Handbook of Statistics, Vol. 26
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(06)26004-8

Reliability Coefficients and Generalizability Theory

Noreen M. Webb, Richard J. Shavelson and Edward H. Haertel

1. Introduction

When a person is tested or observed multiple times, such as a student tested for mathematics achievement or a Navy machinist mate observed while operating engine room equipment, scores reflecting his or her performance may or may not agree. Not only may individuals' scores vary from one testing to another, calling into question the defensibility of using only one score for decision-making purposes, but the rankings of individuals may also disagree. The concern of reliability studies is to estimate the consistency of scores across repeated observations. Reliability coefficients quantify the consistency among the multiple measurements on a scale from 0 to 1.

In this chapter we present reliability coefficients as developed in the framework of classical test theory, and describe how the conception and estimation of reliability was broadened in generalizability theory. Section 2 briefly sketches foundations of classical test theory (see the chapter by Lewis for a thorough development of the theory) and focuses on traditional methods of estimating reliability. Section 3 reviews generalizability theory, including applications and recent theoretical contributions.

2. Reliability Coefficients in Classical Test Theory

Classical test theory's reliability coefficients are widely used in behavioral and social research. Each provides an index of measurement consistency ranging from 0 to 1.00 and their interpretation, at first blush, is relatively straightforward: the proportion of observed-score variance attributable to true-scores (stable or nonrandom individual differences) (see Lewis chapter for definitions in Classical Test Theory). Coefficients at or above 0.80 are often considered sufficiently reliable to make decisions about individuals based on their observed scores, although a higher value, perhaps 0.90, is preferred if the decisions have significant consequences. Of course, reliability is never the sole consideration in decisions about the appropriateness of test uses or interpretations.

Coefficient alpha (also known as "Cronbach's alpha") is perhaps the most widely used reliability coefficient. It estimates test-score reliability from a single test administration using information from the relationship among test items. That is, it provides an

estimate of reliability based on the covariation among items internal to the test; hence it is also called an internal-consistency coefficient. Cronbach's (2004) (for context, see Shavelson, 2004) reflections on 50 years of coefficient alpha's use foreshadows our treatment of reliability coefficients in this chapter – as a useful but limited tool for practice. More importantly for us, reliability coefficients give insight into limitations of classical test theory and provide background for our treatment of classical test theory's successor, generalizability theory:

> *I no longer regard the alpha formula as the most appropriate way to examine most data. Over the years, my associates and I developed the complex generalizability (G) theory (*Cronbach et al., 1963*;* Cronbach et al., 1972*; see also* Brennan, 2001*;* Shavelson and Webb, 1991*) which can be simplified to deal specifically with a simple two way matrix and produce coefficient alpha* (Cronbach, 2004, p. 403).

In this section, we develop classical test theory's notion of reliability and ways in which it is estimated and applied (see Lewis' chapter for details of this theory). We then show that each traditional estimation method – test-retest, parallel (equivalent) forms, and internal consistency – defines reliability somewhat differently; none is isomorphic with the theoretical definition. We conclude this section by presenting "popular" reliability coefficients and their application before turning to generalizability theory.

2.1. Theoretical definition of reliability

Classical test theory (CTT) is often introduced by assuming an infinite population of people, each observed an infinite number of times (see Table 1). The cells of this "infinite matrix" (Cronbach, 2004) contain observed scores – X_{pi} is the observed total score, for example, the total number of mathematics items correct for person p observed under condition i (e.g., person p's total score on occasion i). The average of a person's scores in the infinite matrix (i.e., the average over i of the X_{pi} for person p) is defined as that person's true score – T_p. Note that we denote the true score without specifying the particular condition because of CTT's steady-state assumption – for the period of observation, the person is in a steady state, that is, the person's true score is constant (later on we relax this definition). Because a person's observed scores vary from one condition to the next even though the true score is constant, this variation must be due, in theory, to random measurement error: $e_{pi} = X_{pi} - T_p$. Consequently, a person's observed score is comprised of a true score and error: $X_{pi} = T_p + e_{pi}$. Moreover, the variance of persons' observed scores for condition i equals true-score variance plus error variance, as the covariance (or correlation) of true scores and errors is zero (as is the covariance of errors on two parallel tests): $s^2_{X_i} = s^2_T + s^2_{e_i}$.

We now have sufficient information from CTT to define the reliability coefficient. The reasoning goes like this. If there is little variability in observed scores from one condition to the next, each observed score must be pretty close to the person's true score. Hence, assuming persons vary in their true scores, the correlation between observed and true scores should be quite high. If, however, there is a large bounce in a person's observed score from one condition to the next, the correspondence between observed and true scores must be minimal, error must be substantial, and as a consequence the correlation between observed and true scores must be low.

Table 1
Person × condition score (X_{pi}) "infinite" ("population-universe") matrix[a]

Person	Condition					
	1	2	...	i	...	$k \to \infty$
1	X_{11}	X_{12}	...	X_{1i}	...	X_{1k}
2	X_{21}	X_{22}	...	X_{2i}	...	X_{2k}
...	...	...	...	...	...	...
p	X_{p1}	X_{p2}	...	X_{pi}	...	X_{pk}
...	...	...	...	...	...	...
$n \to \infty$	X_{n1}	X_{n2}	...	X_{ni}	...	X_{nk}

[a]Adapted from Cronbach (2004, p. 401).

Reliability in CTT – the proportion of observed-score variance due to variance among persons' true scores – is defined as the square of the correlation, in the population, between observed and true scores, ρ^2_{XT}:

$$\rho_{XT} = \frac{\sigma_{XT}}{\sigma_X \sigma_T} = \frac{\sigma_{TT}}{\sigma_X \sigma_T} = \frac{\sigma^2_T}{\sigma_X \sigma_T} = \frac{\sigma_T}{\sigma_X} \quad \text{Reliability Index,} \tag{1.1a}$$

$$\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X} \quad \text{Reliability Coefficient.} \tag{1.2b}$$

2.2. *Estimation of the reliability coefficient*

The theoretical reliability coefficient is not practical; we do not know each person's true score. Nevertheless, we can estimate the theoretical coefficient with the sample correlation between scores on two parallel tests. Assume that X and X' are two strictly parallel tests (for simplicity) – that is, tests with equal means, variances, covariances with each other, and equal covariances with any outside measure (see chapter by Lewis). The Pearson product-moment correlation between parallel tests produces an estimate of the theoretical reliability coefficient:

$$r_{XX'} = \frac{s_{XX'}}{s_X s_{X'}} = \frac{s_{(T+e)(T+e')}}{s^2_X} = \frac{s^2_T}{s^2_X}. \tag{2}$$

While we have assumed strictly parallel tests for simplicity in deriving 2, we need not make such strong assumptions. Indeed, reliability coefficients can be obtained with increasingly lenient assumptions (see Lewis' Chapter 2 for details):

- Tau equivalent tests – each person is assumed to have a constant true score over tests but the error variances may vary across tests.
- Essentially tau equivalent tests – each person's true score under one condition (g) differs by an additive constant under a second condition (h): $T_{pg} = T_{ph} + c_{gh}$. (Note: while the true scores differ by c_{gh} across conditions, true-score variance is constant.) Error variances may differ.

Table 2a
Schematic sample person × condition score (X_{pi}) matrix

Person	Condition 1	Condition 2	...	Condition i	...	Condition k
1	X_{11}	X_{12}	...	X_{1i}	...	X_{1k}
2	X_{21}	X_{22}	...	X_{2i}	...	X_{2k}
...	...	...	...	...	...	...
p	X_{p1}	X_{p2}	...	X_{pi}	...	X_{pk}
...	...	...	...	...	...	...
N	X_{N1}	X_{N2}	...	X_{Ni}	...	X_{Nk}

Table 2b
Person × condition score matrix with hypothetical data

Person	Condition 1	Condition 2
1	2	1
2	2	3
3	2	3
4	4	3
5	6	5
6	6	5
7	8	6
8	8	6
9	9	8
10	8	9
Mean	5.500	4.900
Variance	7.833	6.100

- Congeneric tests – each person's true score may vary across tests but there is an additive constant (c_{gh}) and a positive regression constant (b_{gh}) that relates true-scores across any two tests: $T_{ph} = b_{gh}T_{pg} + c_{gh}$. Neither true-score nor error variances need be equal.

2.3. Reliability study designs and corresponding reliability coefficients

To estimate test-score reliability, at a minimum one needs at least two observations (scores) on the same set of persons (Tables 2a and 2b). The correlation between one set of observations with the second, then, provides a reliability coefficient. (Internal consistency reliability estimates follow a slightly more complicated procedure.) From this simple requirement, a wide variety of reliability studies could be designed.

2.3.1. Test–retest reliability coefficient

In designing a reliability study to produce two sets of observations, one might give the same test on two occasions, separated (say) by two weeks. In this case, Condition 1 in Table 2b becomes Occasion 1, and Condition 2 becomes Occasion 2. Following

Cronbach et al. (1972) and Thorndike (1947), notice that whatever is lasting and general about the test taker (e.g., ability, testwiseness) will contribute to the consistency of the measurement from one occasion to another – i.e., contribute to "wanted" true-score variance and reliability. Something that is lasting but specific about a test taker – e.g., knowledge pertaining to a particular item or set of items on the test – will also contribute to consistency in scores. However, something temporary but general (e.g., disruptions by a lawnmower on a particular test occasion) will contribute to inconsistency from one occasion to another. Something temporary and specific, such as a very recent event (e.g., recent car engine repair creating a negative mental set on a science-test item on ideal gas law illustrated with a piston) might affect performance at time 1 but not at time 2 and hence also give rise to inconsistency in performance. Finally, chance success (or failure!) such as random guessing will also and always contribute to inconsistency from one time to the next. In summary, consistency or reliability with this reliability study design involves lasting general and lasting specific conditions of the test; temporary conditions or events (general and specific) and chance events contribute to error.

The correlation between scores at occasion 1 and occasion 2 produces a test-retest reliability coefficient – a coefficient of stability. From Table 2b, we find the correlation between scores at time 1 and time 2 – i.e., the reliability of the test scores – is 0.908.

2.3.2. Parallel or equivalent forms reliability

A second way to design a reliability study is to create two parallel forms of the test, say Form 1 and Form 2, and give the two forms of the test on the same day. In this case, the correlation between scores on Forms 1 and 2 yields the test's (either form's) reliability, also referred to as a coefficient of equivalence. From Table 2b, considering Condition 1 as Form 1 and Condition 2 as Form 2, the reliability of either form is 0.908.

The parallel forms reliability study defines consistency or reliability somewhat differently than does the retest study. For both retest and parallel forms estimates (a) the lasting-general effect contributes to measurement consistency – i.e., consistent individual differences; (b) the temporary-specific effect – e.g., the car engine example above – contributes to inconsistency across forms; and of course (c) the other random influences contribute to inconsistency. In contrast to retest reliability: (a) the lasting-specific effect – e.g., knowing particular items on one form of the test – will *not* contribute to consistent performance across test forms; and (b) the temporary-general ("lawnmower") effect may contribute to test-score consistency when the two forms of the test are taken at the same time.

2.3.3. Delayed parallel or equivalent forms reliability

It is possible to design a reliability study in which two parallel forms of a test are developed and administered to a sample of people on two different occasions. This produces a design in which the lasting-general effect contributes to consistency but the other effects contribute to error. The correlation between scores on the two forms taken at different times produces a reliability coefficient that we would expect to be less than or

equal to either the stability or equivalent forms reliability coefficients above. This may be referred to as a coefficient of stability and equivalence.[1]

2.3.4. Internal-consistency reliability

Both retest and parallel-forms reliability are logistically inconvenient. Either the same sample of people must be tested on two occasions or more than one form of a test needs to be developed to find the reliability of a single form. The solution to how to calculate reliability of a single form of a test has a long and illustrious history (e.g., Cronbach, 2004). All solutions involve recognizing that a full length test, given on a single occasion, can be divided into (parallel) parts.

Split-half reliability

Initially tests were divided into half parts and a host of split-half reliability coefficients were derived (for reviews and details, see, for example, Feldt and Brennan, 1989; Haertel, in press). The oldest and probably most widely used split-half coefficient calculates the correlation between scores on two halves of the test (X_1 and X_2) and estimates the correlation – reliability – for a full length test with the Spearman–Brown "prophecy" formula, assuming strictly parallel test halves:

$$SB_2 r_{XX'} = \frac{2r_{X_1X_2}}{1 + r_{X_1X_2}}. \tag{3}$$

If we treat Conditions 1 and 2 in Table 2b as corresponding to persons' total scores on two halves of the same test, we can find the reliability of the full length test – i.e., double the length of the half tests – with Spearman–Brown:

$$SB_2 r_{XX'} == \frac{2 \cdot 0.908}{1 + 0.908} = 0.952.$$

While all this seems straightforward, it is not and herein lies the controversy. One source of the controversy lies in how to define halves of a test – the first versus the last half? Odd–even halves? Random halves? While the debate goes on, consensus is that the first versus last half method is not a good idea, especially if fatigue and speed apply toward the end of the test, and that content should be stratified in the allocation of items to halves. The debate, in a way, is also over how to define true-score and error variance. Internal consistency coefficients such as the split half method, like parallel-forms coefficients, treat lasting-general (e.g., ability) and temporary-general ("lawnmower") effects as consistency (reliability) in the measurement for most such splits, and lasting- and temporary-specific effects along with random guessing as error.

The second source of controversy revolves around the strictly parallel assumption of the Spearman–Brown formula. To this end, a whole host of split-half reliability coefficients have been developed historically that relax the assumptions (e.g., Flanagan's formula, which requires only the essential tau equivalence assumption). Since, in practice, these coefficients are not as widely used as is a general method, coefficient alpha, we do not provide these formulas (see Feldt and Brennan, 1989; Haertel, in press).

[1] For both the equivalent forms and the delayed equivalent forms designs, better estimates of the reliability of forms g and h are given by $\hat{\sigma}_{gh}/\hat{\sigma}_g^2$ and $\hat{\sigma}_{gh}/\hat{\sigma}_h^2$, respectively.

Multi-part reliability and coefficient alpha

Up to this point, we have considered tests split in half to produce two strictly parallel forms to be used to estimate reliability. Scores on one half can be correlated with scores on the other half. If the test is longer than two items, and most tests are, it is possible to split it into thirds, fourths, fifths, and so on until each test item is considered a parallel test itself. This is just the reasoning used in creating general-use internal consistency reliability coefficients, Cronbach's alpha being the most widely known and used.

To see the reasoning behind alpha, we extend the Spearman–Brown application from split-half reliability to multi-part reliability. Rather then adjusting the correlation between scores on test halves, we use the correlation between individual test items – the reliability of a single item test – and adjust this correlation with a generalized Spearman–Brown formula. The Spearman–Brown formula then provides the reliability – internal consistency – of the full-length test with all k items. In making this adjustment using the Spearman–Brown formula, we must assume equal variances and correlations between items. The general Spearman–Brown formula for a test of k items is:

$$_{SB_k}r_{XX'} = \frac{kr_1}{1 + (k-1)r_1}, \tag{3a}$$

where SB_k refers to the Spearman–Brown adjustment for the reliability of a k-item test, k refers to the number of items, and r_1 refers to the inter-item correlation or the reliability of a single-item test.

To see how (3a) works, let's continue the example from Table 2b where Condition 1 refers to a 12-point Likert-type rating on questionnaire item 1 and Condition 2 refers to a similar rating on a strictly parallel questionnaire item 2. In this case, we know that $r_1 = 0.908$ is the reliability of a single item questionnaire. So an estimate of the reliability of the two-item questionnaire would be:

$$_{SB_{K=2}}r_{XX'} = \frac{2 \cdot 0.908}{1 + (2-1) \cdot 0.908} = 0.952.$$

We can also use (3a) to prophesize what the reliability of the test would be with 5 items (or any other number of items):

$$_{SB_{K=5}}r_{XX'} = \frac{5 \cdot 0.908}{1 + (5-1) \cdot 0.908} = 0.980.$$

In general, as parallel items are added to a test, the reliability of the test increases in the manner given by (3a). Conceptually it is helpful to conceive of internal consistency coefficient alpha as the inter-item correlation on a test adjusted to the full length of the test.

In practice, however, the assumption of strictly parallel items is too restrictive. Cronbach (1951) derived an internal consistency coefficient, coefficient alpha, based on a weaker assumption, that of essential tau equivalence; violation of the assumption tends to lower the coefficient.

The general formula for coefficient alpha is typically written as (Cronbach, 2004):

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{s_t^2}\right), \tag{4}$$

where k refers to the number of items on the test, s_i^2 refers to the variance of item i, and s_t^2 refers to the variance of total scores (summed over items) on the test. Again, referring to the data in Table 2b and considering Conditions 1 and 2 to be two questionnaire items, we can calculate the item and total score variance: $s_1^2 = 7.833$, $s_2^2 = 6.100$, and $s_t^2 = 26.489$. Using this information and $k = 2$ in (4) we have:

$$\alpha = \frac{2}{2-1}\left(1 - \frac{7.833 + 6.100}{26.489}\right) = 2\left(1 - \frac{13.933}{26.489}\right) = 0.948.$$

Note that the alpha coefficient gives a lower result than the Spearman–Brown formula. If the sample variances for the two conditions (Table 2b) were identical, the results from the two formulas would be the same.[2]

Coefficient alpha provides a general solution to most all split-half methods that relax the strictly-parallel test assumptions. Moreover, whereas earlier internal consistency formulas such as Kuder–Richardson formulae 20 and 21 dealt with the special case of dichotomously scored items (see Feldt and Brennan, 1989, and Haertel, in press, for details), Cronbach's alpha applies to the more general case of items scored dichotomously or otherwise (e.g., Likert-type scale as in the example above).

Components of variance approach to coefficient alpha

The components of variance approach to alpha begins with the infinite matrix (Table 1) and assumes that persons (rows) and conditions (columns) are sampled from this matrix at random – i.e., the variance-component approach assumes randomly parallel conditions. "The ... assumptions ... are not true in any strict sense, and a naive response would be to say that if the assumptions are violated, the alpha calculations cannot be used. No statistical work would be possible, however, without making assumptions and, as long as the assumptions are not obviously grossly inappropriate to the data, the statistics calculated are used, if only because they can provide a definite result that replaces a hand-waving interpretation" (Cronbach, 2004, p. 404).

In the random model with persons crossed with conditions (for example, persons crossed with items), the observed score for person p on a single item i, X_{pi}, can be divided into four components:

$$X_{pi} = \mu + (\mu_p - \mu) + (\mu_i - \mu) + (X_{pi} - \mu_p - \mu_i + \mu). \tag{5}$$

The first component on the right is the grand mean, μ, constant for all persons. The next term is person p's true score expressed in deviation form $(\mu_p - \mu)$. The metric for the person's true score is the item, where μ_p denotes person p's mean item score over the items in the infinite matrix. (This contrasts with T_p in earlier sections, which was expressed in the total score metric.) The third component is the item effect $(\mu_i - \mu)$ – the average deviation score for item i. And the last term, the residual $(X_{pi} - \mu_p - \mu_i + \mu)$, reflects the departure of observed scores from what would be expected given the first three terms in (5).

[2] If item scores are standardized, the result is what some statistical computer programs call "*standardized alpha*."

Table 3
Expected mean squares and variance component estimates

Expected mean square	Variance component estimators
$E(MS_p) = \sigma^2_{RES} + n_i\sigma^2_p$	$\hat{\sigma}^2_p = \frac{MS_p - \hat{\sigma}^2_{RES}}{n_i}$
$E(MS_i) = \sigma^2_{RES} + n_p\sigma^2_i$	$\hat{\sigma}^2_i = \frac{MS_i - \hat{\sigma}^2_{RES}}{n_p}$
$E(MS_{RES}) = \sigma^2_{RES}$	$\hat{\sigma}^2_{RES} = MS_{RES}$

Except for the first component, μ, each of the components of the observed score varies – from one person to another, from one item to another, or from one person-item combination to another. For this simple design, it is readily shown that these last three score components, as defined, must be uncorrelated. Hence, the observed-score variance can be decomposed into three *components of variance* (or simply, variance components):

$$\sigma^2_{X_{pi}} = \sigma^2_p + \sigma^2_i + \sigma^2_{RES}. \tag{6}$$

The first variance component on the right (σ^2_p) is the systematic, error-free variance among persons' score components. The i variance component (σ^2_i) describes the extent to which the item means vary. And the residual variance component (σ^2_{RES}) is what is commonly thought of as variance due to measurement error. (We speak of a residual because in practice, from sample data as in Table 2a, we cannot disentangle the $p \times i$ interaction from random error e.)

Notice that (6) departs from the usual decomposition of observed score variance into true-score variance plus error variance. If we assume strictly parallel tests with equal means, the variance due to conditions (σ^2_i) is zero and drops out. As Cronbach (2004, p. 406) explains, "Spearman started the tradition of ignoring item characteristics because he felt that the person's position on the absolute score scale was of no interest"; condition means were arbitrary. More generally, if our interest is in the performance of a group of examinees relative to one another, and if all of them take the same form of the test, then the score component associated with item difficulties is irrelevant to our intended inferences and should be ignored.

The variance components in (6) can be estimated from sample data (e.g., Tables 2a, 2b) using the random-effects model of the analysis of variance. Variance component estimates can be calculated by setting the observed mean squares in the ANOVA equal to their expectations and solving for the variance components (Table 3). In Table 3, in preparation for the factorial ANOVA approach used in generalizability theory (next section), we use n_i instead of k to represent the number of items and n_p instead of N to represent the number of persons.

With the expected mean squares and estimated variance components in hand, we can write coefficient alpha – proportion of observed-score variance (assuming that $\sigma^2_i = 0$)

due to true-score variance – as:

$$\alpha = \frac{MS_P - MS_{RES}}{MS_P} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{RES}^2}{n_i}}. \tag{7}$$

As an example, consider the data in Table 2b. Estimating the variance with a repeated-measures ANOVA yields: $MS_{RES} = 0.689$, $MS_i = 1.800$, and $MS_p = 13.244$. Solving for the estimated variance components, we have:

$$\hat{\sigma}_{RES}^2 = MS_{RES} = 0.689,$$
$$\hat{\sigma}_i^2 = (MS_i - MS_{RES})/N = (1.800 - 0.689)/10 = 0.111,$$
$$\hat{\sigma}_p^2 = (MS_p - MS_{RES})/k = (13.244 - 0.689)/2 = 6.278.$$

Inserting the appropriate estimated variance components into (7) and assuming $n_i = 2$, we have:

$$\alpha = \frac{6.278}{6.278 + \frac{0.689}{2}} = 0.948.$$

2.3.5. Reliability coefficients for special occasions

Up to this point, we have dealt with widely used, standard reliability formulas that apply to total test scores or item scores. Nevertheless, we have not dealt directly with scores on open-ended ("constructed response") tests where "raters" or examiners provide the scores. Moreover, scores other than observed scores, such as difference scores or various kinds of composite scores, are widely used in social, behavioral and educational research. Reliability coefficients have been adapted or developed for such "special" occasions. We review, briefly, a number of "special-occasion" reliability formulas here.

Interrater reliability

Often with essay and other constructed-response tests, two (or more) raters score individuals' performance. Each rater would provide a performance score for each individual (e.g., a total score summed over open-ended test items). Assuming "parallel raters" are created by selection and training, akin to parallel tests, the correlation between the scores assigned by Rater 1 (Condition 1 in Table 2b) and by Rater 2 (Condition 2) to persons can be used to determine interrater reliability – the consistency among raters (in the sense of relative or deviation scores assigned). From the data in Table 2b, the reliability of scores assigned by a single rater (cf. single form) would be 0.908.

Reliability of difference scores

Often interest attaches to an individual's change over time. One way to reflect change is to calculate a difference score – post-test–pretest: $X_{post} - X_{pre}$. In this case:

$$X_{post} = T_{post} + e_{post},$$
$$X_{pre} = T_{pre} + e_{pre}, \quad \text{so that } X_D = (T_{post} - T_{pre}) + (e_{post} - e_{pre}).$$

Assuming that the covariance of the errors on the pre- and post-test is zero, we can write the true-score variance and error variance for the difference score as follows:

$$\sigma^2_{T_D} = \sigma^2_{T_{pre}} + \sigma^2_{T_{post}} - 2\sigma_{T_{pre}T_{post}},$$
$$\sigma^2_{e_D} = \sigma^2_{e_{pre}} + \sigma^2_{e_{post}}.$$

Note that usually the covariance between pre- and post-test scores will be positive, reflecting the positive covariance between true scores on the pre- and post-tests. This implies that the true-score variance of difference scores will be less than the sum of the true-score variances for the pre- and post-tests. However, the error variance of the difference score will be equal to the sum of the pre- and post-test error variances. Consequently, often difference scores have lower reliability than observed scores.

Defining reliability as true (difference)-score variance divided by observed (difference)-score variance, a simple, general computational formula is (Haertel, in press):

$$\rho_{DD'} = 1 - \frac{(1 - \rho_{X_{pre}X'_{pre}})\sigma^2_{X_{pre}} + (1 - \rho_{X_{post}X'_{post}})\sigma^2_{X_{post}}}{\sigma^2_{(X_{post}-X_{pre})}}. \tag{8}$$

If $\sigma^2_{X_{pre}} = \sigma^2_{X_{post}}$, we can write (8) as:

$$\rho_{DD'} = \frac{\bar{\rho}_{XX'} - \rho_{X_{pre}X_{post}}}{1 - \rho_{X_{pre}X_{post}}}, \tag{8a}$$

where $\bar{\rho}_{XX'}$ is the average of the reliabilities of the pre- and post-test scores.

Difference scores have been criticized not only for low reliability but for being problematic in the sense that individuals with high pretest scores are likely to improve less than individuals with low pretest scores (i.e., difference scores are often negatively correlated with pretest scores). Nevertheless, difference scores have been viewed favorably as they provide an unbiased estimator of the true difference for a randomly chosen individual. Moreover, low difference-score reliability does not necessarily mean low statistical power for mean comparisons (see Haertel, in press, for details). Because differences are expressed on a scale with a true zero point (representing no change), the absolute magnitude of the difference is usually of greater interest than the stability of a rank ordering by difference scores. For this reason, the reliability coefficient for the difference score is often the wrong statistic to use; the standard error of the difference score is more informative (see Haertel, in press, for details).

Reliability of composite scores: Weighted and unweighted

The reliability of difference scores is the simplest, most straightforward case of the reliability of composite scores where the difference is a linear combination of two component scores with fixed weights -1 and $+1$: $-1X_{pre} + 1X_{post}$. At the other extreme, a composite score might be formed as a weighted linear combination of a set of tests (components) measuring distinct constructs (e.g., subtests of the Wechsler intelligence test). The weights might be statistically (e.g., multiple regression) determined or defined a priori by expert judges or from theory.

(i) Weighted composites

The composite score, C_p, for person p formed with k component scores, $X_{p1}, \ldots, X_{pk}$ and weights $w_1, \ldots, w_k$ may be written as follows (an additive constant w_0 may also be present):

$$C_p = w_0 + \sum_{i=1}^{k} w_i X_{pi}. \tag{9}$$

To develop a reliability coefficient, we assume that composite scores consist of a true-score component and an error component with errors uncorrelated. Moreover, we conceive of parallel composite scores, C and C' formed from parallel component scores $X_{p1}, \ldots, X_{p2}, \ldots, X_k$ and $X'_{p1}, \ldots, X'_{p2}, \ldots, X'_k$, respectively. With these assumptions, the error variance for the composite score is simply a weighted sum of the error variances of the component scores; the true-score variance, however, is a bit more complicated as it is a weighted sum of the true-component-score variances and covariances. This said, the reliability of composite scores may be written (Haertel, in press):

$$\begin{aligned}\rho_{CC'} &= 1 - \frac{\sigma^2_{E_C}}{\sigma^2_C} \\ &= 1 - \frac{\sum_{i=1}^{k} w_i^2 \sigma^2_{E_i}}{\sigma^2_C} = 1 - \frac{\sum_{i=1}^{k} w_i^2 \sigma^2_{X_i}(1 - \rho_{X_i X'_i})}{\sigma^2_C}.\end{aligned} \tag{10}$$

(ii) Unweighted composites or test battery composite scores

Often a test battery composite will consist of an unweighted composite of subtest mean scores – i.e., simply add up a person's scores on a set of subtests. The reliability of this unweighted composite – often referred to as the reliability of a battery composite score – is a special case of (9) and (10) with all the w_i equal. If the components of a composite are each weighted 1, the formula for weighted composites can be simplified to that for a battery composite score:

$$\rho_{CC'} = 1 - \frac{\sum_{i=1}^{k} \sigma^2_{X_i}(1 - \rho_{X_i X'_i})}{\sigma^2_C} \tag{11}$$

If the components or subtests in a test battery composite score are scaled to have identical variances, we can simplify (11) to:

$$\rho_{CC'} = \frac{\left(\frac{1}{k-1}\right)\bar{\rho}_{X_i X'_i} + \bar{\rho}_{X_i X_j}}{\left(\frac{1}{k-1}\right) + \bar{\rho}_{X_i X_j}} \quad (i \neq j). \tag{11a}$$

From this formula we see that the battery composite reliability is a function of the number of subtests, the individual subtest reliabilities and the intercorrelation among subtests.

Stratified coefficient alpha

Tests may contain different types of items that can be categorized. For example, reading achievement tests may include both vocabulary and reading comprehension items.

Or, a battery of tests might be categorized into subsets as is done with the Wechsler intelligence test with both verbal (e.g., information, comprehension, arithmetic) and performance (e.g., block design, picture arrangement, matrix reasoning) categories. When components (items or subtests) fall within categories or strata, we might view the composite as a result of stratified random sampling of subtests or items.

When we have such stratification, we would expect that items or subtests within a stratum would correlate more highly with each other than with items or subtests in other strata. Consider a case with k strata, $i = 1, \ldots, k$, with stratum-level total scores X_i and reliabilities α_i. Let X represent the total score obtained as a weighted sum of the X_i. A reliability estimate that takes account this stratification is stratified coefficient alpha:

$$_{stratified}\alpha = 1 - \frac{\sum_{i=1}^{k} \sigma_{X_i}^2 (1 - \alpha_i)}{\sigma_X^2}. \tag{12}$$

When the stratification holds up and correlations within strata are higher than those between strata, coefficient alpha will be smaller than stratified alpha.

2.4. Concluding comments: The need for more comprehensive reliability theory

What became evident over 60 years ago, and what we have tried to make clear, is that there are many possible reliability coefficients for any single test. Moreover, different reliability coefficients define true-score and error differently. And these coefficients sometimes produce contradictory estimates of reliability. As Cronbach et al. (1972, p. 6) citing Goodenough (1936) pointed out:

> *[T]he investigator who compares two administrations of the same list of spelling words asks a different question than the investigator who compares performance on two different lists. Inconsistency of observers, inconsistency in the subject's response to different stimulus lists, inconsistency of his response to the same stimulus on different occasions – all of these may be sources of error, but any one comparison will detect some inconsistencies and not others.*

In the next part of this chapter, we present a theory that deals with definitional and interpretative issues of true scores, sources of error, and variation in reliability estimates. This theory is Generalizability theory; it draws on the components of variance approach to reliability.

3. Generalizability theory

As the previous section makes clear, classical test theory treats measurement error as undifferentiated random variation. Cronbach et al. (1972) (see also Brennan, 1997, 2001; Shavelson and Webb, 1981, 1991) introduced the theory of generalizability to make it possible to assess *multiple* sources of measurement error in order to characterize the measurement and improve its design (Brennan, 1997, 2001; Cronbach et al., 1972; Shavelson and Webb, 1981, 1991). Generalizability theory pinpoints the sources of measurement error, disentangles them, and estimates each one.

In G theory, a behavioral measurement (e.g., a test score) is conceived of as a sample from a *universe of admissible observations*, which consists of all possible observations on an *object of measurement* (typically a person) that a decision maker considers to be acceptable substitutes for the observation in hand. Each characteristic of the measurement situation (e.g., test form, test item, rater, test occasion) is called a *facet*. The universe of admissible observations is usually defined by the Cartesian product of the levels (called *conditions* in G theory) of the facets.

In order to evaluate the dependability of behavioral measurements, a *generalizability (G) study* is designed to isolate and estimate variation due to the object of measurement and as many facets of measurement error as it is reasonably and economically feasible to examine. A *decision (D) study* uses the information provided by the G-study to design the best possible application of the measurement for a particular purpose. In planning the D-study, the decision maker defines a *universe of generalization*, the set of facets and their levels to which he or she wants to generalize, and specifies the proposed interpretation of the measurement. The decision maker uses the information from the G-study to evaluate the effectiveness of alternative designs for minimizing error and maximizing reliability.

3.1. One-facet designs

Consider a one-facet universe of admissible observations with items (i). The decision maker is interested in the performance of persons drawn from a particular population. The *object of measurement*, then, is persons. Persons are not a source of error, hence not a facet. Assume all persons respond to all items so that persons (p) are crossed with items (i). We denote this crossed design as $p \times i$. With generalization over all admissible items taken from an indefinitely large universe, the decomposition of the observed score is the same as in (5):

$$\begin{aligned} X_{pi} &= \mu && \text{(grand mean)} \\ &\quad + \mu_p - \mu && \text{(person effect)} \\ &\quad + \mu_i - \mu && \text{(item effect)} \\ &\quad + X_{pi} - \mu_p - \mu_i + \mu && \text{(residual effect).} \end{aligned}$$

In G theory, μ_p is called the *universe score* instead of the true score, recognizing that a person may have different universe scores depending on the characterization of the universe.[3] The universe score, then, is defined as the expected value (E) of a person's observed score over the universe of items:

$$\mu_p \equiv \mathrm{E}_i X_{pi}. \tag{13}$$

The population mean for item i is:

$$\mu_i \equiv \mathrm{E}_p X_{pi}. \tag{14}$$

[3] Strictly speaking, the universe score is not defined until the objects of measurement are identified in the decision (D) study. A single G-study can give rise to different D studies with different objects of measurement, as noted later in this paper.

Table 4
Estimates of variance components for a random-effects, one-facet $p \times i$ G-study design

Source of variation	Mean square	Expected mean square	Estimated variance component
Person (p)	MS_p	$\sigma^2_{pi,e} + n_i\sigma^2_p$	$\hat{\sigma}^2_p = \frac{MS_p - MS_{pi,e}}{n_i}$
Item (i)	MS_i	$\sigma^2_{pi,e} + n_p\sigma^2_i$	$\hat{\sigma}^2_i = \frac{MS_i - MS_{pi,e}}{n_p}$
pi, e	$MS_{pi,e}$	$\sigma^2_{pi,e}$	$\hat{\sigma}^2_{pi,e} = MS_{pi,e}$

And the mean over both the population and the universe (the "grand mean") is:

$$\mu \equiv \mathrm{E}_p\mathrm{E}_i X_{pi}. \tag{15}$$

As defined in (6), each score component, except for the grand mean, has a distribution. The distribution of $\mu_p - \mu$ has mean zero and variance $\mathrm{E}_p(\mu_p - \mu)^2 = \sigma^2_p$, which is called the universe-score variance. Similarly, the component for item has mean zero and variance $\mathrm{E}_i(\mu_i - \mu)^2 = \sigma^2_i$. The residual component has mean zero and variance $\mathrm{E}_p\mathrm{E}_i(X_{pi} - \mu_p - \mu_i + \mu)^2 = \sigma^2_{pi,e}$, which indicates the person $\times$ item interaction (pi) confounded with unsystematic or unmeasured error (e). The collection of observed item-level scores, X_{pi}, has a variance $\mathrm{E}_p\mathrm{E}_i(X_{pi} - \mu)^2 = \sigma^2_{X_{pi}}$, which equals the sum of the variance components:

$$\sigma^2_{X_{pi}} = \sigma^2_p + \sigma^2_i + \sigma^2_{pi,e}. \tag{16}$$

Because the nature of the residual component changes depending on the design of the generalizability study, we use the notation pi, e instead of *RES*.

3.1.1. Generalizability and decision studies with a crossed design

The variance components are estimated from an analysis of variance of sample data in the *generalizability (G) study*. For a random-effects $p \times i$ (person by item) design in which a random sample of n_p persons responds to each of n_i randomly sampled items, numerical estimates of the variance components are obtained in the same way as in the components of variance approach to coefficient alpha, that is, by setting the expected mean squares equal to the observed mean squares and solving the set of simultaneous equations that appear in Table 4.

In the *decision (D) study*, decisions usually will be based on the mean over multiple observations rather than on a single observation. The mean score over a sample of n'_i items is denoted as X_{pI} in contrast to a score on a single item, X_{pi}. (Note the switch in notation from k in classical test theory to n'_i in generalizability theory.) A one-facet, crossed D-study design where decisions are to be made on the basis of X_{pI} is, then, denoted as $p \times I$. Note that the number of items used in the D-study, n'_i, may differ from the number of items used in the G-study, n_i. As is seen here, G theory typically uses the mean score metric (e.g., mean of multiple item scores) rather than the total score metric (e.g., sum of multiple item scores).

G theory recognizes that the decision maker might want to make two types of decisions based on a behavioral measurement: relative (norm-referenced) and absolute

(criterion- or domain-referenced). A *relative decision* concerns the relative ordering of individuals (e.g., norm-referenced interpretations of test scores). For relative decisions, the error in a random effects $p \times I$ design is defined as:

$$\delta_{pI} \equiv (X_{pI} - \mu_I) - (\mu_p - \mu), \tag{17}$$

where $\mu_p = \mathrm{E}_I X_{pI}$ and $\mu_I = \mathrm{E}_p X_{pI}$. The variance of the errors for relative decisions is:

$$\sigma_\delta^2 = \mathrm{E}_p \mathrm{E}_I \delta_{pI}^2 = \sigma_{pI,e}^2 = \frac{\sigma_{pi,e}^2}{n_i'}. \tag{18}$$

Notice that the main effect of item does not enter into error for relative decisions because all people respond to all items so any difference in items affects all persons and does not change their relative standing. In order to reduce σ_δ^2, n_i' may be increased (analogous to the Spearman–Brown prophesy formula in classical test theory and the standard error of the mean in sampling theory).

Although G theory stresses the importance of variance components and measurement error, it provides a *generalizability coefficient* that is analogous to the reliability coefficient in classical test theory (ratio of universe-score variance to the expected observed-score variance). For the $p \times I$ random-effects design, the generalizability coefficient is

$$\mathrm{E}\rho^2(X_{pI}, \mu_p) = \mathrm{E}\rho^2 = \frac{\mathrm{E}_p(\mu_p - \mu)^2}{\mathrm{E}_I \mathrm{E}_p (X_{pI} - \mu_I)^2} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2}. \tag{19}$$

Sample estimates of the parameters in (19) are used to estimate the generalizability coefficient:

$$\mathrm{E}\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_\delta^2}. \tag{20}$$

$\mathrm{E}\hat{\rho}^2$ is a biased but consistent estimator of $\mathrm{E}\rho^2$. When $n_i' = n_i$, $\mathrm{E}\hat{\rho}^2$ is the same as coefficient alpha in (7).

An *absolute decision* focuses on the absolute level of an individual's performance independent of others' performance (cf. criterion- or domain-referenced interpretations). For absolute decisions, the error in a random-effects $p \times I$ design is defined as:

$$\Delta_{pI} \equiv X_{pI} - \mu_p \tag{21}$$

and the variance of the errors is:

$$\sigma_\Delta^2 = \mathrm{E}_p \mathrm{E}_I \Delta_{pI}^2 = \sigma_I^2 + \sigma_{pI,e}^2 = \frac{\sigma_i^2}{n_i'} + \frac{\sigma_{pi,e}^2}{n_i'}. \tag{22}$$

Note that, with absolute decisions, the main effect of items – how difficult an item is – does influence absolute performance and so is included in the definition of measurement error. Also note that $\sigma_\Delta^2 \geqslant \sigma_\delta^2$.

G theory provides an *index of dependability* (Kane and Brennan, 1977) for absolute decisions:

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\Delta^2}. \tag{23}$$

For criterion-referenced decisions involving a fixed cutting score (λ), it is possible to define a loss function based on the squared distance from the cut score. In such applications, assuming that λ is a constant that is specified a priori, the error of measurement is:

$$\Delta_{pI} = (X_{pI} - \lambda) - (\mu_p - \lambda) = X_{pI} - \mu_p, \tag{24}$$

and an index of dependability may be defined as:

$$\Phi_\lambda = \frac{\mathrm{E}_p(\mu_p - \lambda)^2}{\mathrm{E}_I\mathrm{E}_p(X_{pI} - \lambda)^2} = \frac{\sigma_p^2 + (\mu - \lambda)^2}{\sigma_p^2 + (\mu - \lambda)^2 + \sigma_\Delta^2}. \tag{25}$$

An unbiased estimator of $(\mu-\lambda)^2$ is $(\bar{X}-\lambda)^2-\hat{\sigma}^2_{\bar{X}}$ where $\bar{X}$ is the observed grand mean over sampled persons and sampled items in the D-study and $\hat{\sigma}^2_{\bar{X}}$ is the error variance involved in using the observed grand mean $\bar{X}$ as an estimate of the grand mean over the population of persons and the universe of items (μ). For the $p \times I$ random-effects design, $\hat{\sigma}^2_{\bar{X}}$ is:

$$\hat{\sigma}^2_{\bar{X}} = \frac{\hat{\sigma}_p^2}{n'_p} + \frac{\hat{\sigma}_i^2}{n'_i} + \frac{\hat{\sigma}^2_{pi,e}}{n'_p n'_i}. \tag{26}$$

The estimate of Φ_λ is smallest when the cut score λ is equal to the observed grand mean $\bar{X}$; in that case, $\hat{\Phi}_\lambda = \hat{\Phi}$.

Table 4a gives the results of generalizability analyses of the data in Table 2b. For $n'_i = 2$, $\mathrm{E}\hat{\rho}^2 = 0.948$ is the same as coefficient alpha using (7). The error variances for absolute decisions ($\hat{\sigma}^2_\Delta$) and coefficients of dependability ($\hat{\Phi}$) provide information that has no counterpart in classical test theory, that is, consistency of persons' absolute performance independent of others' performance rather than consistency of persons' relative standing. In this example, because the main effect for items ($\hat{\sigma}_i^2 = 0.11111$) is small relative to the other estimated variance components, there is little difference between $\mathrm{E}\hat{\rho}^2$ and $\hat{\Phi}$. The results in Table 4a show diminishing returns in $\mathrm{E}\hat{\rho}^2$ and $\hat{\Phi}$ when the length of the test is increased. Estimated levels of generalizability and dependability for a 10-item test ($\mathrm{E}\hat{\rho}^2 = 0.987$ and $\hat{\Phi} = 0.989$) are not appreciably higher than for a 5-item test ($\mathrm{E}\hat{\rho}^2 = 0.979$ and $\hat{\Phi} = 0.975$).

3.1.2. Generalizability and decision studies with a nested design

In a nested one-facet G-study design (items nested within persons, denoted as $i : p$), each person responds to a different set of n_i items. A nested G-study design may be used because the investigator considers the universe of admissible observations to be naturally nested or, when the universe of admissible observations is crossed, for cost or

Table 4a
Generalizability analysis of the data in Table 2b (1-facet, random effects $p \times i$ design)

Source of variation	G-study $\hat{\sigma}^2$	Size of decision study		
		$n'_i = 2$	$n'_i = 5$	$n'_i = 10$
Estimated variance components				
Person (p)	6.27778	6.27778	6.27778	6.27778
Item (i)	0.11111	0.05556	0.02222	0.01111
pi, e	0.68889	0.34444	0.13778	0.06889
Error variances				
$\hat{\sigma}^2_\delta$		0.34444	0.13778	0.06889
$\hat{\sigma}^2_\Delta$		0.40000	0.16000	0.08000
Coefficients				
$\mathrm{E}\hat{\rho}^2$		0.948	0.979	0.989
$\hat{\Phi}$		0.940	0.975	0.987

Table 5
Estimates of variance components for a random-effects one-facet $i : p$ G-study design

Source of variation	Mean square	Expected mean square	Estimated variance component
Person (p)	MS_p	$\sigma^2_{i,pi,e} + n_i\sigma^2_p$	$\hat{\sigma}^2_p = \frac{MS_p - MS_{i,pi,e}}{n_i}$
pi, e	$MS_{i,pi,e}$	$\sigma^2_{i,pi,e}$	$\hat{\sigma}^2_{i,pi,e} = MS_{i,pi,e}$

logistical considerations. A person's observed score is decomposed into the following effects:

$$\begin{aligned} X_{pi} &= \mu && \text{(grand mean)} \\ &\quad + (\mu_p - \mu) && \text{(person effect)} \\ &\quad + (X_{pi} - \mu_p) && \text{(residual effect).} \end{aligned} \tag{27}$$

Unlike the one-facet crossed $p \times i$ design, the nested $i : p$ design has no separate term for the item effect. Because μ_i and μ_{pi} a re confounded in this design, the item effect is part of the residual term. The variance component for persons, the universe-score variance, is defined in the same way as for crossed designs. The variance component for the residual effect is:

$$\sigma^2_{i,pi,e} = \mathrm{E}_p\mathrm{E}_i(X_{pi} - \mu_p)^2. \tag{28}$$

The variance of the collection of observed scores X_{pi} for all persons and items is:

$$\sigma^2_{X_{pi}} = \sigma^2_p + \sigma^2_{i,pi,e}. \tag{29}$$

Numerical estimates of the variance components are obtained by setting the expected mean squares equal to the observed mean squares and solving the set of simultaneous equations as shown in Table 5.

A one-facet, nested D-study design where decisions are to be made on the basis of X_{pI} is denoted as $I : p$. Because the item (I) effect is confounded with the usual residual term (pI, e), error variance for relative decisions (σ^2_δ) is the same as the error variance for absolute decisions (σ^2_Δ):

$$\sigma^2_\delta = \sigma^2_\Delta = \sigma^2_{I,pI,e} = \frac{\sigma^2_{i,pi,e}}{n'_i}. \tag{30}$$

Consequently, for this design, the generalizability and phi coefficients are the same ($\mathrm{E}\rho^2 = \Phi$).

3.1.3. A crossed generalizability study and a nested decision study

Generalizability theory allows the decision maker to use different designs in the G- and D-studies. Although G-studies should use crossed designs whenever possible to avoid confounding of effects, D-studies may use nested designs for convenience or for increasing sample size, which typically reduces estimated error variance and, hence, increases estimated generalizability. A one-facet, crossed $p \times i$ G-study can be used to estimate variance components, error variance, and generalizability and phi coefficients for a one-facet, nested $I : p$ D-study. The variance component for the effects that are confounded in the nested design (i and pi, e) is the sum of the corresponding variance components in the crossed G-study:

$$\sigma^2_{i,pi,e} = \sigma^2_i + \sigma^2_{pi,e}. \tag{31}$$

The estimated residual variance for use in the D-study, then, is the sum of the following terms from the G-study:

$$\hat{\sigma}^2_{I,pI,e} = \frac{\hat{\sigma}^2_i}{n'_i} + \frac{\hat{\sigma}^2_{pi,e}}{n'_i}. \tag{32}$$

Universe-score variance (σ^2_p) remains the same in the crossed and nested designs.

3.2. Multifacet designs

The procedures for one-facet designs are easily extended to multifacet designs. Consider a two-facet crossed $p \times t \times o$ (person by task by observer) design in which tasks and observers are randomly sampled from infinitely large universes (random-effects design).

Table 6
Expected mean square equations for a random-effects, two-facet, crossed $p \times t \times o$ design

Source of variation	Variance component	Expected mean square equation
Persons (p)	σ^2_p	$EMS_p = \sigma^2_{pto,e} + n_o\sigma^2_{pt} + n_t\sigma^2_{po} + n_t n_o\sigma^2_p$
Tasks (t)	σ^2_t	$EMS_t = \sigma^2_{pto,e} + n_p\sigma^2_{to} + n_o\sigma^2_{pt} + n_p n_o\sigma^2_t$
Observers (o)	σ^2_o	$EMS_o = \sigma^2_{pto,e} + n_p\sigma^2_{to} + n_t\sigma^2_{po} + n_p n_t\sigma^2_o$
pt	σ^2_{pt}	$EMS_{pt} = \sigma^2_{pto,e} + n_o\sigma^2_{pt}$
po	σ^2_{po}	$EMS_{po} = \sigma^2_{pto,e} + n_t\sigma^2_{po}$
to	σ^2_{to}	$EMS_{to} = \sigma^2_{pto,e} + n_p\sigma^2_{to}$
pto, e	$\sigma^2_{pto,e}$	$EMS_{pto,e} = \sigma^2_{pto,e}$

The observed score for a particular person (X_{pto}) is:

$$
\begin{aligned}
X_{pto} = {} & \mu && \text{(grand mean)} \\
& + \mu_p - \mu && \text{(person effect)} \\
& + \mu_t - \mu && \text{(task effect)} \\
& + \mu_o - \mu && \text{(observer effect)} \\
& + \mu_{pt} - \mu_p - \mu_t + \mu && \text{(person } \times \text{ task effect)} \\
& + \mu_{po} - \mu_p - \mu_o + \mu && \text{(person } \times \text{ observer effect)} \\
& + \mu_{to} - \mu_t - \mu_o + \mu && \text{(task } \times \text{ observer effect)} \\
& + X_{pto} - \mu_{pt} - \mu_{po} - \mu_{to} + \mu_p \\
& + \mu_t + \mu_o - \mu && \text{(residual),} \qquad (33)
\end{aligned}
$$

where $\mu = \mathrm{E}_p\mathrm{E}_t\mathrm{E}_o X_{pto}$ and $\mu_p = \mathrm{E}_t\mathrm{E}_o X_{pto}$ and other terms in (33) are defined analogously. The collection of observed scores, X_{pto} has a variance $\mathrm{E}_p\mathrm{E}_t\mathrm{E}_o(X_{pto} - \mu)^2 = \sigma^2_{X_{pto}}$, which equals the sum of the variance components:

$$\sigma^2_{X_{pto}} = \sigma^2_p + \sigma^2_t + \sigma^2_o + \sigma^2_{pt} + \sigma^2_{po} + \sigma^2_{to} + \sigma^2_{pto,e}. \qquad (34)$$

Estimates of the variance components in (34) can be obtained by setting the expected mean squares equal to the observed mean squares and solving the set of simultaneous equations shown in Table 6.

In a fully crossed $p \times O \times T$ decision (D) study, error for relative decisions is defined as:

$$\delta_{pTO} = (X_{pTO} - \mu_{TO}) - (\mu_p - \mu), \qquad (35)$$

where $\mu_p = \mathrm{E}_T\mathrm{E}_O X_{pTO}$ and $\mu_{TO} = \mathrm{E}_p X_{pTO}$. The variance of the errors for relative decisions is:

$$\sigma^2_\delta = \mathrm{E}_p\mathrm{E}_T\mathrm{E}_O\delta^2_{pTO} = \sigma^2_{pT} + \sigma^2_{pO} + \sigma^2_{pTO,e} = \frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pto,e}}{n'_t n'_o}. \qquad (36)$$

Table 7
Possible random-effects D-study designs from a random-effects, two-facet G-study design

G-study design	D-study design
$p \times t \times o$	$p \times T \times O$
	$p \times (T : O)$ or $p \times (O : T)$
	$(O : p) \times T$ or $(T : p) \times O$
	$O : (p \times T)$ or $T : (O \times p)$
	$(O \times T) : p$
	$O : T : p$ or $T : O : p$
$p \times (o : t)$	$p \times (O : T)$
	$O : T : p$
$p \times (t : o)$	$p \times (T : O)$
	$T : O : p$
$o : (p \times t)$	$O : (p \times T)$
	$O : T : p$
$t : (p \times o)$	$T : (p \times O)$
	$T : O : p$
$(o \times t) : p$	$(O \times T) : p$
	$O : T : p$ or $T : O : p$

The "main effects" of observer and task do not enter into error for relative decisions because, for example, all people are observed on the same tasks so any difference in task difficulty affects all persons and does not change their relative standing.

For absolute decisions, the error in a random-effects $p \times T \times O$ design is defined as:

$$\Delta_{pTO} \equiv X_{pTO} - \mu_p, \tag{37}$$

and the variance of the errors is:

$$\begin{aligned}\sigma^2_\Delta = \mathrm{E}_p\mathrm{E}_T\mathrm{E}_O\Delta^2_{pTO} &= \sigma^2_T + \sigma^2_O + \sigma^2_{pT} + \sigma^2_{pO} + \sigma^2_{TO} + \sigma^2_{pTO,e} \\ &= \frac{\sigma^2_t}{n'_t} + \frac{\sigma^2_o}{n'_o} + \frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{to}}{n'_t n'_o} + \frac{\sigma^2_{pto,e}}{n'_t n'_o}.\end{aligned} \tag{38}$$

With absolute decisions, the main effects of observers and tasks – the strictness of the observers or the difficulty of the tasks – do influence absolute performance and so are included in the definition of measurement error.

With these definitions of error variance, the generalizability coefficient ($\mathrm{E}\rho^2$) and phi coefficient (Φ) are defined as in (20) and (23).

The results of a two-facet crossed G-study can be used to estimate error variance and generalizability and phi coefficients for a wide range of D-study designs. Any G-study can be used to estimate the effects in a D-study design with the same or more nesting than in the G-study design. Table 7 lists the possible two-facet G- and D-study designs for which p is the object of measurement and is not nested within a facet.

A decision-maker who uses a crossed $p \times t \times o$ generalizability study design may, for example, choose to use a $p \times (T : O)$ design in the decision study for convenience. In this D-study design, each observer rates different tasks and the decision-maker can

sample a larger number of tasks in the study than would be the case if tasks and observers were crossed (e.g., if two observers rate performance on 5 tasks, then 10 tasks can be represented). Using the variance components from the $p \times t \times o$ G-study, error variances for relative and absolute decisions for a $p \times (T : O)$ D-study design are:

$$\sigma_\delta^2 = \sigma_{pO}^2 + \sigma_{pT:O}^2 = \frac{\sigma_{po}^2}{n'_o} + \frac{\sigma_{pt,pot,e}^2}{n'_o n'_t}, \tag{39}$$

$$\sigma_\Delta^2 = \sigma_O^2 + \sigma_{pO}^2 + \sigma_{T:O}^2 + \sigma_{pT:O}^2 = \frac{\sigma_o^2}{n'_o} + \frac{\sigma_{po}^2}{n'_o} + \frac{\sigma_{t,to}^2}{n'_t n_o} + \frac{\sigma_{pt,pto,e}^2}{n'_t n'_o}, \tag{40}$$

where $\sigma_{t,to}^2$ is the sum of σ_t^2 and σ_{to}^2 from the $p \times t \times o$ G-study, and $\sigma_{pt,pto,e}^2$ is the sum of σ_{pt}^2 and $\sigma_{pto,e}^2$. Even if the number of tasks and the number of raters per task are held constant, nesting observers within tasks will result in improved precision, because a larger number of observers is sampled. Thus, the variance components σ_o^2 and σ_{po}^2 are divided by $n'_t n'_o$ instead of n'_o, for example, as illustrated in Example 1 below.

EXAMPLE 1 (Generalizability study of job performance measurements of Navy machinist mates). Navy machinist mates ($n_p = 26$) aboard ship were observed by two examiners as they carried out 11 tasks in the engine room (concerning knowledge of gauges, equipment operation, casualty control; Webb et al., 1989). The task score was the proportion of steps in a task (ranging from 12 to 49 steps per task) that were judged to be completed correctly. Two examiners observed each machinist mate on all tasks, so the design of the generalizability study was persons (machinist mates) crossed with tasks and observers ($p \times t \times o$). Tasks and observers were considered random facets.

Table 8 gives the estimated variance components from the G-study, and estimated error variances and generalizability and dependability coefficients for fully crossed, $p \times T \times O$ D-study designs with different numbers of tasks (n'_t) and observers (n'_o). Observer was a negligible source of variation. The variance components for the main effect of observer, the interaction between persons and observers, and the interaction between observers and tasks were all zero or near zero. Observers not only rank ordered machinist mates similarly ($\hat{\sigma}_{po}^2 = 0$), but they also gave nearly identical scores, on average, to the machinist mates ($\hat{\sigma}_o^2 = 0$). The task, on the other hand, was a substantial source of variation. The relatively large component for the main effect of task ($\hat{\sigma}_t^2 = 0.00970$) shows that tasks differed in difficulty level. Furthermore, the very large variance component for the interaction between persons and tasks ($\hat{\sigma}_{pt}^2 = 0.02584$; 60% of the total variation) suggests that the relative standing of persons differed across tasks (cf. Shavelson et al., 1993; Shavelson et al., 1999).

The estimated variance components from the G-study suggest that a single observer would probably provide dependable ratings but that multiple tasks are needed to represent job requirements. As can be seen in Table 8, averaging over 11 tasks (the same number as used in the G-study) with a single observer yields moderate estimated generalizability and dependability coefficients ($E\hat{\rho}^2 = 0.716$; $\hat{\Phi} = 0.649$). Using two observers has no appreciable effect on the results. A very large number of tasks, of

Table 8
Generalizability study of job performance scores (2-facet, random effects $p \times t \times o$ design)

Source of variation	G-study $\hat{\sigma}^2$	Size of decision study		
		$n'_t = 11$ $n'_o = 1$	$n'_t = 11$ $n'_o = 2$	$n'_t = 17$ $n'_o = 1$
Estimated variance components				
Person (p)	0.00626	0.00626	0.00626	0.00626
Task (t)	0.00970	0.00088	0.00088	0.00057
Observer (o)	0.00000[a]	0.00000	0.00000	0.00000
pt	0.02584	0.00235	0.00235	0.00152
po	0.00000[a]	0.00000	0.00000	0.00000
to	0.00003	0.00000	0.00000	0.00000
pto, e	0.00146	0.00013	0.00007	0.00009
Error variances				
$\hat{\sigma}^2_\delta$		0.00248	0.00242	0.00161
$\hat{\sigma}^2_\Delta$		0.00339	0.00330	0.00218
Coefficients				
$\mathrm{E}\hat{\rho}^2$		0.716	0.721	0.795
$\hat{\Phi}$		0.649	0.655	0.742

[a]Negative estimated variance component set equal to zero.

questionable feasibility, would be needed to obtain a reasonable level of generalizability (with 17 tasks and one observer, $\mathrm{E}\hat{\rho}^2 = 0.795$; $\hat{\Phi} = 0.742$). Even this level of dependability may be insufficient for individual decision making, however. Using the square root of absolute error variance, the standard error of measurement (SEM) for absolute decisions, to calculate a confidence interval for machinist mates' universe scores shows that observed scores are likely to vary considerably over 17-task samples. For 17 tasks and one observer, $\hat{\sigma}_\Delta = 0.047$; a 95% confidence interval is $X_{pTO} \pm 0.092$. Compared to the possible range of performance scores (0 to 1), the width of this interval is substantial.

For convenience, the decision maker may elect to use a $p \times (T : O)$ decision study design, in which machinist mates perform all tasks, but each examiner is responsible for judging machinist mates on only a subset of tasks. For example, consider a $p \times (T : O)$ decision study in which two examiners each observe machinist mates performing 11 tasks and the tasks are different for each examiner. The partially nested $p \times (T : O)$ design with $n'_t = 11$ and $n'_o = 2$, then, actually involves using 22 tasks. This design yields $\hat{\sigma}^2_\delta = 0.00124$ and $\hat{\sigma}^2_\Delta = 0.00168$ and $\mathrm{E}\hat{\rho}^2 = 0.835$ and $\hat{\Phi} = 0.788$. Comparing these values to those for the crossed D-study with the same number of observations ($n'_t = 11$ and $n'_o = 2$) shows smaller error variances and larger generalizability coefficients for the partially nested D-study than for the crossed D-study. The error variances in the partially nested design are reduced by virtue of the larger denominators. Because $\hat{\sigma}^2_t$ is

confounded with $\hat{\sigma}^2_{to}$ in the partially nested design, it is divided by the number of tasks *and* the number of observers ($n'_t n'_o$). Similarly, $\hat{\sigma}^2_{pt}$ is divided by $n'_t n'_o$ by virtue of its being confounded with $\hat{\sigma}^2_{pto,e}$. The partially nested $p \times (T : O)$ design has an advantage over the fully crossed $p \times T \times O$ design in terms of estimated generalizability as well as convenience.

3.3. Random and fixed facets

G theory is essentially a random effects theory. Typically a random facet is created by randomly sampling levels of a facet (e.g., tasks from a job in observations of job performance). When the levels of a facet have not been sampled randomly from the universe of admissible observations but the intended universe of generalization is infinitely large, the concept of exchangeability may be invoked to consider the facet as random (de Finetti, 1964). Formally,

> *The random variables $X_1, \ldots, X_n$ are exchangeable if the $n!$ permutations $(X_{k1}, \ldots, X_{kn})$ have the same n-dimensional probability distribution. The variables of an infinite sequence X_n are exchangeable if $X_1, \ldots, X_n$ are exchangeable for each n* (Feller, 1966, p. 225).

Even if conditions of a facet have not been sampled randomly, the facet may be considered to be random if conditions not observed in the G-study are exchangeable with the observed conditions.

A fixed facet (cf. fixed factor in analysis of variance) arises when the decision maker: (a) purposely selects certain conditions and is not interested in generalizing beyond them, (b) finds it unreasonable to generalize beyond the conditions observed, or (c) when the entire universe of conditions is small and all conditions are included in the measurement design. G theory typically treats fixed facets by averaging over the conditions of the fixed facet and examining the generalizability of the average over the random facets (Brennan, 2001; Cronbach et al., 1972). When it does not make conceptual sense to average over the conditions of a fixed facet, a separate G-study may be conducted within each condition of the fixed facet (Shavelson and Webb, 1991) or a full multivariate analysis may be performed with the conditions of the fixed facet comprising a vector of dependent variables (Brennan, 2001; see below).

G theory recognizes that the universe of admissible observations in a G-study may be broader than the universe of generalization of interest in a D-study. The decision maker may reduce the levels of a facet (creating a fixed facet), select (and thereby control) one level of a facet, or ignore a facet. A facet is fixed in a D-study when $n' = N' < \infty$, where n' is the number of levels for a facet in the D-study and N' is the total number of levels for a facet in the universe of generalization. Consider a person by rater by subject matter ($p \times r \times s$) random-effects G-study in which raters (r) observe the behavior of teachers (p) while teaching multiple subject matters (s). The universe of admissible observations is defined by facets r and s of infinite size. Fixing the subject matter facet s in the D-study and averaging over the n_s conditions of facet s in the G-study ($n_s = n'_s$) yields the following universe-score variance:

$$\sigma^2_\tau = \sigma^2_p + \sigma^2_{pS} = \sigma^2_p + \frac{\sigma^2_{ps}}{n'_s}, \tag{41}$$

where σ_τ^2 denotes universe-score variance in generic terms. When facet s is fixed, the universe score is based on a person's average score over the finite subset of levels of facet s, so the generic universe-score variance in (41) is the variance over persons' mean scores for *just those levels*. Any person effects specific to those particular levels therefore become part of the universe-score variance. Hence, (41) includes σ_{pS}^2 as well as σ_p^2. $\hat{\sigma}_\tau^2$ is an unbiased estimate of universe-score variance for the mixed model only when the same levels of facet s are used in the G and D studies (Brennan, 2001). The relative and absolute error variances, respectively, are:

$$\sigma_\delta^2 = \sigma_{pR}^2 + \sigma_{pRS}^2 = \frac{\sigma_{pr}^2}{n_r'} + \frac{\sigma_{prs,e}^2}{n_r' n_s'}, \tag{42}$$

$$\sigma_\Delta^2 = \sigma_R^2 + \sigma_{pR}^2 + \sigma_{RS}^2 + \sigma_{pRS}^2 = \frac{\sigma_r^2}{n_r'} + \frac{\sigma_{pr}^2}{n_r'} + \frac{\sigma_{rs}^2}{n_r' n_s'} + \frac{\sigma_{prs,e}^2}{n_r' n_s'}. \tag{43}$$

And the generalizability coefficient and index of dependability, respectively, are:

$$\mathrm{E}\rho^2 = \frac{\sigma_p^2 + \frac{\sigma_{ps}^2}{n_t'}}{\sigma_p^2 + \frac{\sigma_{ps}^2}{n_s'} + \frac{\sigma_{pr}^2}{n_r'} + \frac{\sigma_{prs,e}^2}{n_r' n_s'}}, \tag{44}$$

$$\Phi = \frac{\sigma_p^2 + \frac{\sigma_{ps}^2}{n_s'}}{\sigma_p^2 + \frac{\sigma_{ps}^2}{n_s'} + \frac{\sigma_r^2}{n_r'} + \frac{\sigma_{pr}^2}{n_r'} + \frac{\sigma_{rs}^2}{n_r' n_s'} + \frac{\sigma_{prs,e}^2}{n_r' n_s'}}. \tag{45}$$

EXAMPLE 2 (Generalizability study of teacher behavior). Elementary school teachers were observed by three raters in two subject matters: reading and mathematics (cf. Elrich and Shavelson (**?**)). Teachers (p) and raters (r) are considered random. Subject matter (s) is considered fixed because these subject matters were specifically selected to represent standard curriculum areas and it is probably not reasonable to generalize from teachers' behavior (for example, the number of high-level questions teachers ask) when teaching reading and mathematics to their behavior when teaching other subject matters. We consider a crossed design in which three raters observed behavior of all teachers in both reading and mathematics ($p \times r \times s$).

Table 9 gives the estimated variance components from the G-study treating raters (r) and subject matter (s) as random. With subject matter treated as fixed, and substituting values from Table 9 into (41), $\hat{\sigma}_\tau^2 = 2.34375 \cdot (1.17262 + 2.34226/2)$ for the universe score based on a teacher's average score over reading and mathematics. Substituting values from Table 9 into (42)–(45) yields estimated error variances and generalizability and phi coefficients for different numbers of raters in the decision study (n_r'). Results for one and two raters appear in Table 10. When averaging over teacher behavior in reading and mathematics, two raters are needed to reach an estimated level of generalizability of 0.80.

Because teacher behavior varied substantially across reading and mathematics (see, especially, the relatively large $\hat{\sigma}_{ps}^2$ in Table 9, showing that relative standing of teachers

Table 9
Generalizability study of teacher behavior ratings (2-facet $p \times r \times s$ design with all facets treated as random)

Source of variation	$\hat{\sigma}^2$
Teacher (p)	1.17262
Rater (r)	0.01488
Subject matter (s)	0.15774
pr	0.50595
ps	2.34226
rs	0.03869
prs, e	0.87798

Table 10
Decision study error variances and coefficients for study of teacher behavior (2-facet, mixed-effects $p \times r \times s$ design with s fixed)

	Averaging over subject matters	
	$n'_r = 1$	$n'_r = 2$
Universe-score variance		
$\hat{\sigma}^2_\tau$	2.34375	2.34375
Error variances		
$\hat{\sigma}^2_\delta$	0.94494	0.47247
$\hat{\sigma}^2_\Delta$	0.97917	0.48959
Coefficients		
$\mathrm{E}\hat{\rho}^2$	0.713	0.832
$\hat{\Phi}$	0.705	0.827

differed substantially from reading to mathematics), it may be reasonable to analyze teacher behavior for each subject matter separately, as well as averaging over them. Table 11 gives the results for decision studies of teacher behavior separately for reading and mathematics. When subject matters are analyzed separately, three raters are needed to reach an estimated level of generalizability of 0.80 for reading, but only one rater is needed to reach that level of generalizability for mathematics.

3.4. Symmetry

The purpose of psychological measurement has typically been to differentiate individuals. Recognizing that the focus of measurement may change depending on a particular decision-maker's purpose, Cardinet et al. (1976, 1981) spoke of the *principle of symmetry*: "The principle of symmetry of the data is simply an affirmation that each of the facets of a factorial design can be selected as an object of study, and that the operations defined for one facet can be transposed in the study of another facet" (Cardinet

Table 11
Generalizability study of teacher behavior for each subject matter (one-facet, random-effects $p \times r$ design for each subject matter)

Source of variation	Reading			Mathematics	
	$n'_r = 1$	$n'_r = 2$	$n'_r = 3$	$n'_r = 1$	$n'_r = 2$
Estimated variance components					
Person (p)	3.14286	3.14286	3.14286	3.88691	3.88691
Rater (r)	0.15476	0.07738	0.05159	0[a]	0
pr, e	2.26191	1.13096	0.75397	0.50595	0.25298
Error variances					
$\hat{\sigma}^2_\delta$	2.26191	1.13096	0.75397	0.50595	0.25298
$\hat{\sigma}^2_\Delta$	2.41667	1.20834	0.80556	0.50595	0.25298
Coefficients					
$E\hat{\rho}^2$	0.581	0.735	0.807	0.885	0.939
$\hat{\Phi}$	0.565	0.722	0.796	0.885	0.939

[a]Estimated variance component set equal to zero.

et al., 1981, p. 184). In a persons (p) $\times$ items (i) $\times$ occasions (o) design, while persons may be the focus of measurement for evaluators wishing to make dependable judgments about persons' performance, items may be the focus for curriculum developers wishing to calibrate items for use in item banks (e.g., Wood, 1976). In the latter case, individual differences among persons represent error variation, rather than universe-score variation, in the measurement.

The principle of symmetry leads to the possibility of multifaceted populations. (Cardinet et al., 1976, 1981). Educational evaluators may be interested in scholastic achievement of classes, schools, or districts, or in comparisons across years. Or the focus may be on items corresponding to different content units in which the universe-score of interest is that of items (i) nested within content units (c).

Another instance of multifaceted populations is stratified objects of measurement (Brennan, 2001). The population of interest may be persons nested within, say, geographic region, gender, or socio-economic status. If variation due to these attributes of persons is negligible, the decision maker can assume a homogeneous population and so reduce the design of the D-study. If the variation is large, the decision maker may calculate separate variance component estimates for each subgroup.

3.4.1. Generalizability of group means

A common area of application of the principle of symmetry is the generalizability of group means (e.g., scholastic achievement of classes, schools, or school districts; student evaluations of teachers; see Kane et al., **?**; Kane and Brennan, 1977, for a detailed explication of the generalizability of class means). In these cases, the group, not the individual student, is the object of study and the sampling of students within groups gives

rise to errors of measurement. In a fully random design with persons (p) nested within groups (g) and crossed with items (i), denoted as $(p : g) \times i$, the generalizability coefficient with groups as the object of measurement and when generalizing over persons and items is:

$$\mathrm{E}\rho_g^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{p:g}^2}{n'_p} + \frac{\sigma_{gi}^2}{n'_i} + \frac{\sigma_{pi:g,e}^2}{n'_p n'_i}}. \tag{46}$$

In (46), variation among persons within groups ($\sigma_{p:g}^2$) contributes to error variation but not universe-score variation. Designs that fail to recognize explicitly persons as a source of variation (i.e., a groups × items design) would result in misestimation of generalizability due to the confounding of σ_p^2 and $\sigma_{p:g}^2$.

In some cases, decision makers may be interested in restricted universes of generalization in which one or more sources of variation are fixed. For example, persons may be treated as a fixed facet. In industrial settings where supervisors' ratings of employees (p) in organizations (g) are gathered, employees may be considered as a fixed facet within some period of time. Similarly, in a study of school performance, interest may lie in estimating the performance of the student body during a particular year (Cronbach et al., 1997). For the $(p : g) \times i$ design in which persons are fixed (n_p persons per group) and items are random, the generalizability coefficient for groups is:

$$\mathrm{E}\rho_g^2 \text{ with } p \text{ fixed} = \frac{\sigma_g^2 + \frac{\sigma_{p:g}^2}{n_p}}{\sigma_g^2 + \frac{\sigma_{p:g}^2}{n_p} + \frac{\sigma_{gi}^2}{n'_i} + \frac{\sigma_{pi:g,e}^2}{n_p n'_i}}. \tag{47}$$

Note, however, that Cronbach et al. (1997) caution that in the typical case, when the intended inference concerns the performance of the school with potentially different students, persons (students) should be treated as a random facet, even if all students present in the school are tested.

Alternatively, the focus may be on a fixed set of items. For the $(p : g) \times i$ design in which items are fixed (a fixed set of n_i items) and persons are random, the generalizability coefficient for groups is:

$$\mathrm{E}\rho_g^2 \text{ with } i \text{ fixed} = \frac{\sigma_g^2 + \frac{\sigma_{gi}^2}{n_i}}{\sigma_g^2 + \frac{\sigma_{gi}^2}{n_i} + \frac{\sigma_{p:g}^2}{n_p} + \frac{\sigma_{pi:g,e}^2}{n'_p n_i}}. \tag{48}$$

Cronbach et al. (1997) describe more complex multilevel designs with persons nested within classes, and classes nested within schools. They consider estimation of variance components and school-level standard errors when generalizing to infinite and finite populations. As an alternative to estimating standard errors for school means, Yen (**?**) showed how to estimate standard errors for the percent of students in a school above a cutpoint (percent above cutpoint, or PAC) and compared the results for a variety of generalizability study models.

Challenging conventional wisdom that reliability for groups is larger than reliability for persons, Brennan (1995) provides a number of conditions in which the generalizability of group means may be *less* than the generalizability of scores for persons. For

the fully random $(p : g) \times i$ design in which *person* is the object of measurement, the generalizability coefficient is:

$$\mathrm{E}\rho_p^2 = \frac{\sigma_g^2 + \sigma_{p:g}^2}{\sigma_g^2 + \sigma_{p:g}^2 + \frac{\sigma_{gi}^2}{n_i'} + \frac{\sigma_{pi:g,e}^2}{n_i'}}. \tag{49}$$

As described by Brennan (1995, p. 393), $\mathrm{E}\rho_g^2$ in (49) will tend to be less than $\mathrm{E}\rho_p^2$ when group means are similar (small σ_g^2), the number of items (n_i') is large, the number of persons within each group (n_p') is small, or variation among persons within groups ($\sigma_{p:g}^2$) is relatively large. Brennan (1995) also discusses conditions under which the generalizability of group means may be less than the generalizability of persons for restricted universes (e.g., persons are fixed and items are random).

3.5. Multivariate generalizability

For behavioral measurements involving multiple scores describing individuals' personality, aptitudes, skills, or performance, multivariate generalizability can be used to (a) estimate the reliability of difference scores, observable correlations, or universe-score and error correlations for various D-study designs and sample sizes (Brennan, 2001), (b) estimate the reliability of a profile of scores using multiple regression of universe scores on the observed scores in the profile (Brennan, 2001; Cronbach et al., 1972), or (c) produce a composite of scores with maximum generalizability (Shavelson and Webb, 1981). For all of these purposes, multivariate G theory decomposes both observed variances and covariances into components. Consider a one-facet, crossed $p \times r$ design with teacher behavior divided into two dependent variables – behavior in reading and behavior in mathematics – and the same raters observing teacher behavior in both subject matters. The observed scores for the reading and mathematics variables for person p observed under condition r can be denoted as ${}_1(X_{pr})$ and ${}_2(X_{pr})$, respectively. The variances and covariance of observed scores, $\sigma^2_{{}_1(X_{pr})}$ and $\sigma^2_{{}_2(X_{pr})}$ are decomposed as follows:

$$\begin{aligned}
&\begin{bmatrix} \sigma^2_{{}_1(X_{pr})} & \sigma_{{}_1(X_{pr}),{}_2(X_{pr})} \\ \sigma_{{}_1(X_{pr}),{}_2(X_{pr})} & \sigma^2_{{}_2(X_{pr})} \end{bmatrix} \\
&\quad = \begin{bmatrix} \sigma^2_{{}_1 p} & \sigma_{{}_1 p,{}_2 p} \\ \sigma_{{}_1 p,{}_2 p} & \sigma^2_{{}_2 p} \end{bmatrix} && \text{(person)} \\
&\qquad + \begin{bmatrix} \sigma^2_{{}_1 r} & \sigma_{{}_1 r,{}_2 r} \\ \sigma_{{}_1 r,{}_2 r} & \sigma^2_{{}_2 r} \end{bmatrix} && \text{(rater)} \\
&\qquad + \begin{bmatrix} \sigma^2_{{}_1(pr,e)} & \sigma_{{}_1(pr,e),{}_2(pr,e)} \\ \sigma_{{}_1(pr,e),{}_2(pr,e)} & \sigma^2_{{}_2(pr,e)} \end{bmatrix} && \text{(residual).}
\end{aligned} \tag{50}$$

In (50) the term $\sigma_{{}_1 p,{}_2 p}$ is the covariance between universe scores for behavior in reading and behavior in mathematics. Dividing this covariance by the product of the square root of the corresponding variances ($\sigma_{{}_1 p} * \sigma_{{}_2 p}$) produces the disattenuated correlation between teacher behavior in reading and mathematics. The remaining covariance terms

in (50) are error covariance components. The term $\sigma_{1r,2r}$, for example, is the covariance between teacher behavior scores in reading and mathematics due to consistency of rater effects across the two subject matters. Estimates of the covariance components can be obtained by solving the expected mean product equations in analogous fashion to their univariate counterparts.

An important aspect of the development of multivariate G theory is the distinction between linked and unlinked conditions. The expected values of error covariance components are nonzero when levels are linked or jointly sampled (e.g., the same raters are used to produce scores on the two variables in the teacher behavior profile). The expected values of error covariance components are zero when conditions for observing different variables are unlinked, that is, selected independently (e.g., the raters used to obtain scores on one variable in a profile, teaching behavior in reading, are selected independently of the raters used to obtain scores on another variable, teaching behavior in mathematics).

Joe and Woodward (1976) presented a generalizability coefficient for a multivariate composite that maximizes the ratio of universe-score variation to universe-score plus error variation by using statistically derived weights for each dependent variable (e.g., teacher behavior in reading and mathematics). For a one-facet, crossed $p \times r$ design, their multivariate analogue to a univariate generalizability coefficient is:

$$\mathrm{E}\rho^2 = \frac{\boldsymbol{a}'\boldsymbol{V}_p\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{V}_p\boldsymbol{a} + \frac{\boldsymbol{a}'\boldsymbol{V}_{pr,e}\boldsymbol{a}}{n'_r}}, \tag{51}$$

where $\boldsymbol{V}$ is a matrix of variance and covariance components, n'_r is the number of conditions r in the D-study, and $\boldsymbol{a}$ is the vector of weights that maximizes the ratio of between-person to between-person plus within-person variance component matrices. $\mathrm{E}\hat{\rho}^2$ and $\boldsymbol{a}$ can be obtained by solving the following set of equations:

$$\left[\boldsymbol{V}_p - \rho_k^2(\boldsymbol{V}_p + \boldsymbol{V}_\delta)\right]\boldsymbol{a} = 0, \tag{52}$$

where the subscript k refers to the m ($k = 1, \ldots, m$) roots of (52) and $\boldsymbol{V}_\delta$ is the multivariate analogue to σ_δ^2. For each multivariate generalizability coefficient corresponding to a characteristic root in (52), there is a set of canonical coefficients that defines a composite of the scores. By definition, the first composite is the most reliable while the last composite is the least reliable. Because the data, not the investigators, define the composites of maximum generalizability, this approach would be used in an exploratory, rather than confirmatory, context.

Alternatives to maximizing the generalizability of a composite are to determine variable weights on the basis of expert judgment or use weights derived from a confirmatory factor analysis (Marcoulides, 1994; Short et al., **?**). The most straightforward approach to estimating the generalizability of a composite with weights $\boldsymbol{a}$ is a univariate rather than multivariate analysis. The results of a univariate generalizability analysis would be identical to those of a multivariate generalizability analysis in which the weights of the composite define the $\boldsymbol{a}$ vector in (52).

EXAMPLE 3 (Multivariate generalizability study of teacher behavior). In Example 2, we presented the results of a univariate generalizability study of teacher behavior in

Table 12
Estimated variance and covariance components for multivariate generalizability study of teacher behavior ratings (1-facet, random-effects $p \times r$ design)

Source of variation		(1) Reading	(2) Mathematics
Persons (p)	(1)	3.14286	1.17262
	(2)	1.17262	3.88691
Raters (r)	(1)	0.15476	0.01488
	(2)	0.01488	0[a]
pr, e	(1)	2.26191	0.50595[b]
	(2)	0.50595	0.50595[b]

[a]Negative estimated variance component set equal to zero.

[b]The equal values for these components are correct.

which teachers (p) and raters (r) were considered random and subject matter (s) was considered fixed. Here we present the results of a multivariate generalizability analysis of the same data with teacher behavior in reading and mathematics considered as two dependent variables in a multivariate profile. The design is persons (p) crossed with raters (r).

Table 12 gives the estimated variance and covariance components from the multivariate generalizability study. The estimated variance components (on the main diagonal) are the same as those produced by the univariate analyses of teacher behavior in reading and mathematics (see Table 11). The estimated covariance components provide new information. The positive, nonzero estimated covariance for persons ($\hat{\sigma}_{1p,2p} = 1.17262$) indicates that raters found consistencies in teacher behavior across reading and mathematics. The estimated variance and covariance component for persons can be used to estimate the correlation between teacher behavior in reading and mathematics corrected for attenuation:

$$\frac{\hat{\sigma}_{1p,2p}}{\sqrt{\hat{\sigma}^2_{1p}\hat{\sigma}^2_{2p}}}. \tag{53}$$

Substituting values from Table 12 gives a corrected correlation equal to 0.34, indicating that teachers' universe scores for behavior are modestly correlated in reading and mathematics.

The very small estimated covariance component for raters ($\hat{\sigma}_{1r,2r} = 0.01488$) shows that any disagreements among raters in the overall frequency of teacher behavior observed did not carry over from reading to mathematics. The substantial estimated covariance component for the residual ($\hat{\sigma}_{1(pr,e),2(pr,e)} = 0.50595$) suggests that disagreements among raters about the relative standing of teachers are similar in reading and mathematics; unexplained factors that contribute to the variation of behavior ratings within

a subject matter contribute to the covariation between behavior ratings in reading and mathematics, or both.

In this example, a composite that weights mathematics more than reading will have higher estimated generalizability than will a composite that weights the two subject matters equally; compare $E\hat{\rho}^2 = 0.902$ for a composite with weights of 0.25 for reading and 0.75 for math (2 raters in the D-study) with $E\hat{\rho}^2 = 0.832$ for a composite in which reading and mathematics are equally weighted (from Table 10).

3.6. Additional issues in generalizability theory

3.6.1. Variance component estimates

We treat three issues of concern when estimating variance components: (1) the variability of variance-component estimates, (2) negative estimated variance components, and (3) complexity of estimation in unbalanced designs.

Variability of variance-component estimates

The first concern is that estimates of variance components may be unstable with usual sample sizes (Cronbach et al., 1972). Assuming that mean squares are independent and score effects have a multivariate normal distribution, the sampling variance of an estimated variance component ($\hat{\sigma}^2$) is

$$\operatorname{var}(\hat{\sigma}^2) = \frac{2}{c^2}\sum_q \operatorname{var} MS_q = \frac{2}{c^2}\sum_q \frac{EMS_q^2}{df_q}, \tag{54}$$

where c is the constant associated with the estimated variance component, EMS_q is the expected value of the mean square, MS_q, and df_q is the degrees of freedom associated with MS_q (Searle, 1971; Smith, 1978). In (54), the summation is taken over all MS_q that are used to obtain the ANOVA estimate of the variance component.

The variances of estimated variance components will be larger with smaller numbers of observations (reflected in smaller df_q). Moreover, the more mean squares that are involved in estimating variance components, the larger the estimated variances are likely to be. In a two-facet, random-effects $p \times t \times o$ design, for example, σ_p^2 is estimated by $(MS_p - MS_{pt} - MS_{po} + MS_{pto,e})/(n_t n_o)$, c in (54) refers to $n_t n_o$, and MS_q refers to MS_p, MS_{pt}, MS_{po}, and $MS_{pto,e}$. Following Smith (1978), the variance of $\hat{\sigma}_p^2$ is:

$$\begin{aligned}\operatorname{var}(\hat{\sigma}_p^2) = \frac{2}{n_p - 1}\Bigg[&\left(\sigma_p^2 + \frac{\sigma_{pt}^2}{n_t} + \frac{\sigma_{po}^2}{n_o} + \frac{\sigma_{pro,e}^2}{n_t n_o}\right)^2 \\ &+ \frac{1}{n_t - 1}\left(\frac{\sigma_{pt}^2}{n_t} + \frac{\sigma_{pto,e}^2}{n_t n_o}\right)^2 \\ &+ \frac{1}{n_o - 1}\left(\frac{\sigma_{po}^2}{n_o} + \frac{\sigma_{pto,e}^2}{n_t n_o}\right)^2 \\ &+ \frac{1}{(n_t - 1)(n_o - 1)}\left(\frac{\sigma_{pto,e}^2}{n_t n_o}\right)^2\Bigg].\end{aligned} \tag{55}$$

In contrast, the variance of $\hat{\sigma}^2_{pto,e}$ is based on only a single mean square:

$$\text{var}\left(\hat{\sigma}^2_{pto,e}\right) = \frac{2}{(n_p - 1)(n_t - 1)(n_o - 1)}\sigma^4_{pto,e}. \tag{56}$$

For the estimated variance components in Table 8 (generalizability study of job performance measurements of Navy machinist mates with a $p \times t \times o$ design), compare estimated standard errors produced by replacing variance components in (55) and (56) with ANOVA estimates of them: $\hat{\sigma}(\hat{\sigma}^2_p) = 0.00246$ (for $\hat{\sigma}^2_p = 0.00626$) and $\hat{\sigma}(\hat{\sigma}^2_{pto,e}) = 0.00013$ (for $\hat{\sigma}^2_{pto,e} = 0.00146$). (The preceding estimates are biased. Unbiased estimates of the standard errors, obtained by replacing df_q with $df_q + 2$ in (54), yield 0.00237 and 0.00013 for $\hat{\sigma}(\hat{\sigma}^2_p)$ and $\hat{\sigma}(\hat{\sigma}^2_{pto,e})$, respectively.)

Smith (1981) also pointed out that the stability of variance component estimates varies across design complexity (the number of facets) and design configurations (the extent of nesting in the design). As an example of the latter, assuming the same number of observations, the sampling variance of $\hat{\sigma}^2_p$ would be smaller in a partially nested $p \times (t : o)$ design than in a fully crossed $p \times t \times o$ design. Compare $\text{var}(\hat{\sigma}^2_p)$ in (57) below for the partially nested $p \times (t : o)$ design to $\text{var}(\hat{\sigma}^2_p)$ in (55) above for the fully crossed $p \times t \times o$ design:

$$\text{var}\left(\hat{\sigma}^2_p\right) = \frac{2}{n_p - 1}\left[\left(\sigma^2_p + \frac{\sigma^2_{po}}{n_o} + \frac{\sigma^2_{pto,e}}{n_t n_o}\right)^2 + \frac{1}{(n_o - 1)}\left(\frac{\sigma^2_{po}}{n_o} + \frac{\sigma^2_{pto,e}}{n_t n_o}\right)^2\right]. \tag{57}$$

Smith (1981) used this result to make recommendations for increasing the sampling stability of variance component estimates. He suggested implementing multiple designs in which the number of mean square estimates used to compute variance component estimates is smaller than would be used in a single, more complex design. As an illustration, Smith (1981) considered a situation in which pupils (p) within schools (s) were administered items (i) within test forms (f). A single design configuration that would make it possible to estimate all possible variance components is $(p : s) \times (i : f)$ in which a sample of pupils from each school is administered samples of items from multiple test forms. As an alternative to this comprehensive design, Smith described two designs that require fewer mean squares to compute variance component estimates deemed likely to have the largest errors in the comprehensive design: a $p : [(s \times (i : f)]$ design (different sets of pupils within each school are administered each item in each test form, that is, there is a different set of pupils for each school-item combination) and a $i : [f \times (p : s)]$ design (for each school, samples of items from each test form are administered to a sample of pupils in that school such that there is a different set of items for each form–pupil combination). Together, these alternative designs make it possible to estimate all of the variance components in the single, comprehensive design. Smith showed that estimates of variance components from randomly sampled replications of the designs were more stable for the multiple designs (although less practical) than for the single, comprehensive design (using a fixed number of observations).

Another strategy for increasing the stability of variance component estimates involves the allocation of measurements (e.g., n_i and n_j). In a $p \times i \times j$ design, for example, as $\hat{\sigma}^2_{pij,e}$ increases relative to $\hat{\sigma}^2_{pi}$ and $\hat{\sigma}^2_{pj}$, the optimal solution tends toward $n_i = n_j$; as $\hat{\sigma}^2_{pij,e}$ decreases relative to $\hat{\sigma}^2_{pi}$ and $\hat{\sigma}^2_{pj}$, the optimal solution is to make n_i and n_j proportional to $\hat{\sigma}^2_{pi}/\hat{\sigma}^2_{pj}$ (Smith, 1978).

In contrast to ANOVA-based procedures for estimating standard errors of estimated variance components, approaches for estimating standard errors that do not make any distributional-form assumptions include the jackknife (Tukey, 1958) and bootstrap resampling methods (Efron, 1982; Efron and Tibshirani, 1986; Shao and Tu, 1995). Both methods involve creating multiple samples from a dataset, conducting separate analyses of the samples, and examining the variability of the estimates across the samples. The jackknife procedure creates samples by omitting one data point at a time; the bootstrap procedure samples with replacement. Extending these procedures to obtain estimated standard errors for variance components is not straightforward for multidimensional datasets, however (Brennan, 2001). For example, in a $p \times i$ (persons crossed with items) design it is possible to obtain bootstrap samples by sampling persons with replacement but not items; by sampling items with replacement but not persons; by sampling persons and items with replacement; or by sampling persons, items, and residuals with replacement (which may produce different results; see Brennan, 1994, 2001; see also Wiley, 2000, for guidelines on choosing among bootstrap strategies to obtain the most accurate estimates of standard errors). Furthermore, bootstrap procedures are likely to lead to biased estimates of variance components and standard errors. Wiley (2000) (summarized in Brennan, 2001, p. 188) derived adjustments for the estimates of variance components that eliminate the bias from different bootstrap procedures.

Although exact confidence intervals for variance components are generally unavailable (due to the inability to derive exact distributions for variance component estimates; see Searle, 1971), Satterthwaite (1941, 1946) and Ting et al. (1990) developed procedures for obtaining approximate confidence intervals (see also Burdick and Graybill, 1992). In simulations of $p \times i$ G-study estimated standard errors and confidence intervals with normal data, Brennan (2001) showed that confidence intervals based on the Satterthwaite, Ting et al., and jackknife procedures were fairly accurate, while the results using bootstrap procedures were mixed. Citing Burdick and Graybill (1992), Brennan (2001) noted theoretical reasons to prefer the Ting et al. procedure to the Satterthwaite procedure. For non-normal dichotomous data, the jackknife procedure produced confidence intervals that were more accurate than those obtained with the other methods (Brennan, 1994, 2001).

Negative estimated variance components

Negative estimates of variance components can arise because of sampling errors or because of model misspecification (Searle, 1971). Possible solutions when negative estimates are small in magnitude are to (1) substitute zero for the negative estimate and carry through the zero in other expected mean square equations from the analysis of variance (which produces biased estimates, Cronbach et al., 1972), (2) set any negative estimates of variance components to zero (Brennan, 2001), (3) use a Bayesian approach that sets a lower bound of zero on the estimated variance component (Box and Tiao,

1973; Fyans, 1977), and (4) use maximum likelihood (ML) or restricted maximum likelihood (REML) methods, which preclude negative estimates (Searle, 1971). Although the latter two approaches require a distributional form assumption, usually normality (Brennan, 1994), Marcoulides (1987, 1990) showed in simulations of data using a variety of crossed and nested, balanced and unbalanced designs that REML estimates were more accurate than ANOVA estimates for both non-normal (dichotomous, skewed, and j-shaped) and normal distributions.

Variance component estimation with unbalanced designs

An unbalanced design has unequal numbers of observations in its subclassifications. A nested design may be unbalanced purposively (e.g., achievement test items nested within content categories), dictated by the context itself (e.g., pupils nested within classrooms where class size is not constant), or by unforeseen circumstance (e.g., observers nested within occasions, with an unequal number of observers present at each occasion). Unbalanced situations can also arise in crossed and nested designs due to missing observations.

Estimating variance components in unbalanced designs is not straightforward. Some or all methods have problems of computational complexity, distributional assumptions, biased estimation, require decisions that cannot be justified in the context of generalizability theory, or produce results that are inconclusive (Brennan, 2001). While Henderson's Method 1 produces unbiased estimates for random effects models, two of Henderson's (1953) ANOVA-like methods produce biased estimates with models involving fixed effects (Method 1) or cannot be applied to models with interactions between random and fixed effects (Method 2). Henderson's Method 3 does not have these problems but produces different results depending upon the order in which variance components are estimated (Chiu and Wolfe, 2002), with no obvious basis for choosing among different orders. Maximum likelihood methods assume normality and are computationally complex. MINQUE (minimum norm quadratic unbiased estimation) does not assume normality and does not involve iterative estimation, thus reducing computational complexity, but estimates may be negative and are usually biased. Maximum likelihood procedures and MINQUE are often not viable, however, because sample sizes in G theory tend to be large, necessitating inverting matrices closely related to the number of observations in the data. Moreover, MINQUE requires the investigator to specify a priori weights corresponding to the relative sizes of the variance components (Brennan, 2001). The results may differ for different choices of a priori weights, without a compelling reason to choose among solutions (Searle et al., 1992).

Other approaches to estimating variance components with unbalanced designs involve analyzing one or more balanced subsets of the dataset, the major drawback being that the data analyzed may not be representative of the full dataset. To remedy this problem, Chiu and Wolfe (2002) proposed a method in which the dataset is exhaustively subdivided into all possible fully crossed subsets, into all possible nested subsets, and into all subsets with a modified balanced incomplete block design (*MBIB*, Chiu, 2001). The resulting variance component estimates from the different subsets are then weighted by the subset sample size and averaged. Chiu and Wolfe's (2002) analyses of data sets with two facets and large amounts of missing data (in which they assumed that data were missing at random) produced unbiased and consistent results.

3.6.2. Hidden facets

All studies have sources of variance that are not explicitly recognized or represented in the design (Cronbach et al., 1997; Shavelson et al., 1999; Webb et al., 2000). These are called implicit or hidden facets. In some cases, the same condition underlies all observations in the G-study, for example, when all persons are tested on a single occasion. Estimates of generalizability from such data do not take into account the possibility that persons' scores might be different on another occasion and, consequently, overestimate generalizability when interest lies in generalizing over occasions.

In other cases, a facet is linked or confounded with another source of variance (the object of measurement, another facet, or both) such that as the levels of one facet vary, correspondingly the levels of the other source of variance do too. As an example of a hidden facet confounded with the object of measurement, Brennan (2001) describes a piano recital in which students are evaluated by the same panel of judges but each student plays a different musical selection. Here, persons and musical selections are confounded, making it impossible to disentangle the effects of pianist ability and the difficulty of the particular musical selection. When interest lies in generalizing over musical selections, the design with musical selection as a hidden facet may underestimate or overestimate generalizability depending on the magnitudes of the (unknown) effects of musical selection.

As an example of a hidden facet that is confounded with another facet, consider again the notorious hidden facet, occasion. While a test administration might be conceived as a single occasion, an alternative conception of occasion is possible. As a person proceeds through a test or performs a series of tasks, his performance occurs over time. While task is varying, occasion of testing may be considered to vary too. Variability in performance from task to task typically is interpreted as task-sampling variability (Shavelson et al., 1993), but it may also be due to occasion-sampling variability, as pointed out by Cronbach et al. (1997). If occasion is the cause of variability in performance over tasks, then adding tasks to address the task-sampling problem will not improve the dependability of the measurement.

To examine the effects of occasion of testing, Shavelson et al. (1999) re-analyzed Shavelson et al.'s (1993) educational performance assessment data with occasions explicitly recognized as a source of variability in the design. Analysis of the design with persons crossed with tasks, raters, and occasions showed that both the task (person $\times$ task) and occasion (person $\times$ occasion) facets contributed variability, but the lion's share came from task sampling (person $\times$ task), and joint task and occasion sampling (person $\times$ task $\times$ occasion). The person $\times$ task $\times$ occasion interaction was the largest effect by far. In a different study, Webb et al. (2000) reported similar results. These studies showed that explicitly recognizing occasion as a facet of error variance altered the interpretation about the substantial sources of error in the measurement and called into question the interpretation of G studies in which the hidden occasion facet was ignored.

Occasion may be hidden in other respects as well. When a judge scores a particular examinee's performance once (as is usually the case), the occasion of scoring becomes an additional hidden facet (Cronbach et al., 1997), possibly leading to misinterpretation of effects involving judges.

3.6.3. Nonconstant error variance for different true scores

The descriptions of error variance given in previous sections implicitly assume that variance of measurement error is constant for all persons, regardless of universe score (true score in Classical Test Theory). The assumption of constant error variance for different universe scores was first criticized more than half a century ago in the context of Classical Test Theory, including by Lord (1955), who derived a formula for conditional error variance that varies as a function of true score. Lord's approach produced a concave-down quadratic form in which error variances are smaller for very high and very low true scores than for true scores that are closer to the mean. While persons with true scores close to the mean may produce scores that fluctuate from item to item, persons with true scores at the extremes have little opportunity for errors to influence their observed scores.

Brennan (1998) pointed out that Lord's conditional error variance is the conditional error variance for absolute decisions in G theory for a one-facet design with dichotomously-scored items and n_i' equal to n_i. For polytomously-scored items, estimated conditional error variance for absolute decisions for a one-facet $p \times I$ design (Brennan, 2001) is:

$$\hat{\sigma}^2_{\Delta_p} = \frac{\sum_i (X_{pi} - X_{pI})^2}{n_i'(n_i - 1)}. \tag{58}$$

An approximation for estimated conditional error variance for relative decisions for a $p \times I$ one-facet design based on large n_p (Brennan, 2001) is:

$$\hat{\sigma}^2_{\delta_p} = \hat{\sigma}^2_{\Delta_p} + \frac{\hat{\sigma}_i^2}{n_i'} - \frac{2\,\mathrm{cov}(X_{pi}, X_{Pi}|p)}{n_i'}, \tag{59}$$

where

$$\mathrm{cov}(X_{pi}, X_{Pi}|p) = \frac{\sum_i (X_{pi} - X_{pI})(X_{Pi} - X_{PI})}{n_i - 1}. \tag{60}$$

Following Brennan (2001), for a multi-facet random design (e.g., $p \times T \times O$ with persons crossed with tasks and observers), the conditional error variance for absolute decisions is:

$$\sigma^2_{\Delta_p} = \frac{\sigma^2_{t_p}}{n_t'} + \frac{\sigma^2_{o_p}}{n_o'} + \frac{\sigma^2_{to_p}}{n_t' n_o'}, \tag{61}$$

where $\sigma^2_{t_p}$, $\sigma^2_{o_p}$, and $\sigma^2_{to_p}$ a re-estimated using the $T \times O$ data for person p only. Brennan (1996, 1998) gives complicated formulas for conditional error variances for relative decisions for multifacet random designs. A simplified estimator that assumes very large n_p (Brennan, 2001) is:

$$\hat{\sigma}^2_{\delta_p} = \hat{\sigma}^2_{\Delta_p} + \left(\hat{\sigma}^2_{\Delta} - \hat{\sigma}^2_{\delta}\right). \tag{62}$$

Brennan (2001) also shows procedures for estimating condition relative and absolute error variance for multivariate G studies and for unbalanced designs.

Applying (62) to the generalizability study of job performance measurements of Navy machinist mates (Example 1) for a D-study with 11 tasks and 2 observers gives

$\hat{\sigma}^2_{\Delta_p} = 0.00027$ for a machinist mate with a high observed score (0.96 out of a maximum possible of 1.00) and $\hat{\sigma}^2_{\Delta_p} = 0.00731$ for a machinist mate with an observed score close to the mean (0.75). (As pointed out by Brennan, 2001, estimates of conditional error variances are usually expressed with respect to $\bar{X}_p$ because universe scores are unknown.) These values can be compared to the overall $\hat{\sigma}^2_{\Delta} = 0.00330$ in Table 8. These results show considerable variability in estimated error variance for persons with different estimated universe scores.

3.6.4. Optimizing the decision (D) study design

As shown in previous sections, when designing a decision study that will yield a reasonable level of estimated generalizability, the decision maker may consider both the design of the D-study and the number of observations to be used for each facet in the design. The resources available for the D-study may not accommodate the number of observations needed for a desired level of estimated generalizability, however. In this case, the decision maker may want to know the number of observations per facet that will minimize error (and maximize generalizability) for a fixed number of observations per person (Woodward and Joe, **?**) or for a given set of cost constraints. Marcoulides and Goldstein (1991, 1992), Marcoulides (1993, 1995) (see also Sanders, 1992; Sanders et al., 1991) describe methods for minimizing error variance in decision studies under cost constraints, both total costs and per-facet costs, for univariate and multivariate designs.

In the case of a fixed total number of observations for a multivariate profile or composite in which the observations for each score in the profile are not the same, Brennan (2001, p. 312) describes procedures for allocating observations to minimize error variance of the composite. In Example 3, the multivariate generalizability study of teacher behavior, suppose that the D-study will have six raters who will rate teacher behavior in either reading or mathematics (but not both). If a composite score is desired with weights 0.67 and 0.33 (reading weighted twice as heavily as mathematics), applying Brennan's procedure yields the smallest estimated absolute error variance of the composite ($\hat{\sigma}^2_{\Delta} = 0.27113$) and highest level of dependability ($\hat{\Phi} = 0.885$) for $n'_r = 5$ in reading and $n'_r = 1$ in mathematics.

3.7. Linking generalizability theory and item response theory

Item response theory (IRT), also known as latent trait theory, is a widely used approach to test development that focuses on estimating a person's ability (or other latent trait) and describing properties of items on the basis of assumed mathematical relationships between the latent trait and item responses (Embertson and Reise, 2000; Hambleton et al., 1991; Lord, 1980). A central assumption of IRT is that an individual's probability of responding to an item in a certain way is a function only of the individual's ability. Generally speaking, estimates of a person's ability level do not depend on the group of persons being examined, the particular items used, or other conditions of measurement, such as the occasions of testing or the raters used to judge performance; moreover, estimates of item (or occasion or rater) attributes do not depend on the ability distribution of the examinees. The property of invariance of ability and item (or judge or occasion)

parameters makes IRT well suited for addressing issues related to test or scale construction, score equating, computer adaptive testing, and item bias. IRTs focus on items (or other conditions of measurement) as *fixed* entities (Brennan, 2001), however, makes it impossible to estimate sources of error variation that can be used to optimize the design of a future measurement using a different sample of items, raters, and occasions. Hence, Brennan (2001, p. 166) characterizes IRT as principally a scaling model while G theory is truly a measurement model.

Kolen and Harris (1987) and Briggs and Wilson (**?**) proposed an approach that brings together the sampling model of G theory with the scaling model of IRT. For example, in a person $\times$ item design in which persons and items are assumed to be randomly drawn from a population and universe of interest, respectively, specifying prior distribution functions for person and item effects makes it possible to define and estimate IRT analogues to variance components used in G theory as well as estimate traditional IRT parameters. Expected, rather than observed, item responses are used to estimate variance components.

Bock et al. (2002) described how to link G theory and IRT in situations where the IRT assumption of local independence – holding constant the abilities influencing test performance, persons' responses to any pair of items are statistically independent – is violated. In the partially nested $p \times (r : i)$ design that is common in large-scale performance assessment, for example, in which different sets of raters evaluate the items, the multiple ratings of each item do not represent distinct items and are not locally independent. Bock et al. (2002) note that treating multiple ratings as if they were separate items in an IRT analysis would produce underestimated standard errors of IRT scale-score and item parameter estimates; they suggest corrections for this downward bias using G theory results.

A number of other studies have used G theory and IRT approaches sequentially, first carrying out generalizability analyses to identify important sources of variation and then applying IRT techniques to diagnose aberrant persons or conditions or combinations of them to help guide test revision and rater training. Bachman et al. (1995), for example, applied generalizability theory and an extension of the Rasch model known as many-facet Rasch measurement (also called multi-faceted Rasch measurement or MFRM) to examine variability in tasks and rater judgments of examinees' foreign language speaking ability (for similar sequences of analyses, see also Lynch and McNamara, 1998; Smith and Kulikowich, 2004; Stahl and Lunz, **?**). Bachman et al.'s generalizability analysis showed a non-negligible examinee $\times$ rater interaction, indicating that raters were not consistent in their rank orderings of examinees. The fit statistics from the many-facet Rasch analysis (e.g., infit, outfit) were used to identify individual raters who were not internally self-consistent in their ratings, that is, who gave unexpectedly high or low ratings compared to their usual rating style, and "bias" statistics (using the FACETS program, Linacre and Wright, 1993), which pointed to specific examinee-rater combinations with scores that deviated from expectation. Consistent with other application studies, Bachman et al. (1995) described the value of diagnostic information from MFRM as feedback to raters to improve their future ratings. Pinpointing discrepant behavior of raters in a G-study may also have ramifications for designing effective training for potential raters (not just those used in the original study), which may reduce variabil-

ity among future raters, and so reduce the number needed for dependable measurement in a D-study.

In another sequential application of the two theories, Marcoulides and Drezner (2000) developed an alternative model for estimating latent traits such as examinee ability, rater severity, and item difficulty. This model uses a weighted Euclidean distance function to provide both latent trait estimates and two-dimensional graphical representations of conditions or combinations of conditions for use in identifying unusually performing persons, raters, or items (or combinations of them). Applying the model to the teacher behavior data in Example 2, Marcoulides (1999) provided detailed insights into the behavior of raters, showing clearly that raters were more consistent in their ratings of teachers in mathematics than in reading. Marcoulides (1999) also showed that the latent trait estimates produced using the Marcoulides–Drezner model are not sensitive to unusually performing persons or conditions, in contrast to estimation procedures based on the many-facet Rasch model, which makes the assumption of no interactions.

3.8. Computer programs

A number of popular computer packages and programs provide estimates of variance components in generalizability studies. These include SAS (SAS Institute, 1996), SPSS (SPSS, 1997), and S-Plus (MathSoft, 1997). However, only one such program has been developed specifically for generalizability theory: GENOVA (GENeralized analysis Of Variance; Brennan, 2001; Crick and Brennan, 1983). GENOVA handles complete balanced designs (with up to five facets), urGENOVA handles designs that are unbalanced due to nesting and some designs with missing data, and mGENOVA performs multivariate generalizability and decision analyses for a selected set of designs that are balanced or unbalanced due to nesting.

The programs listed above use a variety of methods for estimating variance components, including the ANOVA procedure using expected mean square equations (EMS, Cornfield and Tukey, 1956), Henderson's Method 1 and Method 3 (Henderson, 1953), MINQUE (minimum norm quadratic unbiased estimation) with equal weighting of all variance components (MINQUE(0)) or equal weights for all variance components except the residual (MINQUE(1)), maximum likelihood, and restricted maximum likelihood. GENOVA also provides estimates using Cronbach et al.'s (1972) recommendation to substitute zero for negative estimated variance components wherever they appear in the EMS equations. As noted by Brennan (2001), processing times and memory requirements vary widely, depending on the complexity of the design, the numbers of observations (which can easily exceed 10,000), the estimation method used, and the computational algorithm employed.

4. Concluding remarks

Generalizability theory has sometimes been mischaracterized as "merely" an application of random-effects ANOVA models to item response data. This characterization is unfortunate and misleading (Brennan, 2000). In addition to computational procedures,

G theory provides a conceptual framework and notational system for characterizing alternative intended universes of generalization, distinguishing fixed versus random facets, and designing efficient measurement procedures, among other uses. G theory has also clarified the use and interpretation of coefficients and estimated standard errors from classical test theory. This is of value because, despite the greater sophistication of G theory, and the important applications described in this chapter, classical test theory remains the dominant tool for estimating reliability.

In closing, it bears repeating that reliability is just one consideration in the responsible use and interpretation of tests, and not the most important consideration, at that. The meaning, appropriateness, and consequences of testing must be weighed in any practical situation, and the most reliable measurement procedure may not be the best. That said, it is a commonplace that reliability is prerequisite to validity, and the statistical methods described in this chapter remain critically important for sound measurement practice.

Uncited references

(Briggs, 2004) (de Finetti, 1937) (Linacre and Wright, 2002) (Searle, 1987) (Shavelson et al., 1991) (Short et al., 1986) (Stahl and Lunz, 1992)

References

Bachman, L.F., Lynch, B.K., Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing* **12**, 238–257.

Bock, R.D., Brennan, R.L., Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement* **26**, 364–375.

Box, G.E.P., Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA.

Brennan, R.L. (1994). Variance components in generalizability theory. In: Reynolds, C.R. (Ed.), *Cognitive Assessment: A Multidisciplinary Perspective*. Plenum Press, New York, pp. 175–207.

Brennan, R.L. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement* **32**, 385–396.

Brennan R.L. (1996). Conditional standard errors of measurement in generalizability theory. Iowa Testing Programs Occasional Paper No. 40. Iowa Testing Programs, University of Iowa, Iowa City, IA.

Brennan, R.L. (1997). A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice* **16**, 14–20.

Brennan, R.L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement* **22**, 307–331.

Brennan, R.L. (2000). (Mis)conceptions about generalizability theory. *Educational Measurement: Issues and Practice* **19**, 5–10.

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

Briggs, D.C. (2004). Generalizability in item response modeling. Paper presented at the 2004 Annual Meeting of the Psychometric Society, Pacific Grove, CA.

Burdick, R.K., Graybill, F.A. (1992). *Confidence Intervals on Variance Components*. Dekker, New York, NY.

Cardinet, J., Tourneur, Y., Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement* **13**, 119–135.

Cardinet, J., Tourneur, Y., Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement* **18**, 183–204.

Chiu, C.W.T., Wolfe, E.W. (2002). A method for analyzing sparse data matrices in the generalizability theory framework. *Applied Psychological Measurement* **26**, 321–338.

Chiu, C.W.T. (2001). *Generalizability Theory: Scoring Large-Scale Assessments based on Judgments*. Kluwer Academic, Boston, MA.

Cornfield, J., Tukey, J.W. (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics* **27**, 907–949.

Crick, J.E., Brennan, R.L. (1983). GENOVA: A generalized analysis of variance system [Computer software and manual]. University of Iowa, Iowa City, IA. Available from http://www.education.uiowa.edu/casma/.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* **16**, 297–334.

Cronbach, L.J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement* **64**, 391–418.

Cronbach, L.J., Rajaratnman, N., Gleser, G.C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology* **16**, 137–163.

Cronbach, L.J., Gleser, G.C., Nanda, H., Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements*. Wiley, New York.

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

de Finetti, B.D. (1937). Foresight: Its logical laws, its subjective sources. *Annales de l'Institut Henri Poincare* **7**, 95–158.

de Finetti, B.D. (1964). Foresight: Its logical laws, its subjective sources. In: Kyburg, H.E., Smokler, G.E. (Eds.), *Studies in Subjective Probability*. Wiley, New York, NY, pp. 95–158.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. SIAM, Philadelphia, PA.

Efron, B., Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* **1**, 54–77.

Embertson, S.E., Reise, S.P. (2000). *Item Response Theory as Model-Based Measurement*. In: *Item Response Theory for Psychologists*. Erlbaum, Mahwah, NJ.

Feldt, L.S., Brennan, R.L. (1989). Reliability. In: Linn, R.L. (Ed.), *Educational measurement*, 3rd ed. The American Council on Education, Macmillan, pp. 105–146.

Feller, W. (1966). *An Introduction to Probability Theory and its Applications*. Wiley, New York, NY.

Fyans Jr., L.J. (1977). A new multiple level approach to cross-cultural psychological research. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Goodenough, F.L. (1936). A critical note on the use of the term 'reliability' in mental measurement. *Journal of Educational Psychology* **27**, 173–178.

Haertel, E.H. (in press). Reliability. In: Brennan, R.L. (Ed.), *Educational Measurement*, 4th ed. Greenwood, Westport, CT.

Hambleton, R.K., Swaminathan, H., Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Sage Publications, Newbury Park, CA.

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics* **9**, 227–252.

Joe, G.W., Woodward, J.A. (1976). Some developments in multivariate generalizability. *Psychometrika* **41**, 205–217.

Kane, M.T., Brennan, R.L. (1977). The generalizability of class means. *Review of Educational Research* **47**, 267–292.

Kolen, M.J., Harris, D.J. (1987). A multivariate test theory model based on item response theory and generalizability theory. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

Linacre, J.M., Wright, B.D. (1993). *A User's Guide to FACETS: Rasch Model Computer Program, Version 2.4 for PC-Compatible Computers*. MESA Press, Chicago, IL.

Linacre, J.M., Wright, B.D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement* **3**, 486–512.

Lord, F.M. (1955). Estimating test reliability. *Educational and Psychological Measurement* **16**, 325–336.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum, Hillsdale, NJ.

Lynch, B.K., McNamara, T.F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing* **15**, 158–180.

Marcoulides, G.A., Drezner, Z. (2000). A procedure for detecting pattern clustering in measurement designs. In: Wilson, M., Engelhard, G. (Eds.), In: *Objective Measurement: Theory into Practice*, vol. 5. Ablex, Stamford, CT, pp. 287–303.

Marcoulides, G.A., Goldstein, Z. (1991). Selecting the number of observations in multivariate measurement studies under budget constraints. *Educational and Psychological Measurement* **51**, 573–584.

Marcoulides, G.A., Goldstein, Z. (1992). The optimization of multivariate generalizability studies with budget constraints. *Educational and Psychological Measurement* **52**, 301–309.

Marcoulides, G.A. (1987). An alternative method for variance component estimation: Applications to generalizability theory. Unpublished doctoral dissertation, University of California, Los Angeles.

Marcoulides, G.A. (1990). An alternate method for estimating variance components in generalizability theory. *Psychological Reports* **66**, 379–386.

Marcoulides, G.A. (1993). Maximizing power in generalizability studies under budget constraints. *Journal of Educational Statistics* **18**, 197–206.

Marcoulides, G.A. (1994). Selecting weighting schemes in multivariate generalizability studies. *Educational and Psychological Measurement* **54**, 3–7.

Marcoulides, G.A. (1995). Designing measurement studies under budget constraints: Controlling error of measurement and power. *Educational and Psychological Measurement* **55**, 423–428.

Marcoulides, G.A. (1999). Generalizability theory: Picking up where the Rasch IRT model leaves off? In: Embretson, S.E., Hershberger, S.L. (Eds.), *The New Rules of Measurement: What Every Psychologist Should Know*. Erlbaum, Mahwah, NJ, pp. 129–152.

MathSoft, Inc. (1997). *S-Plus 4.5 Standard Edition*. Author, Cambridge, MA.

Sanders, P.F. (1992). Alternative solutions for optimizing problems in generalizability theory. *Psychometrika* **57**, 351–356.

Sanders, P.F., Theunissen, T.J.J.M., Baas, S.M. (1991). Maximizing the coefficient of generalizability under the constraint of limited resources. *Psychometrika* **56**, 87–96.

SAS Institute, Inc. (1996). *The SAS System for Windows Release 6.12*. Author, Cary, NC.

Satterthwaite, F.E. (1941). Synthesis of variance. *Psychometrika* **6**, 309–316.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics* **2**, 110–114.

Searle, S.R., Casella, G., McCulloch, C.E. (1992). *Variance Components*. Wiley, New York, NY.

Searle, S.R. (1971). *Linear Models*. Wiley, New York, NY.

Searle, S.R. (1987). *Linear Models for Unbalanced Data*. Wiley, New York, NY.

Shao, J., Tu, D. (1995). *The Jackknife and the Bootstrap*. Springer-Verlag, New York, NY.

Shavelson, R.J. (2004). Editor's preface to Lee J. Cronbach's "My current thoughts on coefficient alpha and successor procedures". *Educational and Psychological Measurement* **64**, 389–390.

Shavelson, R.J., Webb, N.M. (1981). Generalizability Theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology* **34**, 133–165.

Shavelson, R.J., Webb, N.M. (1991). *Generalizability Theory: A Primer*. Sage Publications, Newbury Park, CA.

Shavelson, R.J., Baxter, G.P., Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement* **30**, 215–232.

Shavelson, R.J., Baxter, G.P., Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education* **4**, 347–362.

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

Short, L.M., Shavelson, R.J., Webb, M.N. (1986). Issues in multivariate generalizability: Weighting schemes and dimensionality. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Smith, E.V., Kulikowich, J.M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement* **64**, 617–639.

Smith, P.L. (1978). Sampling errors of variance components in small sample multifacet generalizability studies. *Journal of Educational Statistics* **3**, 319–346.

Smith, P.L. (1981). Gaining accuracy in generalizability theory: Using multiple designs. *Journal of Educational Measurement* **18**, 147–154.

SPSS, Inc. (1997). *SPlanet. Space Sci. for Windows Release 8.0.0*. Author, Chicago, IL.

Stahl, J.A., Lunz, M.E. (1992). A comparison of generalizability theory and multi-faceted Rasch measurement. Paper presented at the Midwest Objective Measurement Seminar, University of Chicago.

Thorndike, R.L. (1947). Research Problems and Techniques. Report No. 3. AAF Aviation Psychology Program Research Reports. U.S. Government Printing Office, Washington, DC.

Ting, N., Burdick, R.K., Graybill, F.A., Jeyaratnam, S., Lu, T.C. (1990). Confidence intervals on linear combinations of variance components that are unrestricted in sign. *Journal of Statistical Computational Simulation* **35**, 135–143.

Tukey, J.W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics* **29**, 614.

Webb, N.M., Schlackman, J., Sugrue, B. (2000). The dependability and interchangeability of assessment methods in science. *Applied Measurement in Education* **13**, 277–301.

Webb, N.M., Shavelson, R.J., Kim, K.S., Chen, Z. (1989). Reliability (generalizability) of job performance measurements: Navy machinists mates. *Military Psychology* **1**, 91–110.

Wiley, E. (2000). Bootstrap strategies for variance component estimation: Theoretical and empirical results. Unpublished doctoral dissertation, Stanford University.

Wood, R. (1976). Trait measurement and item banks. In: DeGruijter, D.N.M., van der Kamp, L.J.T. (Eds.), *Advances in Psychological and Educational Measurement*. Wiley, New York, NY, pp. 247–263.