9^2(278.2) + 1.5^2(52.75424) + 2(.9)(1.5)(116.324) = 658.11384,\\
\hat\sigma_c^2(\delta+) &= .9^2(18.434) + 1.5^2(13.79048) = 45.96012,\\
\hat\sigma_c^2(\Delta+) &= .9^2(21.822) + 1.5^2(16.59540) = 55.01547,\\
\boldsymbol{E}\hat\rho_c^2+ &= 658.11384/(658.11384 + 45.96012) = .93472, \text{ and}\\
\widehat\Phi_c+ &= 658.11384/(658.11384 + 55.01547) = .92285.
\end{aligned}$$

3.2 Case (b)

$$v_m = \left(\frac{.6}{1}\right)150 = 90 \quad \text{and} \quad v_f = \left(\frac{.4}{10}\right)150 = 6.$$

Using the mean-score metric D study statistics,

$$\begin{aligned}
\hat\sigma_c^2(p+) &= 90^2(.02782) + 6^2(3.29714) + 2(90)(6)(.29081) = 658.11384,\\
\hat\sigma_c^2(\delta+) &= 90^2(.00184) + 6^2(.86191) = 45.96012,\\
\hat\sigma_c^2(\Delta+) &= 90^2(.00218) + 6^2(1.03721) = 55.01547,\\
\boldsymbol{E}\hat\rho_c^2+ &= 658.11384/(658.11384 + 45.96012) = .93472, \text{ and}\\
\widehat\Phi_c+ &= 658.11384/(658.11384 + 55.01547) = .92285.
\end{aligned}$$

These results for Case (b) are necessarily the same as those for Case (a). It is important to note, however, that different nominal weights must be used with the two different types of D study variance and covariance components.

3.3 Case (c)

$$w'_m = \frac{.6}{100} = .006 \quad \text{and} \quad w'_f = \frac{.4}{40} = .01.$$

Using the total-score metric D study statistics, it follows that

$$\begin{aligned}
\hat{\sigma}^2_c(p*) &= .006^2(278.2) + .01^2(52.75424) + 2(.006)(.01)(116.324) = .02925, \\
\hat{\sigma}^2_c(\delta*) &= .006^2(18.434) + .01^2(13.79048) = .00204, \\
\hat{\sigma}^2_c(\Delta*) &= .006^2(21.822) + .01^2(16.59540) = .00245, \\
\boldsymbol{E}\hat{\rho}^2_c* &= .02925/(.02925 + .00204) = .93472, \ \text{and} \\
\widehat{\Phi}_c* &= .02925/(.02925 + .00245) = .92285.
\end{aligned}$$

3.4 Case (d)

$$v'_m = \frac{.6}{1} = .6 \quad \text{and} \quad v'_f = \frac{.4}{10} = .04.$$

Using the mean-score metric D study statistics, it follows that

$$\begin{aligned}
\hat{\sigma}^2_c(p*) &= .6^2(.02782) + .04^2(3.29714) + 2(.6)(.04)(.29081) = .02925, \\
\hat{\sigma}^2_c(\delta*) &= .6^2(.00184) + .04^2(.86191) = .00204, \\
\hat{\sigma}^2_c(\Delta*) &= .6^2(.00218) + .04^2(1.03721) = .00245, \\
\boldsymbol{E}\hat{\rho}^2_c* &= .02925/(.02925 + .00204) = .93472, \ \text{and} \\
\widehat{\Phi}_c* &= .02925/(.02925 + .00245) = .92285.
\end{aligned}$$

These results for Case (d) are necessarily the same as those for Case (c).

3.5 Similarities and Differences Among the Cases

The generalizbility coefficients are the same for all four Cases, as are the Phi coefficients. This equivalence necessarily follows from the fact that all four transformations are linear and apply to both the numerator and the denominator of the coefficients.

By contrast, that the error variances for Cases (a) and (b) are substantially different from the error variances for Cases (c) and (d). This follows from that fact that error variances are different for the mean-score and total-score metrics. Note that the corresponding standard errors of measurement are on the scale of the composite ([0,150] for the third quarter of Table 1, and [0,1] for the bottom quarter of the table).

4 Other Considerations

For simplicity, to this point, all score ranges have been defined such that the lowest score is 0. Obviously, any score range is unchanged if the same constant is added to all scores in the range.[4] For example, if MC items were scored [1,2], the results in Table 1 would be unchanged. As discussed next, a somewhat more challenging issue arises if the variables X_m and/or X_f are linearly transformed.

4.1 Linear Transformations of X_m and/or X_f

Consider again Case (a), which uses observed total scores for the parts with $C \in [0, S_c]$. Suppose $\mathcal{X}_m = a_m + b_m X_m$ and $\mathcal{X}_f = a_f + b_f X_f$, and assume that the nominal weights (which we will designate $\ddot{w}_m$ and $\ddot{w}_f$) apply to these transformed variables. That is,

$$\ddot{w}_m \mathcal{X}_m + \ddot{w}_f \mathcal{X}_f = C.$$

To keep matters simple, we will also assume that u_m, u_f, and S_c are unchanged. Since the range is unaffected by the addition of a constant to all scores within the range, it is sufficient to consider $\mathcal{X}_m = b_m X_m$ and $\mathcal{X}_f = b_f X_f$.

Let $\ddot{S}_m = b_m S_m$ and $\ddot{S}_f = b_f S_f$. We wish to find $\ddot{w}_m$ and $\ddot{w}_f$ such that

$$\ddot{w}_m \ddot{S}_m + \ddot{w}_f \ddot{S}_f = C,$$

subject to the constraint in Equation 5.

Following the same logic as in Section 2.1 [Case (a)], we obtain

$$\ddot{w}_m = \left(\frac{u_m}{\ddot{S}_m}\right) S_c = \left(\frac{u_m}{b_m S_m}\right) S_c = \frac{w_m}{b_m}, \tag{18}$$

where the last result follows from Equation 7. Similarly,

$$\ddot{w}_f = \left(\frac{u_f}{\ddot{S}_f}\right) S_c = \left(\frac{u_f}{b_f S_f}\right) S_c = \frac{w_f}{b_f}. \tag{19}$$

In this sense, $\ddot{w}_m$ and $\ddot{w}_f$ absorb the slopes of the transformations.

Returning to the example in Table 1 and Section 2.1 [Case (a)], the computations proceed in the same manner except that w_m and w_f are replaced by $\ddot{w}_m$ and $\ddot{w}_f$, respectively. In particular, the *same* total-score metric D study statistics for MC and FR items are used; these statistics are *not* multipled by the slopes.

The same type of logic applies to Cases (b)–(d). For example, for Case (b) the analogue of Equation 2 is

$$\ddot{v}_m \overline{\mathcal{X}}_m + \ddot{v}_f \overline{\mathcal{X}}_f = C;$$

[4] The focus of this paper is composite score *variances* or functions of them. Clearly results are unchanged if the same constant is added to all scores.

and the weights are

$$\ddot{v}_m = \frac{v_m}{b_m} \tag{20}$$

and

$$\ddot{v}_f = \frac{v_f}{b_f}, \tag{21}$$

where b_m and b_f are the slopes of the transformations for converting $\overline{X}_m$ to $\overline{\mathcal{X}}_m$ and $\overline{X}_f$ to $\overline{\mathcal{X}}_f$. In this instance, computations use the mean-score metric D study statistics for MC and FR items, and these statistics are *not* multiplied by the slopes.

4.2 Item-score Increments Other than 1

Since s_m, s_f, S_m, S_f, and S_c are all ranges, the results provided in this paper do not require that contiguous scores all differ by 1, although the early part of this paper used 1 as the increment. For example, a three-point rubric might have scores of [.1, .2, .3] or [10, 20, 30]. As illustrated below, however, when the increments for MC and FR item scores differ, the composite score scale may have characteristics judged undesirable.

Suppose, there are three MC items scored [0, 1] and two FR items scored [10, 20, 30], with $u_m = .4$, $u_f = .6$, and $S_c = 90$. Using the results in Case (a) it is easy to determine that $w_m = 12$, $w_f = 1.35$, and the possible composite scores are:

27.0	39.0	40.5	51.0	52.5	54.0	63.0	64.5	66.0	67.5
76.5	78.0	79.5	81.0	90.0	91.5	93.0	103.5	105.0	117.0

Note that the distances between contiguous scores differ dramatically, varying between 1.5 and 12.0. This characteristic of the composite score scale may render it less than ideal in some circumstances.

Also, the results derived in this paper do not require that there be a constant difference between contiguous item scores. So, for example, these results apply to formula-scored MC items[5] where the possible scores are [-1/(k-1), 0, 1] with k being the number of alternatives. In such cases, however, almost certainly the composite score scale will have characteristics similar to (or even more extreme than) the example in the preceding paragraph.

4.3 Effective Weights

Effective weights in multivariate G theory are discussed by Brennan (2001a, pp. 306–307). In general, an effective weight provides the proportion of a particular composite variance attributable to a level of a fixed facet. The effective weights considered here are the same for Cases (a)–(d).

[5]The author is not recommending use of formula scoring.

Relative to composite universe score variance, the effective weights for MC and FR items, expressed in terms of Case (a), are:

$$ew_m(\tau) = \frac{w_m^2\,\sigma_m^2(p+) + w_m w_f\,\sigma_{mf}(p+)}{\sigma_c^2(p+)} \tag{22}$$

and

$$ew_f(\tau) = \frac{w_m w_f\,\sigma_{mf}(p+) + w_f^2\,\sigma_f^2(p+)}{\sigma_c^2(p+)}, \tag{23}$$

respectively. For the example,

$$\widehat{ew}_m(\tau) = \frac{.9^2(278.2) + .9(1.50)(116.324)}{658.11384} = .58102$$

and

$$\widehat{ew}_f(\tau) = \frac{.9(1.5)(116.324) + 1.5^2(52.75424)}{658.11384} = .41898.$$

Relative to composite relative error variance, the effective weights for MC and FR items, expressed in terms of Case (a), are:

$$ew_m(\delta) = \frac{w_m^2\,\sigma_m^2(\delta+)}{\sigma_c^2(\delta+)} \quad \text{and} \quad ew_f(\delta) = \frac{w_f^2\,\sigma_f^2(\delta+)}{\sigma_c^2(\delta+)}.$$

Relative to composite absolute error variance, the effective weights for MC and FR items, are:

$$ew_m(\Delta) = \frac{w_m^2\,\sigma_m^2(\Delta+)}{\sigma_c^2(\Delta+)} \quad \text{and} \quad ew_f(\Delta) = \frac{w_f^2\,\sigma_f^2(\Delta+)}{\sigma_c^2(\Delta+)}.$$

The following table provides estimated effective weights for the example along with u_m and u_f (see Equation 5).

	MC	FR
u	.6	.4
$\widehat{ew}(\tau)$	.58102	.41898
$\widehat{ew}(\delta)$	.32488	.67512
$\widehat{ew}(\Delta)$	.32129	.67871

It is evident that the effective weights relative to composite universe score variance are quite close to the proportional weights for the composite score range. However these are quite different from the effective weights relative to composite error variance (both the δ and Δ types), which clearly indicate that FR items contribute more that twice as much as MC items to error variance.

4.4 Composite Generalizability Coefficient and Stratified α

As noted previously and illustrated in Table 1, $(\boldsymbol{E}\hat{\rho}^2_c+) = (\boldsymbol{E}\hat{\rho}^2_c*)$, since all the transformations considered here are linear. Furthermore, as discussed below, for the design considered in this paper, these coefficients are identical to stratified α (Rajaratnam, Cronbach, & Gleser, 1965).

In the terminology and notation of generalizability theory, the design discussed in this paper is the multivariate $p^\bullet \times I^\circ$ design (see Brennan, 2001a, sect. 9.1). This notation means that: (i) for both MC and FR items, the design is $p \times i$; (ii) persons respond to both types of items (signified by the closed circle following p); (iii) any given item is either MC or FR (signified by the open circle following i)—i.e., items are nested within levels of the fixed item-type facet). For this design, Jarjoura and Brennan (1982, 1983) demonstrated that the generalizability coefficient is identical to stratified α.

In terms of the notation for the $[0,S_c]$ metric used in this paper,

$$\hat{\alpha}_{strat} = 1 - \frac{w_m^2\,\hat{\sigma}^2_m(\delta+) + w_f^2\,\hat{\sigma}^2_f(\delta+)}{\hat{\sigma}^2_c(p+) + \hat{\sigma}^2_c(\delta+)}, \tag{24}$$

and for the example

$$\hat{\alpha}_{strat} = 1 - \frac{.9^2(18.43400) + 1.5^2(13.79048)}{658.11384 + 45.96012} = .93472,$$

which is identical to $\boldsymbol{E}\hat{\rho}^2_c+$. Changing "+" to "$*$" in Equation 24 gives the corresponding equation for the [0,1] metric, but does not alter $\hat{\alpha}_{strat}$.

Whether $\boldsymbol{E}\hat{\rho}^2_c+$, $\boldsymbol{E}\hat{\rho}^2_c*$, or $\hat{\alpha}_{strat}$ is used, of course, the nominal weights must be determined, and that is the principal purpose of this paper. Beyond that, however, since all three coefficients are equal, one might ask, "Why even consider the more complicated multivariate G theory approach in this paper?" There are several reasons. First, the multivariate G theory approach provides a much richer set of results than a single coefficient. Second, the multivariate G theory approach for the $p^\bullet \times i^\circ$ design considered here can be used directly with mGENOVA. Third, and probably most importantly, the approach is easily generalized to many other multivariate designs, the most common of which are accommodated by mGENOVA.

5 Concluding Comments

This paper has considered four types of nominal weights in multivariate G theory that might be used to form a composite of MC and FR scores, where the composite has a specified range, and weights are subject to the constraint that MC items contribute r times as much to the composite score range as FR items. Usually nominal weights are defined such that they sum to one. Here there is no such restriction.

Two sets of weights [Cases (b) and (d)] use mean-score metric D study statistics, which makes them directly applicable using mGENOVA. The other two sets of weights, [Cases (a) and (c)] use total-score metric D study statistics. Composite results for Cases (a) and (b) are identical; a similar statement holds for Cases (c) and (d). Composite generalizability coefficients and Phi coefficients are the same for all four Cases.

The types of weights considered in this paper are used in some operational testing programs and appear reasonable in certain circumstances. However, the author offers no general endorsement of such weights.

To keep the discussion relatively simple, this paper has assumed that a composite consists of only two fixed parts or types of items, MC and FR. Results are easily generalized to more than two levels of a single fixed facet, or even multiple fixed facets.

Also, to keep the discussion relatively simple, it has been assumed that s_f is a constant for all n_f FR items. If this is not true, one alternative is to treat all FR items with the same s_f as belonging to the same section or stratum. Then there will be as many FR strata as there are values of s_f. Obviously, is this approach is adopted, then there will have to be weights associated with each of these strata, and choosing those weights will involve the types of considerations addressed in this paper.

6 References

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

Brennan, R. L. (2001b). *mGENOVA* (Version 2.1) [Computer software and manual]. Iowa City, IA: University of Iowa. (Available on https://education.uiowa.edu/centers/casma)

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.

Jarjoura, D. & Brennan, R. L. (1983). Multivariate generalizability models for tests developed according to a table of specifications. In L. J. Fyans (Ed.), *New directions for testing and measurement: Generalizability theory* (No.18) (pp. 83–101). San Francisco: Jossey-Bass.

Powers, S., & Brennan, R. L. (2009, September). *Multivariate Generalizability Analyses of Mixed-format Exams.* (CASMA Research Report No. 29). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/centers/casma)

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika*, *30*, 39–56.