A Variance Components Model for Measurement Procedures Associated with a Table of Specifications

David Jarjoura and Robert L. Brennan
American College Testing Program

Although many tests are developed according to a table of specifications, the literature contains little guidance to measurement specialists for considering the measurement properties of tests developed in this manner. Rather, most of the literature makes the simplistic assumption that the entire set of items are drawn from (or represent) a common undifferentiated domain or universe. This paper presents a variance components model for many measurement procedures that are associated with a table of specifications. In addition, simple procedures are provided for estimating the model parameters (variance components and category means) and functions of them (e.g., composite universe score variance and error variances).

The variance components model presented in this paper focuses on the many test forms that can be generated from a table of specifications; and, thus, the components of interest are not associated with a particular test form but with the overall measurement procedure. Consequently, consistency of measurement among test forms is directly tied to the categories of a table of specifications, and these categories are *fixed* conditions for a measurement procedure. Typically, a specified number of items are selected or generated for each of the categories; and since items differ among test forms, items constitute a random dimension in a measurement procedure. From this perspective the proposed model is within a purview of generalizability theory (see Brennan, in press; Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

The administration of a test form can be represented by a design in which some sample of persons responds to a set of items that fall into separate categories. Usually, such a design is unbalanced in the sense that the number of items associated with each of the categories is not the same. However, the usual complications associated with estimating variance components from unbalanced designs (see Searle, 1971, chaps. 10 & 11) do not pose a problem here (see Brennan, Jarjoura, & Deaton, 1980). The restricted nature of the unbalancing makes possible a straightforward estimation procedure based on a simple set of mean square and mean product statistics.

Modeling tests that are developed according to a table of specifications certainly has precedents but appears not to have received much attention in the measurement literature. Here, the model is developed from a multivariate perspective that is related to Scheffé's (1959, chap. 8) balanced mixed

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 6, No. 2, Spring 1982, pp. 161–171

0146-6216/82/020161-11$1.55

model and to multivariate generalizability theory (Cronbach et al., 1972, chaps. 9 & 10). The problem was first put into a generalizability theory framework by Rajaratnam, Cronbach, and Gleser (1965), but they focused on the derivation of a generalizability coefficient and did not fully develop the variance components model.

Besides modeling and estimation, the following topics will also be discussed: (1) choice of weights for defining a universe score of interest; (2) contributions that categories of the table of specifications make to universe score variance and error variance; (3) options for estimating universe scores and their error variances; (4) optimal choices for the number of items to be used in each category to minimize error variance; and (5) error variance when a test developer deviates from the table of specifications. First, these topics are introduced from a theoretical perspective, and then they are illustrated using data from the Mathematics Subtest of the ACT Assessment Program.

The Model

The sampling model involves the potential observations that would result from the administration of a potential test form. These observations correspond to the responses of P persons to I_+ items. The items fall into C fixed categories or cells of a table of specifications with I_c items in category c, so that $I_+ = \Sigma_{c=1}^{C} I_c$. (Whether the table is considered unidimensional or multidimensional does not affect the results discussed below.) The potential item-level observations are

$$Y_{pic} = \mu_c + \pi_{pc} + \beta_{i:c} + \pi\beta_{pi:c} ,$$

$$p=1, \ldots, P; \quad i=1, \ldots, I_c; \quad c=1, \ldots, C, \qquad [1]$$

where Y_{pic} represents the response of the p^{th} person to the i^{th} item in category c. The μ_c for the C categories are fixed effects or means associated with the population. The other effects are random and are assumed to have expectations of zero. The π_{pc} represent universe score effects, the $\beta_{i:c}$ represent item effects, and the $\pi\beta_{i:c}$ represent residuals. The colon (i:c, pi:c) is used to indicate that i is nested within c.

The variance and covariance components of interest are defined as expectations (over the population of persons and universes of items) of squares and products of the random effects. It is assumed that whatever the sampling process associated with generating the observations, the expectations taken over the sampling distribution are equal to the components of interest. For the universe score effects,

$$\mathcal{E}\pi_{pc}\pi_{pc'} = \sigma(\pi)_{cc'} , \quad c,c'=1, \ldots, C. \qquad [2]$$

With $c = c'$, this equation gives universe score variance for category c; and with $c \neq c'$, it gives the covariance between universe scores for categories c and c'. Sometimes these variance and covariance components will be referred to by the matrix Σ. For item effects,

$$\mathcal{E}\beta^2_{i:c} = \sigma^2(\beta)_c , \quad c=1, \ldots, C; \qquad [3]$$

and for the residual effects,

$$\mathcal{E}\pi\beta^2_{pi:c} = \sigma^2(\pi\beta)_c , \quad c=1, \ldots, C. \qquad [4]$$

Finally, it is assumed that all effects, except for the π_{pc} and $\pi_{pc'}$ in Equation 2, are uncorrelated.

Explicit definitions of effects, in terms of a random sampling process, are a distinguishing feature of generalizability theory. Thus, the mean for category c (μ_c) represents the expected value over random samples of both persons (from the population) and items (from the universe associated with category c). Similarly, the universe score effect for person $p*$ on category c is defined as the expectation over item samples (random) from category c minus μ_c, i.e.,

$$\pi_{p*c} \equiv \mathcal{E}(Y_{pic}|p=p*) - \mu_c \quad , \qquad [5]$$

and the effect for item $i*$ is defined as

$$\beta_{i*:c} \equiv \mathcal{E}(Y_{pic}|i=i*) - \mu_c \quad . \qquad [6]$$

The residual effects, $\pi\beta_{pi:c}$, can be described as person-item interaction plus response error:

$$\pi\beta_{pi:c} \equiv \gamma_{pi:c} + \varepsilon_{pi:c} \quad . \qquad [7]$$

$\gamma_{pi:c}$ represents an interaction effect that is strictly a function of the person and item selected, while $\varepsilon_{pi:c}$ respresents response error. Thus, the $\varepsilon_{pi:c}$ effects are associated with the response process rather than the random sampling process. The response errors are defined here to have expectations of zero for any given person-item combination. It follows that the interaction for person $p*$ and item $i*$ is given as

$$\gamma_{p*i*:c} \equiv \mathcal{E}(Y_{pic}|p=p* , i=i*) - \pi_{p*c} - \beta_{i*:c} - \mu_c \quad . \qquad [8]$$

Thus, to motivate the assumptions of the model, it is sufficient (though not necessary) to assume that the observations are obtained from a simple random sample of persons and from c simple random samples of items from the C categories (with the samples selected independently of one another). Also, the response error effects, $\varepsilon_{pi:c}$, are assumed to be uncorrelated.

Estimation of Variance and Covariance Components

Presented below is a simple procedure for estimating Σ and the two sets of variance components denoted by $\sigma^2(\beta)_c$ and $\sigma^2(\pi\beta)_c$. Attention is restricted here to data from a single administration of a single test form. (In a subsequent example, data from multiple test forms are considered.) Beyond assuming that the data conform to the model and that $I_c \geqslant 2$ for all c, no further assumptions are made. In particular, no distributional form assumptions are made. The statistics used for estimation are linear functions of mean squares and products of the observations, and the estimators are unbiased because they are derived from the expectations of the mean squares and products. (For details of the derivations, see Jarjoura & Brennan, 1981.) With some additional very weak assumptions (e.g., finite fourth moments), consistency of the estimators can be shown also.

For estimating the $\sigma^2(\pi\beta)_c$ and the $\sigma^2(\beta)_c$, each category of a table of specifications can be associated with a separate persons-crossed-with-items ($p \times i$) design. Note that this results in C *balanced* $p \times i$ designs.

The estimators for the set of $\sigma^2(\pi\beta)_c$ are the usual "residual" mean squares for these C $p \times i$ designs:

$$\hat{\sigma}^2(\pi\beta)_c = MS(pi)_c \quad , \quad c=1, \ldots, C. \qquad [9]$$

Similarly, the $\sigma^2(\beta)_c$ are estimated by using the mean squares for items from the C $p \times i$ designs:

$$\hat{\sigma}^2(\beta)_c = [MS(i)_c - MS(pi)_c]/P \quad , \quad c=1, \ldots, C. \qquad [10]$$

To obtain estimates of the $\sigma(\pi)_{cc'}$, the variance-covariance matrix of category mean scores (for persons) is used. This matrix is denoted $\mathbb{S}$, with typical element

$$s_{cc'} = \sum_{p}^{P} (Y_{p\cdot c} - Y_{\cdot\cdot c})(Y_{p\cdot c'} - Y_{\cdot\cdot c'})/(P-1)$$

$$c,c'=1,\ \ldots,\ C; \qquad [11]$$

where $Y_{p\cdot c} = \Sigma_i^{I_c}\ Y_{pic}/I_c$ and $Y_{\cdot\cdot c} = \Sigma_p^P \Sigma_i^{I_c}\ Y_{pic}/(PI_c)$. Estimators of the covariance components in Σ (the off-diagonals) are

$$\hat{\sigma}(\pi)_{cc'} = s_{cc'} \quad \text{for } c \neq c' \quad , c,c'=1,\ \ldots,\ C; \qquad [12]$$

and estimators of the variance components in Σ (the diagonals) are

$$\hat{\sigma}(\pi)_{cc} = s_{cc} - \hat{\sigma}^2(\pi\beta)_c/I_c \quad , \quad c=1,\ \ldots,\ C. \qquad [13]$$

Finally, for estimating the μ_c, the observable means of the categories are used:

$$\hat{\mu}_c = Y_{\cdot\cdot c} \quad , \quad c=1,\ \ldots,\ C. \qquad [14]$$

Defining a Composite Universe Score

As indicated by Equation 5, the variable $\mu_c + \pi_{pc}$ is the expectation of observations for some person p and category c. This universe score for category c is usually not of special interest for measurement procedures associated with a table of specifications. Rather, a single score that takes the form of a linear composite of the C universe scores is of primary interest.

Generally, it seems appropriate to define a composite universe score in terms of a set of a priori weights that reflect the relative importance of each of the categories. Such a composite universe score is

$$\mu_{p\cdot} \equiv \sum_c w_c\ (\mu_c + \pi_{pc}) \quad , \qquad [15]$$

where the w_c (c = 1, ..., C) are the a priori weights and Σ_c denotes $\Sigma_{c=1}^{C}$. For convenience set $\Sigma_c\ w_c = 1$, and without too much loss of generality, $w_c \geqslant 0$ for all c. Thus, the w_c take the form of proportional weights.

The mean of such composite universe scores is simply $\mu. = \Sigma_c\ w_c\ \mu_c$, and the variance $[\varepsilon(\Sigma_c w_c\ \pi_{pc})^2]$ is

$$\sigma^2(\pi)_{\circ} = \sum_c \sum_{c'} w_c\ w_{c'}\ \sigma(\pi)_{cc'} \quad . \qquad [16]$$

Rather than being explicitly defined, the w_c are often implied by the numbers of items that are associated with the categories of a table of specifications. When the reported score is represented by the simple sum of item scores across all I_+ items (divided, say, by I_+), it can be claimed that the implied weights are $w_c = I_c/I_+$, c = 1, ..., C. This popular practice is reasonable when a test developer samples a relatively large number of items for a category that is judged more important than others. However, this practice ignores the possibility of choosing the w_c and the I_c independently.

The w_c are used here to define $\mu_{p\cdot}$ and to reflect the relative importance of the categories in the composite. This allows a free choice of the I_c, and later it is shown that the I_c can be chosen to minimize error variance associated with estimating $\mu_{p\cdot}$. This approach also allows for changes in the I_c from one test form to another while retaining the same definition of the composite universe score.

Many issues are encountered in choosing the w_c. One which seems especially relevant to variance components modeling is the issue of “effective” weights. This issue is related to the fact that the w_c are only partial determiners of the contributions that categories make to $\sigma^2(\pi)$. . The elements of Σ must be considered, also. In their review on weighting, Wang and Stanley (1970) distinguish between the “nominal” and the “effective” weight of a variable in a composite. The nominal weight of a variable can be associated with the w_c. They define the effective weight as the covariance between a weighted variable and the composite variable and refer to it as the variable's contribution to the variance of a composite. Here the effective weight of the c^{th} category is

$$\mathrm{COV}_c = \mathcal{E}(w_c \pi_{pc} \sum_{c'} w_{c'} \pi_{pc'}) = w_c \sum_{c'} w_{c'} \sigma(\pi)_{cc'} \quad . \qquad [17]$$

Note that $\Sigma_c \mathrm{COV}_c = \sigma^2(\pi)$. . Clearly, the w_c and the COV_c will not, in general, be the same.

Estimation of Composite Universe Scores

Choosing the w_c to *define* a composite universe score does not imply that these weights must be used to *estimate* $\mu_{p\cdot}$. Here, however, attention is restricted to the use of the w_c as estimation weights, with the I_c free to be chosen to minimize measurement error variance. The w_c appear to be a reasonable choice because they can be used to specify a simple unbiased estimator of a given person's composite universe score:

$$\hat{\mu}_{p\cdot} = \sum_c w_c Y_{p\cdot c} \quad . \qquad [18]$$

The error associated with this estimator of $\mu_{p\cdot}$ is $\Delta_p \equiv \hat{\mu}_{p\cdot} - \mu_{p\cdot}$; and its variance, $\varepsilon\Delta_p^2$, across samples of persons and items for a given pattern of the I_c is

$$\sigma^2(\Delta) = \sum_c w_c^2 \, [\sigma^2(\beta)_c + \sigma^2(\pi\beta)_c]/I_c \quad . \qquad [19]$$

Since $\sigma^2(\Delta)$ depends on the I_c, the perspective might be taken that the best choice for the I_c, given a fixed value of I_+, is a choice that minimizes $\sigma^2(\Delta)$. Noting that the I_c are positive integers, the minimization can be approached by iteration, given sufficiently precise estimates of the variance components. However, given precise estimates, with I_+ fixed, a simpler solution can be obtained through the following inequality:

$$\begin{aligned} \sigma^2(\Delta) &= \sum_c w_c^2 \, [\sigma^2(\beta)_c + \sigma^2(\pi\beta)_c]/I_c \\ &\geq \left[\sum_c w_c \sqrt{\sigma^2(\beta)_c + \sigma^2(\pi\beta)_c} \right]^2 \Big/ I_+ \quad . \qquad [20] \end{aligned}$$

The I_c can be chosen so that these two terms are equal, which insures that $\sigma^2(\Delta)$ is a minimum. A simple choice that makes them equal is

$$I_c = I_+ w_c \sqrt{\sigma^2(\beta)_c + \sigma^2(\pi\beta)_c} \Big/ \sum_c w_c \sqrt{\sigma^2(\beta)_c + \sigma^2(\pi\beta)_c} \; ,$$

$$c = 1, \ldots, C. \qquad [21]$$

Given that the I_c must be integer valued, it is possible to iterate around the solution in Equation 21 to find a reasonable solution for practical use.

When the w_c are not used as estimation weights, the form of the mean-squared error of measurement changes dramatically. Suppose a different set of weights, a_c ($c = 1, \ldots, C$), is used for estimating composite universe scores. Such estimates could take the form $\tilde{\mu}_{p\cdot} = \sum_c a_c Y_{p\cdot c}$. The resulting mean-squared error is

$$\begin{aligned} \mathcal{E}(\tilde{\mu}_{p\cdot} - \mu_{p\cdot})^2 &= \sum_c \sum_{c'} (a_c - w_c)(a_{c'} - w_{c'})\sigma(\pi)_{cc'} \\ &+ \left[\sum_c (a_c - w_c)\mu_c\right]^2 \\ &+ \sum_c a_c^2 [\sigma^2(\beta)_c + \sigma^2(\pi\beta)_c]/I_c \quad , \end{aligned} \qquad [22]$$

with $a_c = 0$ if there are no items in category c. Thus, there is a non-negative contribution to Equation 22 from the elements of Σ and from the category means (μ_c). For example, the estimation weights could be $a_c = I_c/I_+$, that is, a simple average score could be taken across all I_+ items. Equation 22 would then be useful if the I_c/I_+ fail to correspond precisely to the a priori w_c.

Equation 22 can also be used to gauge the degree to which a table of specifications serves to reduce measurement error. For example, suppose a table of specifications were ignored and all the items for a test form were sampled from just a single category. The resulting increase in error variance can be examined by comparing the mean-squared error, Equation 22, to $\sigma^2(\Delta)$, Equation 19.

Illustrative Analysis for the Mathematics Subtest of the ACT Assessment Program

Generation of multiple forms of the Mathematics Subtest of the ACT Assessment Program (MATH) involves the use of a table of specifications. Thus, MATH is an example of a measurement procedure that might be described by the model. Though all categories in the actual table of specifications are not included in the following analysis, a major classification of the items is represented. Of course, this illustrative analysis is only an incomplete treatment of the measurement procedure.

MATH "emphasizes the solution of practical quantitative problems which are encountered in many post-secondary curricula and includes a sampling of mathematical techniques covered in high school courses" (American College Testing Program, 1980, p. 3). Each of the 40 items of any form of MATH can be classified into one of five categories: Arithmetic and Algebraic Operations (AAO), Arithmetic and Algebraic Reasoning (AAR), Geometry (GEO), Intermediate Algebra (IA), and Number and Numeration Concepts and other Advanced Topics (NA). The numbers of items in these categories are 4, 14, 8, 8, and 6, respectively; and this pattern is constant across all recent forms. Because a simple proportion-correct score is used to arrive at a scaled score, it is presumed here that the $I_c = w_c I_+$; that is, the choice of item numbers is viewed as a direct reflection of the weights that define the composite universe score of interest. Thus, the w_c are .10, .35, .20, .20, and .15, respectively.

The data analyzed here were from eight recent forms of MATH. These data are in the form of right/wrong (1/0) responses. Note that such binary data do not violate assumptions of the model. Each form was administered to independent random samples of approximately 4,500 persons. (Data from such large samples are routinely collected for equating purposes.)

Estimates for MATH

Using the estimators in Equations 9, 10, 12, 13, and 14, estimates of the model components and means were obtained for each of the forms. Then, these estimates were simply averaged over the eight forms to obtain the final estimates for MATH. Other combinations of the multiple form data are possible, but taking a simple average proved convenient for obtaining estimates of the standard errors of the component estimates and category means. These were obtained by calculating the standard deviation of estimates across the eight forms and then dividing by the square root of eight.

Table 1 provides the estimates obtained from a *single* form of MATH. These are presented solely to illustrate the estimation procedure described above. The left section of Table 1 contains the $\mathbb{S}$ matrix and the mean squares with their degrees of freedom. The resulting estimates are in the right section of the table.

The averages of such estimates over the eight forms are provided in Table 2, along with estimates of category means. Note that these average estimates, such as $\bar{\Sigma}$, are denoted with a bar. Table 2 also reports the estimated standard errors, which provide a basis for making inferences about the relative size of components.

From $\bar{\Sigma}$, it is clear that the covariances are close in size to the variances, indicating high correlations among universe scores of different categories. The average of the universe score correlations calculated from $\bar{\Sigma}$ is .93; the highest correlation is .97 (IA with NA); and the lowest is .88 (AAO with AAR). Generally, universe scores for NA have the highest correlations with other categories and those for AAO have the lowest. Some implications of these high correlations are discussed later in the context of potential deviations from the table of specifications. Note also that AAO appears to have a larger universe score variance than the other categories.

From the $\bar{\sigma}^2(\beta)_c$, GEO, IA, and NA appear to have item effect variances that are very close in value, while AAO and AAR appear to have lower and higher values, respectively. From the $\bar{\sigma}^2(\pi\beta)_c$, the residual effect variances of four of the categories are very close, while AAO appears to have lower residual variance. From the $\bar{\mu}_c$, AAO appears to be an easier category than the others, while GEO and IA appear more difficult.

Table 1

Estimates from a Single Form of MATH

Source	df	Mean Square and Product Statistics						Estimates of Variance and Covariance Components				
		.097						.057				
		.041	.054					.041	.040			
$\underset{\sim}{S}$	4557	.036	.037	.059			$\hat{\underset{\sim}{\Sigma}}$	.036	.037	.035		
		.037	.033	.032	.053			.037	.033	.032	.029	
		.046	.037	.033	.033	.066		.046	.037	.033	.033	.037
$MS(i)_c$							$\hat{\sigma}^2(\beta)_c$					
c=1	3	67.370					c=1	.015				
c=2	13	87.277					c=2	.019				
c=3	7	61.880					c=3	.014				
c=4	7	140.305					c=4	.031				
c=5	5	128.520					c=5	.028				
$MS(pi)_c$							$\hat{\sigma}^2(\pi\beta)_c$					
c=1	13671	.158					c=1	.158				
c=2	59241	.190					c=2	.190				
c=3	31899	.197					c=3	.197				
c=4	31899	.193					c=4	.193				
c=5	22785	.175					c=5	.175				

Table 2
Estimates of Variance and Covariance
Components and Category Means[a]

	AAO	AAR	GEO	IA	NA
$\bar{\underset{\sim}{\Sigma}}$					
	.052				
	.004				
	.039	.038			
	.001	*.002*			
	.040	.037	.041		
	.001	*.002*	*.002*		
	.043	.036	.039	.042	
	.003	*.003*	*.003*	*.005*	
	.041	.035	.037	.038	.035
	.002	*.002*	*.002*	*.003*	*.004*
$\bar{\sigma}^2(\beta)_c$					
	.013	.027	.021	.022	.022
	.004	*.003*	*.002*	*.003*	*.003*
$\bar{\sigma}^2(\pi\beta)_c$					
	.176	.185	.187	.183	.187
	.004	*.003*	*.002*	*.003*	*.003*
$\bar{\mu}_c$					
	.607	.501	.456	.446	.501
	.014	*.020*	*.012*	*.018*	*.017*

[a] $\bar{\underset{\sim}{\Sigma}}$ is used to denote the average of the estimates ($\hat{\underset{\sim}{\Sigma}}$) across the eight forms. A similar notation is used to denote the other estimates. The corresponding estimated standard errors are in italics.

Comparing the size of different types of components, it is clear that the residual variances are very much larger than both universe score variances and item effect variances. The large residual variances suggest that many items are required for precise measurement.

Error variance of $\hat{\mu}_p$.

The error variance, $\sigma^2(\Delta)$, can be estimated by substituting the $\bar{\sigma}^2(\pi\beta)_c$ and the $\bar{\sigma}^2(\beta)_c$ for the parameters in Equation 19. The resulting estimate is $\bar{\sigma}^2(\Delta) = .005172$. Recall that for this example the $I_c = w_c I_+$; thus, $\sigma^2(\Delta)$ is the error variance of a proportion-correct score.

To gauge the precision of $\hat{\mu}_{p\cdot}$, $\bar{\sigma}^2(\Delta)$ can be compared to the estimate of composite universe score variance. Substituting the elements of $\bar{\Sigma}$ into Equation 16, $\bar{\sigma}^2(\pi)_{\cdot} = .0380$. Thus, universe score variance is approximately 7.3 times larger than error variance, indicating fairly precise measurement for MATH. In other words, the covariance between $\hat{\mu}_{p\cdot}$ and $\mu_{p\cdot}$ is 7.3 times larger than the variance of the difference $\hat{\mu}_{p\cdot} - \mu_{p\cdot}$.

Contributions to $\bar{\sigma}^2(\pi)$.

Estimates of the effective weights of the categories, denoted $\overline{\mathrm{COV}}_c$, are .0042, .0129, .0077, .0078, and .0055. These sum to $\bar{\sigma}^2(\pi)_{\cdot}$, and dividing them by $\bar{\sigma}^2(\pi)_{\cdot}$ gives estimates of proportional contributions to the composite universe score variance, which can be compared with the w_c. The $\overline{\mathrm{COV}}_c/\bar{\sigma}^2(\pi)_{\cdot}$ are .11, .34, .20, .20, .14, which are very close to the w_c: .10, .35, .20, .20, .15. Thus, in this case the w_c reflect, not only the a priori relative importance of categories, but also (approximately) the contributions that categories make to composite universe score variance. Such a correspondence is of value simply because it indicates that no category contributes too much or too little to composite universe score variance as compared to its *defined* relative importance (w_c).

Minimizing error variance

The I_c that minimize $\sigma^2(\Delta)$ are estimated by substituting $\bar{\sigma}^2(\beta)_c$ and $\bar{\sigma}^2(\pi\beta)_c$ into Equation 21. These are presented in the first row of Table 3. Recall that Equation 21 gives the I_c on a continuous scale; call these the "optimal" I_c. The error variance based on the optimal I_c [$\bar{\sigma}^2(\Delta)_{opt}$] is also provided in Table 3. Below the optimal I_c are an obvious discrete choice for the I_c. These are in fact identical to the I_c used for MATH, and $\bar{\sigma}^2(\Delta)$ is provided again for comparison. Note that there is almost no difference between $\bar{\sigma}^2(\Delta)$ and $\bar{\sigma}^2(\Delta)_{opt}$. This holds because the sums $\bar{\sigma}^2(\beta)_c + \bar{\sigma}^2(\pi\beta)_c$ are close in value. In general, if these sums are equal, the optimal I_c are equal to the $w_c\, I_+$. Furthermore, given equal sums and integer values for the $I_c = w_c\, I_+$, a simple average score taken across all items minimizes $\sigma^2(\Delta)$.

Table 3
Choice of the I_c for Minimization of $\sigma^2(\Delta)$

Choice of I_c	AAO	AAR	GEO	IA	NA	Error Variance
Optimal	3.82	14.17	8.02	7.96	6.03	$\bar{\sigma}^2(\Delta)_{opt} = .005171$
Actual	4	14	8	8	6	$\bar{\sigma}^2(\Delta) = .005172$

Decreasing error variance through adjustments for relative difficulty

The actual score reported for MATH is derived through an equating process. Assuming that a by-product of the process is a precise adjustment in scores for the relative difficulty of a particular test form, an error variance can be defined which considers such an adjustment. From the model and the definition of $\hat{\mu}_{p\cdot}$, the relative difficulty of a form is $\Sigma_c\, w_c\, \Sigma_i\, \beta_{i:c}/I_c$. Given a precise estimate of this term, an adjustment in $\hat{\mu}_{p\cdot}$ can be made:

$$\mathring{\mu}_{p\cdot} = \hat{\mu}_{p\cdot} - Est\left(\sum_c w_c \sum_i \beta_{i:c}/I_c\right) , \qquad [23]$$

where $\dot{\mu}_{p\cdot}$ is the adjusted estimate of $\mu_{p\cdot}$ and *Est* denotes an estimate.

When the error of estimating relative difficulty is small enough to ignore, error variance for $\dot{\mu}_{p\cdot}$ is

$$\sigma^2(\delta) = \mathcal{E}(\dot{\mu}_{p\circ} - \mu_{p\circ})^2 = \sum_c w_c^2\, \sigma^2(\pi\beta)_c/I_c \quad . \qquad [24]$$

For MATH $\bar{\sigma}^2(\delta) = .00461$, which is close to $\bar{\sigma}^2(\Delta)$, indicating that different forms of MATH are similar in difficulty.

Error due to deviations from the table of specifications

With such high correlations among universe scores in the different categories of MATH, it is reasonable to examine the importance of using the categorization for generating multiple forms. As suggested above, Equation 22 can be used to address this issue.

For example, suppose a MATH form were generated with all 40 items falling into the second category, AAR. This would be far from the worst case, since AAR items normally constitute 35% of the test. An estimate of mean-squared error for such test forms can be obtained by setting all the a_c and the I_c in Equation 22 to zero, except for those for AAR that are set to 1 and 40, respectively. Using the estimates in Table 2, the estimated mean-squared error for 40-item tests with only AAR items is .00729, which is 41% larger than $\bar{\sigma}^2(\Delta) = .00517$. Similarly, the estimated mean-squared error for tests with only AAO items is .02459; for GEO it is .00882; for IA it is .01004; and for NA it is .00598. Clearly, it is important to have at least some balance of items across the categories.

In contrast, a pattern of items that more closely resembles the actual I_c produces a mean-squared error quite close to $\bar{\sigma}^2(\Delta)$. For example, taking all the $I_c = 8$ and using a proportion-correct score to estimate $\mu_{p\cdot}$ (i.e., all $a_c = .2$), the resulting estimate of mean-squared error is .00540. Thus, strict adherence to the actual I_c does not seem too critical for MATH.

Signal-noise ratios and reliability-like coefficients

It has been suggested above that $\bar{\sigma}^2(\pi)./\bar{\sigma}^2(\Delta) = 7.3$ indicates the relative precision of the measurement procedure. As such, this statistic is an estimate of the signal-noise ratio for decisions about the absolute value of examinee scores (see Brennan & Kane, 1977). If desired, this statistic can be transformed to a reliability-like coefficient $\bar{\sigma}^2(\pi)./[\bar{\sigma}^2(\pi). + \bar{\sigma}^2(\Delta)] = .88$. Similarly, the signal-noise ratio $\bar{\sigma}^2(\pi)./\bar{\sigma}^2(\delta) = 8.2$ could be examined or transformed to $\bar{\sigma}^2(\pi)./[\bar{\sigma}^2(\pi). + \bar{\sigma}^2(\delta)] = .89$, which has the form of an estimated generalizability coefficient.

Conclusions

A set of a priori weights (the w_c) has been used to define a composite universe score ($\mu_{p\cdot}$). These weights reflect the (judged) relative importance of the categories, which is an intentionally vague concept. It depends on the context of the measurement procedure. However, the definition of the composite universe score must accommodate the intended universe to which generalizations are made. The modeling here stresses that this universe be defined a priori. By contrast, some discussions on weighting in multivariate generalizability theory focus on the estimation of weights that serve to maximize a generalizability coefficient (cf. Joe & Woodward, 1976). In essence, such an approach implies that the universe of generalization is defined a posteriori as the one giving the highest generalizability coefficient.

There certainly are alternatives to choosing a priori the w_c that will accommodate the intended universe of generalization. For example, it may be desired to have the defined relative importance of a category be directly reflected by the contribution the category makes to composite universe score variance. This contribution has been referred to as COV_c. It is possible to choose a priori the COV_c and then to determine the w_c from them. An iterative solution to this problem, assuming Σ known, is provided by Dunnette and Hoggatt (1957) and was originally introduced by Wilks (1938).

References

American College Testing Program. *Content of the tests in the ACT Asessment.* Iowa City IA: Author, 1980.

Brennan, R. L. *Elements of generalizability theory.* Iowa City IA: The American College Testing Program, in press.

Brennan, R. L., Jarjoura, D., & Deaton, E. *Some issues concerning the estimation and interpretation of variance components in generalizability theory* (ACT Technical Bulletin No. 36). Iowa City IA: The American College Testing Program, 1980.

Brennan, R. L., & Kane, M. T. Signal/noise ratios for domain-referenced tests. *Psychometrika,* 1977, *42,* 609-625.

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

Dunnette, M. D., & Hoggatt, A. C. Deriving a composite score from several measures of the same attribute. *Educational and Psychological Measurement,* 1957, *17,* 423-434.

Jarjoura, D., & Brennan, R. L. *Three variance component models for some measurement procedures in which unequal numbers of items fall into discrete categories* (ACT Technical Bulletin No. 37). Iowa City IA: The American College Testing Program, 1981.

Joe, G. W., & Woodward, J. A. Some developments in multivariate generalizability. *Psychometrika,* 1976, *41,* 205-219.

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

Scheffé, H. *The analysis of variance.* New York: Wiley, 1959.

Searle, S. R. *Linear models.* New York: Wiley, 1971.

Wang, M. W., & Stanley, J. C. Differential weighting: A review of methods and empirical studies. *Review of Educational Research,* 1970, *4,* 663-705.

Wilks, S. S. Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika,* 1938, *3,* 23-40.

Acknowledgments

The authors thank Michael T. Kane and D. R. Divgi for their helpful comments concerning a prior version of this paper. Parts of this paper provide a summary of results discussed more fully in Jarjoura and Brennan (1981).

Authors' Address

Send requests for reprints or further information to David Jarjoura or Robert L. Brennan, Measurement Research Department, American College Testing Program, P.O. Box 168, Iowa City IA 52243, U.S.A.