Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 1

Coefficients and Indices in Generalizability Theory*

Robert L. Brennan†

August 2003

*A revised version of a paper presented at the Annual Meeting of the American Statistical Association, San Francisco, August 2003. This paper borrows heavily from Brennan (2001b).

†Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Director, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: robert-brennan@uiowa.edu).

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

Contents

Abstract

Generalizability theory offers an extensive conceptual framework and a powerful set of statistical procedures for addressing numerous measurement issues. Although the statistical aspects of generalizability theory are undeniably important, perhaps the most distinguishing feature of the theory is its conceptual framework, which permits a multifaceted perspective on measurement error and its components. The primary purpose of this paper is to provide discussions and interpretations of the various coefficients and indices that have been proposed for use in generalizability theory, including generalizability coefficients, dependability coefficients, signal-noise ratios, error-tolerance ratios, and various multivariate indices. These coefficients and indices can be very useful if they are interpreted correctly, but almost always they need to be buttressed with additional information—especially information about error variances.

Generalizability theory provides an extensive conceptual framework and set of statistical machinery for quantifying and explaining the consistencies and inconsistencies in observed scores for objects of measurement. To an extent, the theory can be viewed as an extension of classical test theory (see, for example, Feldt & Brennan, 1989, and Lord & Novick, 1969) through the application of certain analysis of variance (ANOVA) procedures. Classical theory postulates that an observed score can be decomposed into a true score and a single undifferentiated random error term. By contrast, generalizability theory substitutes the notion of universe score for true score, and liberalizes classical theory by employing ANOVA methods that allow an investigator to disentangle the multiple sources of error that contribute to the undifferentiated error in classical theory.

There is some truth to the assertion that generalizability theory is the application of ANOVA to measurement problems, but this assertion is perhaps more misleading than informative. On a superficial level it is misleading in that generalizability theory pays no attention to hypothesis testing; rather, generalizability theory focuses on the use and estimation of variance components. On a deeper level the assertion overlooks the conceptual framework of generalizability theory, which is of central importance, as discussed next.

Perhaps the most important and unique feature of generalizability theory is its conceptual framework, which focuses on certain types of studies and universes. A G (*generalizability*) study involves the collection of a sample of data from a *universe of admissible observations* that consists of *facets* defined by an investigator. Typically, the principal result of a G study is a set of estimated random effects variance components for the universe of admissible observations. A D (*decision*) study provides estimates of universe score variance, error variances, certain indexes that are like reliability coefficients, and other statistics for a measurement procedure associated with an investigator-specified *universe of generalization*.

In univariate generalizability theory, there is only one universe (of generalization) score for each object of measurement. In multivariate generalizability theory, each object of measurement has multiple universe scores. Univariate generalizability theory typically employs *random-effects models* primarily . Although *mixed models* can be accommodated in univariate generalizability theory and will be discussed in this paper, strictly speaking mixed models are more closely associated with multivariate generalizability theory.

The initial, defining treatment of generalizability theory was provided over 30 years ago by Cronbach, Gleser, Nanda, and Rajaratnam (1972) in a book entitled *The dependability of behavioral measurements*. Brennan (2001b) provides a recent extended treatment of the theory; Shavelson and Webb (1991) provide a brief monograph that describes the basics of the theory. The essential features of univariate generalizability theory were largely completed with technical reports in 1960–1961 that were revised into three journal articles, each with a different first author (Cronbach, Gleser, and Rajaratnam). Multivariate generalizability theory was developed during the ensuing decade. Research between 1920 and 1955 by Ronald A. Fisher, Cyril Burt, Cyril J. Hoyt, Robert L. Ebel, and E. F.

Lindquist, among others, influenced the development of generalizability theory. (See Brennan, 1997, for a detailed discussion of the history of generalizability theory.)

The primary purpose of this paper is to provide discussions and interpretations of the various coefficients and indices that have been proposed for use in generalizability theory, including generalizability coefficients, dependability coefficients, signal-noise ratios, error-tolerance ratios, and various multivariate indices. These coefficients and indices can be very useful if they are interpreted correctly, but almost always they need to be buttressed with additional information—especially information about error variances.

It is assumed that readers of this paper have some familiarity with ANOVA, and knowledge of classical test theory is also useful. However, it is not assumed that most readers are familiar with generalizability theory. Therefore, the first section provides a simple hypothetical scenario that hints at issues that are discussed more extensively in subsequent sections. The second section treats so-called "single-facet" designs in which measurements (a facet) are obtained for objects of measurement (a population). (In the terminology of generalizability theory, the objects of measurement, which are usually persons, are not called a facet, and the objects of measurement are almost always assumed to random.) The third section treats two-facet designs in which the model is either random (i.e., both facets are random) or mixed (one facet is fixed). The fourth section provides a brief overview of coefficients and indices in multivariate generalizability theory. Only in the second section on single-facet designs is a serious attempt made to derive results under clearly specified assumptions. Elsewhere, emphasis is placed on informing the reader about various procedures for obtaining and interpreting results in generalizability theory, without detailed derivations (see Brennan, 2001b, for a deeper discussion.)

In most of this paper, the context-specific terminology that is used is largely the terminology employed in educational measurement. The issues, however, apply to any field of scientific endeavor. See Brennan (2001b, pp. 17–18) for a list of references that have employed generalizability theory in numerous fields.

1 Hypothetical Scenario: An Introduction

Suppose an investigator, Smith, wants to construct a measurement procedure for evaluating writing proficiency. First, Smith might identify/characterize essay prompts of interest to her, as well as potential raters. In doing so, Smith effectively specifies the facets in her universe of admissible observations. Assume she views these facets as infinite such that, in theory, any person p (i.e., object of measurement) might respond to any essay prompt t, which in turn might be evaluated by any rater r. If so, a likely G study (i.e., a data collection design for estimating variance components in the universe of admissible observations) might be $p \times t \times r$, where "$\times$" is read "crossed with." The associated linear model is

$$X_{ptr} = \mu + \nu_p + \nu_t + \nu_r + \nu_{pt} + \nu_{pr} + \nu_{tr} + \nu_{ptr},$$

where the ν are uncorrelated score effects. This model leads to

$$\sigma^2(X_{ptr}) = \sigma^2(p) + \sigma^2(t) + \sigma^2(r) + \sigma^2(pt) + \sigma^2(pr) + \sigma^2(tr) + \sigma^2(ptr),$$

which is a decomposition of the total observed score variance into seven variance components that are usually estimated using the expected mean squares in a random-effects ANOVA.

Suppose the following estimated variance components are obtained from Smith's G study based on n_t essay prompts and n_r raters:

$$\begin{aligned} \hat{\sigma}^2(p) = .25, \quad \hat{\sigma}^2(t) = .06, \quad \hat{\sigma}^2(r) = .02, \\ \hat{\sigma}^2(pt) = .15, \quad \hat{\sigma}^2(pr) = .04, \quad \hat{\sigma}^2(tr) = .00, \\ \text{and} \quad \hat{\sigma}^2(ptr) = .12. \end{aligned}$$

Suppose also that Smith wants to generalize persons' observed mean scores based on n'_t essay prompts and n'_r raters to persons' scores for a universe of generalization that involves an infinite number of prompts and raters. This is a verbal description of a D study $p \times T \times R$ random effects design. It is much like the $p \times t \times r$ design for Smith's G study, but the sample sizes for the D study need not be the same as the sample sizes for the G study, and the $p \times T \times R$ design focuses on *mean* scores for persons. The expected score for a person over the facets in the universe of generalization is called the person's universe score.

If $n'_t = 3$ and $n'_r = 2$ for Smith's measurement procedure, then the D study estimated random effects variance components are:

$$\begin{aligned} \hat{\sigma}^2(p) = .25, \quad \hat{\sigma}^2(T) = \frac{\hat{\sigma}^2(t)}{n'_t} = .02, \quad \hat{\sigma}^2(R) = \frac{\hat{\sigma}^2(r)}{n'_r} = .01, \\ \hat{\sigma}^2(pT) = \frac{\hat{\sigma}^2(pt)}{n'_t} = .05, \quad \hat{\sigma}^2(pR) = \frac{\hat{\sigma}^2(pr)}{n'_r} = .02, \quad \hat{\sigma}^2(TR) = \frac{\hat{\sigma}^2(tr)}{n'_t n'_r} = .00, \\ \text{and} \quad \hat{\sigma}^2(pTR) = \frac{\hat{\sigma}^2(ptr)}{n'_t n'_r} = .02. \end{aligned}$$

Universe score variance is denoted generically as $\sigma^2(\tau)$, and for the random model

$$\sigma^2(\tau) = \sigma^2(p).$$

For this example, therefore, $\hat{\sigma}^2(p) = .25$ is an estimate of the universe score variance. The other variance components contribute to different types of error variance.

Absolute error, Δ_p, is simply the difference between a person's observed mean score and the person's universe score. For Smith's design and universe, the variance of these errors is

$$\sigma^2(\Delta) = \sigma^2(T) + \sigma^2(R) + \sigma^2(pT) + \sigma^2(pR) + \sigma^2(TR) + \sigma^2(pTR),$$

which gives $\hat{\sigma}^2(\Delta) = .12$. Its square root is $\hat{\sigma}(\Delta) = .35$, which is interpretable as an estimate of the "absolute" standard error of measurement for a randomly selected person.

Relative error, δ_p, is defined as the difference between a person's observed deviation score and his or her universe deviation score. For Smith's design and universe,

$$\sigma^2(\delta) = \sigma^2(pT) + \sigma^2(pR) + \sigma^2(pTR),$$

which gives $\hat{\sigma}^2(\delta) = .09$. Its square root is $\hat{\sigma}(\delta) = .30$, which is interpretable as an estimate of the "relative" standard error of measurement for a randomly selected person.

Two types of reliability-like coefficients are widely used in generalizability theory. One coefficient is called a "generalizability coefficient," which is defined as

$$\boldsymbol{E}\rho^2 = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)}. \tag{1}$$

It is the analogue of a reliability coefficient in classical test theory. The other coefficient is called an "index of dependability," which is defined as

$$\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)}. \tag{2}$$

Since $\sigma^2(\delta) < \sigma^2(\Delta)$, it necessarily follows that $\boldsymbol{E}\rho^2 > \Phi$. For example, using Smith's data, $\boldsymbol{E}\hat{\rho}^2 = .25/(.25 + .09) = .74$, and $\widehat{\Phi} = .25/(.25 + .12) = .68$. Both of these coefficients have interpretations in terms of agreement. Other related indices (discussed later) include signal-noise ratios and error-tolerance ratios.

Suppose Smith decides that she wants to treat essay prompts as fixed. In this case, Smith's universe of generalization consists of only one random facet (raters, R), and it can be shown that universe score variance is

$$\sigma^2(\tau) = \sigma^2(p) + \sigma^2(pT),$$

absolute error variance is

$$\sigma^2(\delta) = \sigma^2(R) + \sigma^2(pR) + \sigma^2(TR) + \sigma^2(pTR),$$

and relative error variance is

$$\sigma^2(\Delta) = \sigma^2(pR) + \sigma^2(pTR).$$

Equations 1 and 2 for $\boldsymbol{E}\rho^2$ and Φ, respectively, still apply; what changes is which variance components contribute to universe score variance and the two error variances. Obviously, with prompts being fixed, $\sigma^2(\tau)$ increases, both $\sigma^2(\Delta)$ and $\sigma^2(\delta)$ decrease, and both $\boldsymbol{E}\rho^2$ and Φ increase. For the illustrative data, it is easy to verify that $\hat{\sigma}^2(\tau) = .30$, $\hat{\sigma}^2(\Delta) = .05$, $\hat{\sigma}^2(\delta) = .04$, $\boldsymbol{E}\hat{\rho}^2 = .88$, and $\widehat{\Phi} = .86$. In other words, fixing a facet narrows the universe of generalization, which makes it easier (i.e., less error-prone) to draw inferences from observed scores to universe scores. Or, stated differently, fixing a facet leads to greater similarity between observed and universe scores.

2 Single-Facet Designs

For single-facet designs, the universe of admissible observations and the universe of generalization involve conditions from the same single facet, which will be called the item facet, here. It will be denoted i or I depending on whether reference is being made to the G study (and universe of admissible observations) or the D study (and universe of generalization), respectively. Also, we will refer to the population of objects of measurement as persons, denoted p. The use of the words "items" to characterize the facet and "persons" to characterize the objects of measurement is merely a convention. The facet could be any similar set of "conditions" of measurement (e.g., raters, forms, occasions, etc.), and the objects of measurement need not be persons.

For a single-faceted universe, there are two designs that might be employed in a G study: the $p \times i$ or the $i{:}p$ design, where the letter p is used to index persons (or examinees), i indexes items, "$\times$" is read "crossed with," and ":" is read "nested within." For the $p \times i$ design, each person is administered the *same* sample of items. For the $i{:}p$ design, each person is administered a *different* sample of items. Similarly, there are two possible D study designs: $p \times I$ and $I{:}p$, where uppercase I is used to emphasize that D study considerations involve mean scores over sets of items. In this section, we focus almost exclusively on crossed designs, only. (See Brennan, 2002, chap. 2 for a treatment of nested designs.)

Technically, it is assumed that persons and items are randomly sampled from an infinite population of persons and an infinite universe of items, respectively. In practice, this assumption is seldom if ever literally true, but it is a useful idealization in most circumstances.

2.1 G Study $p \times i$ Design

Let X_{pi} denote the score for any person in the population on any item in the universe. The expected value of a person's observed score, associated with a process in which an item is randomly selected from the universe, is

$$\mu_p \equiv \underset{i}{\boldsymbol{E}} X_{pi}, \tag{3}$$

where the symbol $\boldsymbol{E}$ is an expectation operator, and the subscript i designates the facet over which the expectation is taken. The score μ_p can be conceptualized as an examinee's "average" or "mean" score over the *universe* of items. In a similar manner, the population mean for item i is defined as

$$\mu_i \equiv \underset{p}{\boldsymbol{E}} X_{pi}, \tag{4}$$

and the mean over both the population and the universe is

$$\mu \equiv \underset{p}{\boldsymbol{E}} \underset{i}{\boldsymbol{E}} X_{pi}. \tag{5}$$

These mean scores are not observable, but any person-item score that *might* be observed (an observable score) can be expressed in terms of them using the linear model:

$$\begin{aligned} X_{pi} &= \mu && \text{(grand mean)} \\ &\quad + \mu_p - \mu && \text{(person effect} = \nu_p) \\ &\quad + \mu_i - \mu && \text{(item effect} = \nu_i) \\ &\quad + X_{pi} - \mu_p - \mu_i + \mu && \text{(residual effect} = \nu_{pi}), \end{aligned} \tag{6}$$

which can be expressed more succinctly as

$$X_{pi} = \mu + \nu_p + \nu_i + \nu_{pi}. \tag{7}$$

In conventional statistical terminology, the design given by Equation 7 might be called a two-way crossed design with a single observation in each cell, although this terminology is seldom used in generalizability theory.

All of the effects (except μ) in Equations 6 and 7 are called *random effects* because they are associated with a process of random sampling from the population and universe. It is important to note that in generalizability theory, these random effects are *defined* in terms of mean scores for the population and universe as indicated in Equation 6.[1] It follows that

$$\underset{p}{\boldsymbol{E}}\, \nu_p = \underset{i}{\boldsymbol{E}}\, \nu_i = \underset{p}{\boldsymbol{E}}\, \nu_{pi} = \underset{i}{\boldsymbol{E}}\, \nu_{pi} = 0.$$

It is "assumed" that all effects in the model are uncorrelated. Letting a prime designate a different person or item, this means that

$$\boldsymbol{E}(\nu_p \nu_{p'}) = \boldsymbol{E}(\nu_i \nu_{i'}) = \boldsymbol{E}(\nu_{pi} \nu_{p'i}) = \boldsymbol{E}(\nu_{pi} \nu_{pi'}) = \boldsymbol{E}(\nu_{pi} \nu_{p'i'}) = 0,$$

and

$$\boldsymbol{E}(\nu_p \nu_i) = \boldsymbol{E}(\nu_p \nu_{pi}) = \boldsymbol{E}(\nu_i \nu_{pi}) = 0.$$

The word "assumed" is in quotes because most of these zero expectations are a direct consequence of the manner in which score effects have been defined in the linear model and/or the random sampling assumptions for the $p \times i$ design. For example, it can be shown that the random sampling assumptions imply that $\boldsymbol{E}(\nu_p \nu_i) = (\boldsymbol{E}\nu_p)(\boldsymbol{E}\nu_i) = 0$.

The above development can be summarized by saying that a G study $p \times i$ design is represented by the random effects linear model in Equation 7 with uncorrelated score effects. This description is adequate *provided* it is understood in the sense discussed above. Note, also, that the modeling has been specified *without* any normality assumptions, and without assuming that score effects are independent, which is a stronger assumption than uncorrelated score effects.

[1] For more complicated designs with mixed models, this definitional issue is particularly crucial.

2.2 G Study Variance Components for $p \times i$ Design

For each score effect, or component, in Equation 7, there is an associated variance of the score effect, or component, which is called a *variance component*. Specifically,

$$\sigma^2(p) = \underset{p}{\boldsymbol{E}}\, \nu_p^2, \quad \sigma^2(i) = \underset{i}{\boldsymbol{E}}\, \nu_i^2, \quad \textit{and} \quad \sigma^2(pi) = \underset{p}{\boldsymbol{E}}\, \underset{i}{\boldsymbol{E}}\, \nu_{pi}^2.$$

These variance components provide a decomposition of the so-called "total" variance:

$$\sigma^2(X_{pi}) = \underset{p}{\boldsymbol{E}}\, \underset{i}{\boldsymbol{E}} (X_{pi} - \mu)^2 = \sigma^2(p) + \sigma^2(i) + \sigma^2(pi).$$

It is important to note that the total variance is a variance of scores for single persons on single items. In this sense, the variance components are for "single" scores, as opposed to mean scores, which are the focus of D study issues.

These variance components might be conceptualized in the following manner. Suppose N_p and N_i were *very* large, but still finite. Under this circumstance, in theory, all items in the universe could be administered to all persons in the population. Given the resulting observed scores, values for the mean scores in Equations 3 to 5 could be obtained. Then $\sigma^2(p)$ could be computed by taking the variance, over the population of persons, of the scores μ_p. Similarly, $\sigma^2(i)$ could be computed by taking the variance, over the universe of items, of the scores μ_i. Finally, $\sigma^2(pi)$ could be computed by taking the variance, over both persons and items, of the scores $X_{pi} - \mu_p - \mu_i + \mu$. This approach to interpreting variance components is clearly a contrived scenario, but it is a helpful conceptual aid.

In generalizability theory, variance components assume central importance. They are the building blocks that provide a crucial foundation for all subsequent results. Numerous procedures might be used to estimate variance components, but by far the most frequently employed procedure in generalizability theory is the so-called "ANOVA" procedure that involves equating mean scores to their expected values, and then solving for the estimated variance components. For the $p \times i$ design, the expected values of the mean squares ($\boldsymbol{E}MS$ equations) are:

$$\begin{aligned} \boldsymbol{E}MS(p) &= \sigma^2(pi) + n_i\, \sigma^2(p) \\ \boldsymbol{E}MS(i) &= \sigma^2(pi) + n_p\, \sigma^2(i) \\ \boldsymbol{E}MS(pi) &= \sigma^2(pi). \end{aligned}$$

Solving these equations for the variance components, and using mean squares in place of their expected values, we obtain the ANOVA estimators:

$$\hat{\sigma}^2(p) = [MS(p) - MS(pi)]/n_i \tag{8}$$

$$\hat{\sigma}^2(i) = [MS(i) - MS(pi)]/n_p \tag{9}$$

$$\hat{\sigma}^2(pi) = MS(pi). \tag{10}$$

2.3 D Studies for the $p \times I$ Design

To this point, discussion has focused on the model and associated variance components for a person's score on a single item (X_{pi}) in the universe of admissible observations. By contrast, if an examinee is administered a sample of n_i' items, *decisions* about the examinee will be based, surely, on his or her mean (or total) score over the n_i' items, not a score on a single item. From the perspective of generalizability theory, the intent of such a decision is to generalize from the examinee's observed mean score to the examinee's *universe score* over all items in the *universe of generalization*.

Another perspective on the universe of generalization is based on the notion of replications of a measurement procedure (see Brennan, 2001a). From this perspective, multiple measurements of an examinee would consist of his or her average score on different random samples of n_i' items from the same universe. Such samples of items, or test forms, are said to be *randomly parallel*, and for the $p \times I$ design, the universe of generalization can be viewed as a universe of such randomly parallel forms.

By convention, in generalizability theory average scores over a sample of conditions are indicated by uppercase letters. Using this notation for the D study $p \times I$ design, the linear model for the decomposition of an examinee's average score over n_i' items is

$$X_{pI} = \overline{X}_p = \mu + \nu_p + \nu_I + \nu_{pI}, \tag{11}$$

where X_{pI} and $\overline{X}_p$ mean the same thing. Equation 11 is completely analogous to Equation 6, the only difference being that i (for a single item) in Equation 6 is replaced everywhere by I (for the mean over a set of n_i' items) in Equation 11.

It is particularly important to note that

$$\mu_p \equiv \underset{I}{\boldsymbol{E}}\, X_{pI}, \tag{12}$$

which means that μ_p is defined as the expected value of X_{pI} over I in the universe of generalization. Alternatively, a person's universe score is the expected value of his or her observable mean score over all randomly parallel instances of the measurement procedure, each of which involves a different random sample of sets of conditions I. In short, the phrase "universe score" refers to the universe of generalization, not the universe of admissible observations.

The variance of the μ_p in Equation 12 is called *universe score variance*:

$$\sigma^2(p) = \underset{p}{\boldsymbol{E}}(\mu_p - \mu)^2 = \underset{p}{\boldsymbol{E}}\, \nu_p^2. \tag{13}$$

Alternatively, universe score variance is the expected value of the covariance between randomly parallel forms I and I'; that is,

$$\sigma^2(p) = \boldsymbol{E}\, \sigma(X_{pI}, X_{pI'}), \tag{14}$$

where the covariance is taken over persons, and the expectation is taken over pairs of randomly parallel forms.

Just as there are G study variance components associated with each of the random effects in Equation 7, so too there are D study variance components associated with the random effects in Equation 11. Definitions of these variance components are obtained by replacing i with I. The person variance component is unchanged by this replacement process, but the other two variance components *are* altered—namely,

$$\sigma^2(I) = \underset{I}{\boldsymbol{E}}(\mu_I - \mu)^2 = \underset{I}{\boldsymbol{E}}\,\nu_I^2 \tag{15}$$

and

$$\sigma^2(pI) = \underset{p}{\boldsymbol{E}}\,\underset{I}{\boldsymbol{E}}(X_{pI} - \mu_p - \mu_I + \mu)^2 = \underset{p}{\boldsymbol{E}}\,\underset{I}{\boldsymbol{E}}\,\nu_{pI}^2. \tag{16}$$

One well-known property of a distribution of mean scores for a set of uncorrelated observations is that the variance of the distribution is the variance of the individual elements divided by the sample size. It follows that

$$\sigma^2(I) = \sigma^2(i)/n_i', \tag{17}$$

and

$$\sigma^2(pI) = \sigma^2(pi)/n_i'. \tag{18}$$

2.4 Error Variances

Absolute error is the error involved in using an examinee's observed mean score as an estimate of his or her universe score. For person p, absolute error is defined as

$$\begin{aligned} \Delta_p &\equiv \overline{X}_p - \mu_p \\ &= \nu_I + \nu_{pI} \end{aligned} \tag{19}$$

since $\mu_p = \mu + \nu_p$. Absolute error is often associated with domain-referenced (or criterion-referenced) interpretations of scores. It is straightforward to show that absolute error variance is

$$\sigma^2(\Delta) = \sigma^2(I) + \sigma^2(pI) = \sigma^2(i)/n_i' + \sigma^2(pi)/n_i'. \tag{20}$$

Sometimes an investigator's interest focuses on the relative ordering of individuals with respect to their test performance, or the adequacy of the measurement procedure for making comparative decisions. In current terminology, such decisions are frequently associated with norm-referenced, or relative, interpretations of test scores. The crucial point about such interpretations is that a person's score is interpreted *not* in isolation but, rather, with respect to some measurement of group performance. The most obvious single measure of group performance is a group mean score, and the associated error is called "relative" error in generalizability theory.

For relative interpretations, a person's raw score X_{pI} carries no inherent meaning. It is the person's deviation score $X_{pI} - \mu_I$ that carries the meaning,

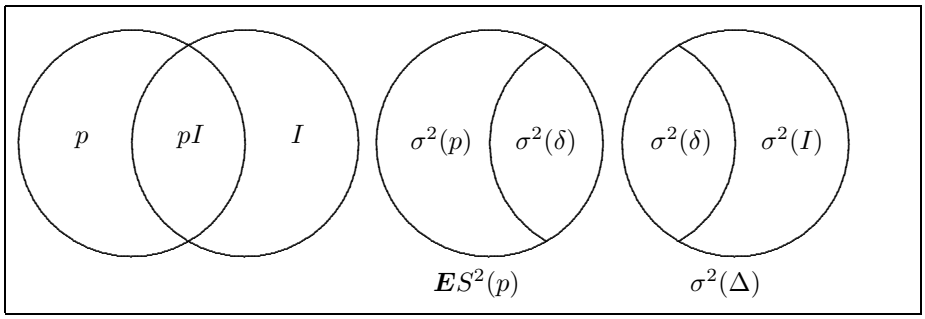

Figure 1: Venn diagrams for $p \times I$ design.

and this deviation score is to be interpreted as an estimate of the person's universe deviation score $\mu_p - \mu$. In accordance with this logic, relative error for person p is defined as

$$\begin{aligned} \delta_p &\equiv (\overline{X}_p - \underset{p}{\boldsymbol{E}}\,\overline{X}_p) - (\mu_p - \underset{p}{\boldsymbol{E}}\,\mu_p) && (21) \\ &= (\overline{X}_p - \mu_I) - (\mu_p - \mu) && (22) \\ &= \nu_{pI}. \end{aligned}$$

It follows immediately that relative error variance is

$$\sigma^2(\delta) = \sigma^2(pI) = \sigma^2(pi)/n'_i. \tag{23}$$

Relative error variance corresponds to the error variance in classical test theory, whereas absolute error variance is related to the "generic" error variance discussed by Lord and Novick (1968, pp. 177–180). From Equations 20 and 23, it is evident that

$$\sigma^2(\Delta) = \sigma^2(\delta) + \sigma^2(I).$$

Clearly, $\sigma^2(\Delta)$ is larger than $\sigma^2(\delta)$ unless $\sigma^2(I)$ is zero. The assumption of classically parallel forms implies that μ_I is a constant for all forms, which means that $\sigma^2(I)$ is zero. It follows that the assumption of classically parallel forms does not permit a formal distinction between $\sigma^2(\delta)$ and $\sigma^2(\Delta)$. In generalizability theory, however, randomly parallel forms can (and usually do) have different means μ_I, and $\sigma^2(I)$ is generally not zero. It follows that, from the perspective of generalizability theory, any test may consist of an especially easy or difficult set of items relative to the entire universe of items. Consequently, when X_{pI} is interpreted as an estimate of μ_p, variability in μ_I *does* contribute to the error variance $\sigma^2(\Delta)$. By contrast, the definition of relative error is such that μ_I is a constant for all persons in the *deviation* scores of interest and, therefore, $\sigma^2(I)$ does not contribute to $\sigma^2(\delta)$, even though $\sigma^2(I)$ may be positive.

Figure 1 provides a Venn diagram perspective on the difference between $\sigma^2(\Delta)$ and $\sigma^2(\delta)$ for the $p \times I$ design. Absolute error variance $\sigma^2(\Delta)$ is associated with the entire I circle, whereas relative error variance $\sigma^2(\delta)$ is that part of the I circle that is contained within the p circle.

2.5 Coefficients and Indices

Probably the most frequently reported statistic in generalizability analyses is a reliability-like coefficient called a generalizability coefficient, which employs $\sigma^2(\delta)$. Another commonly reported statistic is a dependability coefficient, which employs $\sigma^2(\Delta)$. There are also signal-noise ratios that correspond to these coefficients. Recently, Kane (1996) has proposed various error-tolerance ratios that are even more flexible than the traditional coefficients and signal-noise ratios.

2.5.1 Generalizability Coefficients

Classically parallel forms have equal observed score variances. In generalizability theory, randomly parallel forms have no such constraint, but *expected* observed score variance $\boldsymbol{E}S^2(p)$ plays a role like that of observed score variance in classical theory. For the $p \times I$ design,

$$\boldsymbol{E}S^2(p) \equiv \underset{I}{\boldsymbol{E}} \left[\underset{p}{\boldsymbol{E}} (X_{pI} - \mu_I)^2 \right].$$

That is, $\boldsymbol{E}S^2(p)$ is literally the expected value of the observed score variance for person mean scores, with the expectation taken over randomly parallel forms. It follows that

$$\boldsymbol{E}S^2(p) = \sigma^2(p) + \sigma^2(\delta). \tag{24}$$

In terms of the Venn diagrams in Figure 1, expected observed score variance is represented by the entire p circle, which can be partitioned into two parts—universe score variance and relative error variance—as indicated in Equation 24.

Cronbach et al. (1972) define a reliability-like coefficient called a generalizability coefficient, which is denoted $\boldsymbol{E}\rho^2$. A generalizability coefficient can be viewed as the ratio of universe score variance to expected observed score variance. Given the result in Equation 24, it follows that

$$\boldsymbol{E}\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\delta)}. \tag{25}$$

Technically, for the $p \times I$ design, $\boldsymbol{E}\rho^2$ is an intraclass correlation coefficient that is "stepped up" by the Spearman-Brown formula (see Feldt & Brennan, 1989). That is, letting $IC = \sigma^2(p)/[\sigma^2(p) + \sigma^2(pi)]$ be the intraclass correlation,

$$\boldsymbol{E}\rho^2 = \frac{n_i' \, IC}{1 + (n_i' - 1) \, IC} = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\delta)}.$$

As such, $\boldsymbol{E}\rho^2$ is clearly a measure of agreement.

There are other perspectives on $\boldsymbol{E}\rho^2$ that bear upon the interpretability of $\boldsymbol{E}\rho^2$ as an index of agreement. For example, the notation "$\boldsymbol{E}\rho^2$" introduced by Cronbach et al. (1972) is intended to imply that a generalizability coefficient is approximately equal to the expected value (over randomly parallel forms

of length n_i') of the *squared* correlation between observed scores and universe scores. Also, $\boldsymbol{E}\rho^2$ is approximately equal to the expected value of the correlation between pairs of randomly parallel forms of length n_i'. Another interpretation of $\boldsymbol{E}\rho^2$ as an agreement coefficient is discussed at the end of the next section.

In terms of estimates, $\boldsymbol{E}\hat{\rho}^2$ for a $p \times I$ design is identical to Cronbach's (1951) coefficient alpha, which is one of the most frequently cited references in all of the social science literature. When items are scored dichotomously, $\boldsymbol{E}\hat{\rho}^2$ for a $p \times I$ design is KR–20 (Kuder & Richardson, 1937). Note that these equalities are restrictive in two senses: (a) they are for the $p \times I$ design, only; and (b) they are equalities in terms of the estimates (not parameters) that result from replacing Equations 8–10 (the so-called "ANOVA" estimates) in Equation 25.

2.5.2 Dependability Coefficients

It is obvious from Equation 25 that a generalizability coefficient involves relative error variance $\sigma^2(\delta)$, and as such it is appropriate for norm-referenced interpretations of scores in the sense discussed in previously. Brennan and Kane (1977a,b) define a corresponding reliability-like coefficient that involves absolute error variance:

$$\Phi = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\Delta)}, \tag{26}$$

which is called a dependability coefficient. Note that the denominator of Φ is *not* the variance of persons' mean scores: it is the mean-squared deviation for persons, $\boldsymbol{E}(\overline{X}_p - \mu)^2$.

From Equation 26 it is not immediately obvious how to interpret Φ in terms of agreement. That matter has been addressed by Kane and Brennan (1980) in the broader context of a framework for considering many types of agreement coefficients. That framework is outlined next using most of the notational conventions in Kane and Brennan (1980). Then, it is shown that Equation 26 is a particular coefficient that can be derived from this general framework, along with another related coefficient.

For randomly parallel tests (or forms), let the degree of agreement between any two scores, s_i and s_j, be defined by an agreement function, $a(s_i, s_j)$ that satisfies the following three conditions: $a(s_i, s_j) \geq 0$; $a(s_i, s_j) = a(s_j, s_i)$; and $a(s_i, s_i) + a(s_j, s_j) \geq 2\, a(s_i, s_j)$. Note that s_i and s_j may be raw scores or they may be transformed in some way. For that reason, we use s (or S) rather than x (or X) as in other parts of this paper.

The random variable representing the score for person p on the k-th instance of the measurement procedure is denoted S_{pk}; similarly, S_{ql} is the random variable representing the score for person q on the l-th instance of the measurement procedure. It is assumed that for both instances the score points range from

$s_0 \ldots s_v$.[2] Then, expected agreement is defined as

$$A = \underset{p}{\boldsymbol{E}}\,\underset{k}{\boldsymbol{E}}\,\underset{l}{\boldsymbol{E}}\, a(S_{pk}, S_{pl}) = \sum_{i,j=0}^{v} a(s_i, s_j)\Pr(S_{pk} = s_i, S_{pl} = s_j), \tag{27}$$

maximum agreement is defined as

$$A_m = \underset{p}{\boldsymbol{E}}\,\underset{k}{\boldsymbol{E}}\, a(S_{pk}, S_{pk}) = \sum_{i,j=0}^{v} a(s_i, s_i)\Pr(S_{pk} = s_i), \tag{28}$$

and chance agreement is defined as

$$A_c = \underset{p}{\boldsymbol{E}}\,\underset{q}{\boldsymbol{E}}\,\underset{k}{\boldsymbol{E}}\,\underset{l}{\boldsymbol{E}}\, a(S_{pk}, S_{ql}) = \sum_{i,j=0}^{v} a(s_i, s_j)\Pr(S_{pk} = s_i, S_{ql} = s_j). \tag{29}$$

Using these definitions, two agreement coefficients can be specified. One is the proportion of maximum possible agreement:

$$\theta = \frac{A}{A_m}; \tag{30}$$

the other is the chance-corrected proportion of maximum possible agreement

$$\theta_c = \frac{A - A_c}{A_m - A_c}. \tag{31}$$

For example, the agreement function for threshold loss is $t(S_{pk}, S_{ql}) = 1$ if $S_{pk} = S_{ql}$ and 0 otherwise. For this loss function, θ is simply the proportion of consistent decisions, and θ_c is Cohen's (1960) coefficient kappa.

For domain-referenced (or criterion-referenced) interpretations, an examinee's score is typically compared to some mastery, passing, or cut score, λ. Under these circumstances, a domain-referenced agreement function can be defined as

$$d(S_{pI}, S_{qJ}) = (S_{pI} - \lambda)(S_{qJ} - \lambda). \tag{32}$$

Using Equations 27–29 with the $p \times I$ design, it can be shown that

$$A = \sigma^2(p) + (\mu - \lambda)^2,$$

$$A_m = \sigma^2(p) + (\mu - \lambda)^2 + \sigma^2(I) + \sigma^2(pI),$$

and

$$A_c = (\mu - \lambda)^2.$$

It follows from Equation 30 that θ for the domain-referenced agreement function in Equation 32 is

$$\Phi(\lambda) = \frac{\sigma^2(p) + (\mu - \lambda)^2}{\sigma^2(p) + (\mu - \lambda)^2 + \sigma^2(I) + \sigma^2(pI)} = \frac{\sigma^2(p) + (\mu - \lambda)^2}{\sigma^2(p) + (\mu - \lambda)^2 + \sigma^2(\Delta)}, \tag{33}$$

[2] If items are scored dichotomously and raw scores are under consideration, then v here has the same interpretation as n used elsewhere in this paper. Otherwise, however, it is important to note that v here is the ordinal number of the highest possible score.

where $\Phi(\lambda)$ is the notational convention introduced by Brennan and Kane (1977a,b).[3] It follows that, $\Phi(\lambda)$ is interpretable as the proportion of maximum possible agreement given the domain-referenced agreement function in Equation 32. Further, using this agreement function with Equation 31, θ_c is simply

$$\Phi = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\Delta)},$$

which is identical to Equation 26. In other words, Φ is interpretable as a chance-corrected proportion of maximum possible agreement for the agreement function $(S_{pI} - \lambda)(S_{qJ} - \lambda)$.

The Kane and Brennan (1980) framework for agreement coefficients also can be used to obtain $\boldsymbol{E}\rho^2$. Specifically, for the so-called "norm-referenced" agreement function

$$g(S_{pI}, S_{qJ}) = (S_{pI} - \mu_I)(S_{qJ} - \mu_J), \tag{34}$$

it can be shown that $\theta = \theta_c = \boldsymbol{E}\rho^2$. Note that the norm-referenced agreement function is the product of observed scores that are deviated from their respective means, whereas, for the domain-referenced agreement function, observed scores are deviated from the cut score, λ.

2.5.3 Signal-Noise Ratios

Generalizability and dependability coefficients are frequently reported in generalizability analyses. Their popularity is undoubtedly related in part to the fact that they are reliability-like coefficients, and such coefficients have been widely used in measurement contexts since the beginning of the last century. However, other coefficients and indices are sometimes informative and perhaps even more useful in certain contexts. Perhaps the most frequently cited competitor is signal–noise ratios.

In interpreting an error variance it is often helpful to compare its magnitude directly to universe score variance. One way to do so is to form the ratio of universe score variance to error variance, which is called a signal-noise ratio. For absolute error, the signal–noise ratio and its relationships with Φ are:

$$S/N(\Delta) = \frac{\sigma^2(p)}{\sigma^2(\Delta)} = \frac{\Phi}{1 - \Phi} \qquad \text{and} \qquad \Phi = \frac{S/N(\Delta)}{1 + S/N(\Delta)}.$$

Similarly, for relative error, the signal–noise ratio and its relationships with $\boldsymbol{E}\rho^2$ are:

$$S/N(\delta) = \frac{\sigma^2(p)}{\sigma^2(\delta)} = \frac{\boldsymbol{E}\rho^2}{1 - \boldsymbol{E}\rho^2} \qquad \text{and} \qquad \boldsymbol{E}\rho^2 = \frac{S/N(\delta)}{1 + S/N(\delta)}.$$

[3] Estimating $\Phi(\lambda)$ is slightly more complicated than it may appear, because $(\overline{X} - \lambda)^2$ is a biased estimator of $(\mu - \lambda)^2$. An unbiased estimator is $(\overline{X} - \lambda)^2 - \hat{\sigma}^2(\overline{X})$, where

$$\hat{\sigma}^2(\overline{X}) = \sigma^2(p)/n'_p + \sigma^2(i)/n'_i + \sigma^2(pi)/n'_p n'_i.$$

When $\lambda = \overline{X}$, $\widehat{\Phi}(\lambda) =$ KR–21 (see Kuder & Richardson, 1937), which is indicative of the fact that KR–21 involves absolute error variance, not the relative error variance of classical theory.

As discussed by Cronbach and Gleser (1964) and Brennan and Kane (1977b), the signal–noise concept arises naturally in discussing communication systems where the signal–noise ratio compares the strength of the transmission to the strength of the interference. The signal $\sigma^2(p)$ is a function of the magnitude of the intended discriminations $\mu_p - \mu$. These intended discriminations reflect the sensitivity requirements that must be met if the measurement procedure is to achieve its intended purpose. The noise reflects the degree of precision, or the magnitude of the errors that arise in practice. If the signal is large compared to the noise, the intended discriminations are easily made. If the signal is weak compared to the noise, the intended discriminations may be completely lost.

As noted above, there are simple functional relationships between $S/N(\delta)$ and $\boldsymbol{E}\rho^2$, and between $S/N(\Delta)$ and Φ. It is reasonable to ask, therefore, "Do signal-noise ratios have any advantages over $\boldsymbol{E}\rho^2$ and Φ?" This author believes the answer is, "Yes" for at least two reasons. First, signal noise ratios have a simpler mathematical form that is readily interpretable in the manner indicated in the previous paragraph. Second, decreasing the error variance by a factor of k increases the signal-noise ratio by the same multiplicative factor, which is not true for $\boldsymbol{E}\rho^2$ and Φ.

Investigators often seem comforted by the fact that $\boldsymbol{E}\rho^2$ and Φ are scale invariant in the sense that they have values that range from 0 to 1, no matter what the measurement procedure may be. However, this scale invariant characteristic is essentially a result of the fact that $\boldsymbol{E}\rho^2$ and Φ are non-linear transformations of signal-noise ratios. One consequence of this fact that it is much easier to raise a low value of a $\boldsymbol{E}\rho^2$ or Φ by a particular amount than it is to raise a high value by the same amount. For example, to increase $\boldsymbol{E}\rho^2$ from .50 to .55 takes a 22% increase in the number of items, but to increase $\boldsymbol{E}\rho^2$ from .90 to .95 takes an 111% increase in the number of items.

2.5.4 Error-Tolerance Ratios

Kane (1996) begins an extended discussion of measurement precision by noting that:

> Estimates of standard errors provide us with an index of precision that is generic in the sense that it is independent of the demands of a specific use of the measurement procedure (although not independent of the score scale or the population). Such generic indexes can be helpful, but we also need an indication of the tolerance for error in the context in which the measurement is being applied if we are to have a sense of precision in use. That is, ultimately, we need to answer to the following question: Are the measurements precise enough for the particular use under consideration? To answer this question, we need to know how much error can be tolerated in a particular situation before the errors cause serious problems.

To quantify measurement precision, Kane (1996) proposes the use of an index that he calls an "error-tolerance ratio." This index is denoted (E/T) and

defined as the root mean square of the errors of interest divided by the root mean square of the tolerances of interest. For the errors δ and Δ, the error-tolerance ratios are

$$E_\delta/T = \frac{\sigma(\delta)}{\sqrt{\boldsymbol{E}\mathcal{T}^2}} \quad \text{and} \quad E_\Delta/T = \frac{\sigma(\Delta)}{\sqrt{\boldsymbol{E}\mathcal{T}^2}},$$

where $\mathcal{T}$ is the tolerance. Clearly, E/T will tend to be small if errors are small relative to the tolerances, suggesting that measurements have substantial precision for the intended use.

If the tolerance is $\mathcal{T} = \mu_p - \mu$, then the root mean square of the tolerances is simply $\sigma(p)$. If, in addition, the error root mean square is $\sigma(\delta)$, then relationships between E_δ/T and $\boldsymbol{E}\rho^2$ are:

$$E_\delta/T = \sqrt{\frac{1-\boldsymbol{E}\rho^2}{\boldsymbol{E}\rho^2}} \quad \text{and} \quad \boldsymbol{E}\rho^2 = \frac{1}{1+(E_\delta/T)^2}.$$

Similarly, if the error root mean square is $\sigma(\Delta)$, then, relationships between E_Δ/T and Φ are:

$$E_\Delta/T = \sqrt{\frac{1-\Phi}{\Phi}} \quad \text{and} \quad \Phi = \frac{1}{1+(E_\Delta/T)^2}.$$

In general, note that if the error-tolerance ratio is .5, the coefficient is .8.

The "error" in an error-tolerance ratio can also be considered from the perspective of disagreement in observed scores for a person:

$$X_{pI} - X_{pJ} = (\mu + \nu_p + \nu_I + \nu_{pI}) - (\mu + \nu_p + \nu_J + \nu_{pJ}) = (\nu_I - \nu_J) + (\nu_{pI} - \nu_{pJ})$$

The root mean square of these errors is

$$\sigma(X_{pI} - X_{pJ}) = \sqrt{2}\,\sigma(I) + \sqrt{2}\,\sigma(pI) = \sqrt{2}\,\sigma(\Delta)$$

If $I = J$ or, equivalently I and J are "equally difficult,"

$$\sigma(X_{pI} - X_{pJ}) = \sqrt{2}\,\sigma(\delta).$$

As a benchmark for judging the size of the error $X_{pI} - X_{pJ}$, suppose we consider the tolerance to be $\mathcal{T} = \mu_p - \mu_q$, where p and q are different persons. The root mean square of this tolerance is

$$\sigma(\mu_p - \mu_q) = \sqrt{2}\,\sigma(p).$$

Under these circumstances the error-tolerance ratios are

$$E_\Delta/T = \frac{\sqrt{2}\,\sigma(\Delta)}{\sqrt{2}\,\sigma(p)} = \frac{\sigma(\Delta)}{\sigma(p)} \quad \text{and} \quad E_\delta/T = \frac{\sqrt{2}\,\sigma(\delta)}{\sqrt{2}\,\sigma(p)} = \frac{\sigma(\delta)}{\sigma(p)},$$

which are identical to the previously discussed results, although the derivations and rationale are somewhat different.

The notion of an error-tolerance ratio is not restricted to situations involving only square roots of variances, however. For example, for domain-referenced interpretations of scores, often interest focuses on the tolerances $\mathcal{T} = \mu_p - \lambda$, where λ is a cut score. Under these circumstances, the root mean square of the tolerances is

$$\sqrt{\boldsymbol{E}\mathcal{T}^2} = \sqrt{\boldsymbol{E}(\mu_p - \lambda)^2} = \sqrt{\boldsymbol{E}[(\mu_p - \mu) + (\mu - \lambda)]^2} = \sqrt{\sigma^2(p) + (\mu - \lambda)^2}, \tag{35}$$

the standard error of measurement of interest is

$$\sqrt{\boldsymbol{E}[(X_{pI} - \lambda) - (\mu_p - \lambda)]^2} = \sqrt{\boldsymbol{E}(X_{pI} - \mu_p)^2} = \sigma(\Delta),$$

and the error-tolerance ratio is

$$E_\Delta/T(\lambda) = \frac{\sigma(\Delta)}{\sqrt{\sigma^2(p) + (\mu - \lambda)^2}}. \tag{36}$$

Using this error-tolerance ratio, we can obtain the dependability coefficient in Equation 33 in the following manner:

$$\Phi(\lambda) = \frac{1}{1 + \dfrac{\sigma^2(\Delta)}{\sigma^2(p) + (\mu - \lambda)^2}} = \frac{\sigma^2(p) + (\mu - \lambda)^2}{\sigma^2(p) + (\mu - \lambda)^2 + \sigma^2(\Delta)}.$$

Clearly, generalizability coefficients, dependability coefficients, and signal-noise ratios can be expressed as functions of error-tolerance ratios. However, error-tolerance ratios are even more flexible, particularly because the tolerance can take on any form that is meaningful to the investigator. For example, in grading essays it is often the case that discrepancies of more than one point are adjudicated in such a manner that the final ratings are no more discrepant than a single point. Under these circumstances, it may be quite sensible to set the tolerance to 1. This may be a rather simplistic example, but it illustrates the flexibility of error-tolerance ratios. As stated by Kane (1996):

> The tolerance for error depends on the context and intended use of the results of measurement because the consequences of errors of different sizes depend on the context and intended use. In some cases, it may be necessary to keep errors very small in order to achieve the desired outcome. In other cases, even relatively large errors might not interfere with the interpretation. A standard error of an ounce in a bathroom scale represents unnecessary precision. The same standard error in a laboratory scale used to weigh tissue samples would be unacceptably large. (p. 359)

Generalizability coefficients, dependability coefficients, and signal-noise ratios are typically regarded as indices that bear upon the reliability of measurements. Several investigators have noted, however, that generalizability theory

blurs distinctions between reliability and validity (e.g., Cronbach et al., 1972, p. 380; Kane, 1982; Brennan, 2001a; see also the previous discussion of the reliability-validity paradox). This is particularly evident in the case of error-tolerance ratios. As noted by Kane (1996):

> Tolerance is closely connected to the concept of validity because the tolerance for error is determined by the impact of errors of measurement on interpretations of decisions. In general, an evaluation of validity of a proposed interpretations of test scores or decisions based on scores involves an examination of plausibility of the inferences in the interpretation and decision .... To ensure that random errors of measurement are not undermining the intended outcomes, it is necessary to show that the errors are small compared with the tolerance for errors implied by the proposed application. Therefore, an evaluation of the magnitude of errors relative to the tolerance for error is an integral part of any validation effort. The assertion of classical test theory that reliability is necessary but not sufficient for validity is a special case of this general conclusion. (pp. 359–360)

2.5.5 Example: Blood Pressure

Llabre et al. (1988, p. 97) apply generalizability theory to blood pressure measurements "in order to determine the number of readings needed to attain reliable estimates." Their paper provides results for a number of well-designed studies for various universes of generalization. Here, we focus on analyses in which generalization is over only one facet, replications (r) of the measurement procedure.

Each of 40 subjects (p) had their blood pressure taken three times (i.e., $n_r = 3$) using an ambulatory monitor. The design was repeated in three different locations: a laboratory, home, and work. Within a location, readings were taken on the same day. Table 1 provides a summary of the results for systolic and diastolic readings in the three settings.[4] Note that the metric here is millimeters of mercury (mm Hg).

The results for $\widehat{\Phi}$ in Table 1 led Llabre et al. to conclude that

> ...only one reading is necessary whenever generalizations are restricted to the same day in the laboratory. At least six readings of systolic blood pressure are needed at home and at work, and 6 to 10 diastolic blood pressure readings may be required from work and home, respectively.

The Llabre et al. standard for their conclusions seems to be that $\widehat{\Phi}$ be about .80 or greater.

The bottom part of Table 1 provides estimated values for the error-tolerance ratio $E_\Delta/T(\lambda)$ in Equation 36 using $\mathcal{T} = \mu_p - \lambda$, with $\lambda = 120$ for the systolic

[4] Llabre et al. (1988) did not report error-tolerance ratios, which are discussed later.

Table 1: Llabre et al. (1988) Study of Blood Pressure

	Estimated G Study Variance Components					
	Laboratory		Home		Work	
Effect	Sys.	Dias.	Sys.	Dias.	Sys.	Dias.
p	125.33	64.21	143.51	49.33	150.07	57.62
r	.89	.06	6.06	1.55	.00[a]	2.92
pr	22.57	13.66	228.84	111.41	166.82	80.29
$\hat{\sigma}(\Delta)$						
$n'_r = 1$	4.84	3.70	15.33	10.63	12.92	9.12
$n'_r = 2$	3.42	2.62	10.84	7.52	9.13	6.45
$n'_r = 3$	2.80	2.14	8.85	6.14	7.46	5.27
$n'_r = 4$	2.42	1.85	7.66	5.31	6.46	4.56
$n'_r = 5$	2.17	1.66	6.85	4.75	5.78	4.08
$n'_r = 6$	1.98	1.51	6.26	4.34	5.27	3.72
$n'_r = 10$	1.53	1.17	4.85	3.36	4.08	2.88
$\widehat{\Phi}$						
$n'_r = 1$	.84	.82	.38	.30	.47	.41
$n'_r = 2$	.91	.90	.55	.47	.64	.58
$n'_r = 3$	.94	.93	.65	.57	.73	.68
$n'_r = 4$	.96	.95	.71	.64	.78	.73
$n'_r = 5$	.96	.96	.75	.69	.82	.78
$n'_r = 6$	.97	.97	.79	.72	.84	.81
$n'_r = 10$	.98	.98	.86	.81	.90	.87
$\overline{X}$[b]	124	79	122	74	115	76
λ	120	80	120	80	120	80
$Est[E_\Delta/T(\lambda)]$						
$n'_r = 1$	.41	.47	1.33	1.19	1.00	1.11
$n'_r = 2$	.29	.33	.92	.83	.70	.77
$n'_r = 3$	.24	.27	.75	.67	.57	.63
$n'_r = 4$	.21	.23	.65	.58	.49	.54
$n'_r = 5$	.18	.21	.58	.52	.44	.48
$n'_r = 6$	.17	.19	.53	.48	.40	.44
$n'_r = 10$	.13	.15	.41	.37	.31	.33

[a] Slightly negative value set to 0.

[b] Values provided by Maria Llabre.

reading, and $\lambda = 80$ for the diastolic reading. For example, for the laboratory systolic reading with $n'_r = 2$, using Equation 35 and footnote 3, the estimated tolerance is:

$$\begin{aligned}\sqrt{Est[\boldsymbol{E}T^2]} &= \sqrt{\hat{\sigma}^2(p) + (\overline{X} - \lambda)^2 - \hat{\sigma}^2(\overline{X})} \\ &= \sqrt{125.33 + (124 - 120)^2 - \left(\frac{125.33}{40} + \frac{.89}{2} + \frac{22.57}{40 \times 2}\right)} \\ &= 11.72.\end{aligned}$$

Therefore, the estimated value of $E_\Delta/T(\lambda)$ is $3.42/11.72 = .29$, which means that the error is about 29% of the tolerance, given a cut score of 120 for systolic blood pressure.[5]

Here, we are viewing the results in Table 1 as six separate studies—systolic and diastolic readings for each of three locations—with only one facet (replications) in the universe of generalization. Clearly, however, a more sophisticated (but considerably more complicated) analysis might explicitly represent "location" and "type of reading" as additional facets.

2.6 Summary

Table 2 summarizes the most important equations for the $p \times I$ design that have been introduced here. The top of the table provides formulas for errors and their variances. The middle provides formulas for universe score variance and expected observed score variance. The rest of the table provides formulas for the basic coefficients and indices that have been discussed. In most of Table 2 the symbol τ is used instead of p for two reasons. First, in principle, any facet could serve as the objects of measurement. Second, using τ, the formula for $\boldsymbol{E}S^2(\tau)$ and the formulas in the bottom third of the table are quite general; in particular, they apply to multifacet random and mixed model designs, which is the subject of the next section.

3 Multifacet Universes and Designs

The procedures of generalizability theory can be applied to universes and designs with any number of facets. However, to illustrate the theory, it is often sufficient to restrict attention to two facet designs. Here we use t and r to designate levels of the two facets, and we assume the universe of admissible observations has $t \times r$. The reader can think of t as standing for "task" and r as standing for "rater," although the theory to be discussed is blind to these mnenomic conventions. All of the formulas in Table 2 still apply except those that specify which variance components enter $\sigma^2(\tau)$, $\sigma^2(\Delta)$, and $\sigma^2(\delta)$.[6]

[5] Strictly speaking $Est\sqrt{\boldsymbol{E}T^2}$ is not identical to $\sqrt{Est[\boldsymbol{E}T^2]}$, and the estimate of a ratio is not identical to the ratio of the estimates, but we neglect these issues here.

[6] For a much more detailed discussion of multifacet universes and designs, see Brennan, 2001b.

Table 2: Equations for D Study $p \times I$ Design

Absolute Error:	$\Delta = X_{pI} - \mu_p$
Relative Error:	$\delta = (X_{pI} - \mu_I) - (\mu_p - \mu)$
Absolute Error Variance:	$\sigma^2(\Delta) = [\sigma^2(i) + \sigma^2(pi)]/n'_i$
Relative Error Variance:	$\sigma^2(\delta) = \sigma^2(pi)/n'_i$
Universe Score Variance:	$\sigma^2(\tau) = \sigma^2(p)$
Exp. Obs. Score Variance:	$\boldsymbol{E}S^2(\tau) = \sigma^2(\tau) + \sigma^2(\delta)$
Generalizability Coefficient (uses δ):	$\boldsymbol{E}\rho^2 = \dfrac{\sigma^2(\tau)}{\boldsymbol{E}S^2(\tau)} = \dfrac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)}$
Dependability Coefficients (use Δ):	$\Phi = \dfrac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)}$
	$\Phi(\lambda) = \dfrac{\sigma^2(\tau) + (\mu - \lambda)^2}{\sigma^2(\tau) + (\mu - \lambda)^2 + \sigma^2(\Delta)}$
Signal-Noise Ratio using δ:	$S/N_\delta = \dfrac{\sigma^2(\tau)}{\sigma^2(\delta)} = \dfrac{\boldsymbol{E}\rho^2}{1 - \boldsymbol{E}\rho^2}$
Signal-Noise Ratio using Δ:	$S/N_\Delta = \dfrac{\sigma^2(\tau)}{\sigma^2(\Delta)} = \dfrac{\Phi}{1 - \Phi}$
Error-Tolerance Ratio using δ:	$E_\delta/T = \dfrac{\sigma(\delta)}{\sqrt{\boldsymbol{E}T^2}}$
Error-Tolerance Ratios using Δ:	$E_\Delta/T = \dfrac{\sigma(\Delta)}{\sqrt{\boldsymbol{E}T^2}}$
	$E_\Delta/T(\lambda) = \dfrac{\sigma(\Delta)}{\sqrt{\sigma^2(p) + (\mu - \lambda)^2}}$

We will consider two possible G study designs: the fully crossed $p \times t \times r$ design,

$$X_{ptr} = \mu + \nu_p + \nu_t + \nu_r + \nu_{pt} + \nu_{pr} + \nu_{tr} + \nu_{ptr};$$

and the (partially) nested $p \times (r\!:\!t)$ design:

$$X_{ptr} = \mu + \nu_p + \nu_t + \nu_{r:t} + \nu_{pt} + \nu_{pr:t}.$$

We will assume, as well, that G study variance components are estimated for the random model.

To continue to keep matters simple, suppose the universe of generalization has $T \times R$. We will consider two possible D study designs for the decomposition of a person's observed mean score over n'_t and n'_r conditions: the fully crossed $p \times T \times R$ design,

$$X_{pTR} = \mu + \nu_p + \nu_T + \nu_R + \nu_{pT} + \nu_{pR} + \nu_{TR} + \mu_{pTR}; \tag{37}$$

and the (partially) nested $p \times (R{:}T)$ design,

$$X_{pTR} = \mu + \nu_p + \nu_T + \nu_{R:T} + \nu_{pT} + \nu_{pR:T}. \tag{38}$$

As in previous sections, we occasionally use $\overline{X}_p$ as an abbreviation for X_{pTR}.

3.1 Random Models and Infinite Universes of Generalization

D study variance components [denoted generically as $\sigma^2(\bar{\alpha})$] are obtained by dividing the G study variance components [denoted generically as $\sigma^2(\alpha)$] by D study sample sizes. Specifically,

$$\sigma^2(\bar{\alpha}) = \sigma^2(\alpha)/d(\bar{\alpha}), \tag{39}$$

where

$$d(\bar{\alpha}) = \begin{cases} 1 \text{ if } \bar{\alpha} = p, \text{ and, otherwise, the product} \\ \text{of the D study sample sizes } (n') \text{ for all} \\ \text{indices in } \bar{\alpha} \text{ except } p. \end{cases}$$

For the two illustrative designs, the D study variance components are provided in the first columns of Tables 3 and 4.

All of the formulas in Table 2 apply to multi-facet random model designs except those for $\sigma^2(\delta)$ and $\sigma^2(\Delta)$. For these designs, the variance components that enter the universe score variance (denoted generically as $\sigma^2(\tau)$), $\sigma^2(\delta)$ and $\sigma^2(\Delta)$ can be obtained as follows:

- *Rule*: $\sigma^2(\tau)$ is $\sigma^2(p)$;
- *Rule*: $\sigma^2(\Delta)$ is the sum of all $\sigma^2(\bar{\alpha})$ except $\sigma^2(p)$; and
- *Rule*: $\sigma^2(\delta)$ is the sum of all $\sigma^2(\bar{\alpha})$ such that$\bar{\alpha}$ includes p and at least one other index.

For the two illustrative designs, the D study variance components that enter $\sigma^2(\delta)$ and $\sigma^2(\delta)$ are indicated in the second columns of Tables 3 and 4.

3.2 Simplified Procedures for Mixed Models

Fixing a facet results in a *restricted* universe of generalization in the sense that it is smaller than it would be for a random model. Intuitively, restricting a universe should lead to smaller error variances, because the bridge from the observed data to the universe has a shorter span (see the Cornfield & Tukey,

1956, pp. 912–913, "bridge analogy). This intuitive notion is formalized later in this section.

For mixed models, it is theoretically preferable, albeit more complicated, to employ multivariate generalizability theory. However, there are univariate simplified procedures that can be employed, provided that:

1. the estimated G study variance components are for a random model,
2. each facet in the D study is either random or fixed (i.e., no sampling from a finite universe), and
3. the design is balanced in the sense that there is no missing data and, if one facet is nested within another, the number of levels of the nested facet is a constant.

Note that, in any generalizability analysis there must be at least one random facet for the analysis to be meaningful. If all facets were fixed, then no generalization is involved, and all error variances are zero, be definition.

Let an observed mean score be $X_{p\mathcal{RF}}$, where

- $\mathcal{F}$ is the set of fixed facets in the universe of generalization
- $\mathcal{R}$ is the set of random facets in the universe of generalization

The simplified rules for mixed models are:

- *Rule*: $\sigma^2(\tau)$ is the sum of all $\sigma^2(\bar{\alpha})$ such that $\bar{\alpha}$ includes τ and does *not* include any index in $\mathcal{R}$;
- *Rule*: $\sigma^2(\Delta)$ is the sum of all $\sigma^2(\bar{\alpha})$ such that $\bar{\alpha}$ includes at least one of the indices in $\mathcal{R}$; and
- *Rule*: $\sigma^2(\delta)$ is the sum of all $\sigma^2(\bar{\alpha})$ such that $\bar{\alpha}$ includes τ *and* at least one of the indices in $\mathcal{R}$.

Table 3: Random Effects Variance Components that Enter $\sigma^2(\tau)$, $\sigma^2(\delta)$, and $\sigma^2(\Delta)$ for Random and Mixed Model D Studies for the $p \times T \times R$ Design

	D Studies		
$\sigma^2(\bar{\alpha})$	T, R Random	T Fixed	R Fixed
$\sigma^2(p)$	τ	τ	τ
$\sigma^2(T) = \sigma^2(t)/n'_t$	Δ		Δ
$\sigma^2(R) = \sigma^2(r)/n'_r$	Δ	Δ	
$\sigma^2(pT) = \sigma^2(pt)/n'_t$	Δ, δ	τ	Δ, δ
$\sigma^2(pR) = \sigma^2(pr)/n'_r$	Δ, δ	Δ, δ	τ
$\sigma^2(TR) = \sigma^2(tr)/n'_t n'_r$	Δ	Δ	Δ
$\sigma^2(pTR) = \sigma^2(ptr)/n'_t n'_r$	Δ, δ	Δ, δ	Δ, δ

Table 4: Random Effects Variance Components that Enter $\sigma^2(\tau)$, $\sigma^2(\delta)$, and $\sigma^2(\Delta)$ for Random and Mixed Model D Studies for the $p \times (R{:}T)$ Design

	D Studies	
$\sigma^2(\bar{\alpha})$	T, R Random	T Fixed
$\sigma^2(p)$	τ	τ
$\sigma^2(T) = \sigma^2(t)/n'_t$	Δ	
$\sigma^2(R{:}T) = \sigma^2(r{:}t)/n'_r n'_t$	Δ	Δ
$\sigma^2(pT) = \sigma^2(pt)/n'_t$	Δ, δ	τ
$\sigma^2(pR{:}T) = \sigma^2(pr{:}t)/n'_r n'_t$	Δ, δ	Δ, δ

For the two illustrative designs, the D study variance components that enter $\sigma^2(\Delta)$ and $\sigma^2(\delta)$ for mixed models are indicated in the last two columns of Tables 3 and the last column of Table 4. These mixed-model rules are in accord with the intuitive notion mentioned above that fixing a facet leads to smaller error variances over what they would be for a random model.

In addition, these rules demonstrate that fixing a facet increases universe score variance. For example, for the fully crossed $p \times T \times R$ design with T fixed, $\sigma^2(p|T) = \sigma^2(p) + \sigma^2(pt)/n'_t$. An intuitive understanding of why a mixed model leads to an increase in universe score variance can be obtained by recalling that $\boldsymbol{E}S^2(\tau) = \sigma^2(\tau) + \sigma^2(\delta)$ is a constant in any analysis. Therefore, since fixing a facet decreases $\sigma^2(\delta)$, universe score variance must increase. Similarly, $\sigma^2(\overline{X}_\tau) = \sigma^2(\tau) + \sigma^2(\Delta)$ is a constant; so, a decrease in $\sigma^2(\Delta)$ leads to an increase in $\sigma^2(\tau)$.

Mathematically, what happens in that fixing a facet leads to different definitions of the score effects. For example, for the $p \times T \times R$ design, if T is fixed then Equation 37 can be rewritten

$$\begin{aligned} X_{pTR} &= \mu + \nu_p + \nu_T + \nu_R + \nu_{pT} + \nu_{pR} + \nu_{TR} + \nu_{pTR} \\ &= (\mu + \nu_T) + (\nu_p + \nu_{pT}) + (\nu_R + \nu_{TR}) + (\nu_{pR} + \mu_{pTR}) \\ &= \mu^* + \nu_p^* + \nu_R^* + \nu_{pR}^*. \end{aligned}$$

That is, when the T facet is fixed, effects associated with it get absorbed into other effects. Since effects are uncorrelated, it follows, for example, that

$$\sigma^2(\nu_p^*) = \sigma^2(\nu_p + \nu_{pT}) = \sigma^2(\nu_p) + \sigma^2(\nu_{pT}) = \sigma^2(p) + \sigma^2(pt)/n'_t,$$

as noted previously.

It is important to note that in generalizability theory score effects in linear models are *defined* with respect to mean scores. As noted previously, this means that for random effects the expected values are necessarily 0. Similarly, for fixed effects the sums are necessarily 0. In this sense, the zero sums for fixed effects are *integral* to generalizability theory; they are not additional "restrictions" or "constraints" imposed on the model. This perspective on mixed models is similar to

that of Scheffé (1959). By contrast, no zero-sum restrictions for fixed effects are employed in the general linear model (see, for example, Searle, 1971, chap. 9). Consequently, there are a number of procedures for estimating variance components (and computer programs for doing so, including some procedures in SAS) that do not employ these restrictions, and such procedures are not applicable in generalizability theory. To put it bluntly, generalizability theory with fixed facets is not isomorphic with other variance components perspectives on the so-called "general mixed model." (See Brennan, 2001b, pp. 86–88 for more discussion.)

3.3 Performance Assessment Example

Shavelson, Baxter, and Gao (1993, p. 222) provide an instructive example of a performance assessment program in science called the California Assessment Program (CAP). They state that:

> Students were posed five independent tasks. More specifically, students rotated through a series of five self-contained stations at timed intervals (about 15 mins.). At one station, students were asked to complete a problem solving task (determine which of these materials may serve as a conductor). At the next station, students were asked to develop a classification system for leaves and then to explain any adjustments necessary to include a new mystery leaf in the system. At yet another, students were asked to conduct tests with rocks and then use the results to determine the identity of an unknown rock. At the fourth station, students were asked to estimate and measure various characteristics of water (e.g., temperature, volume). And at the fifth station, students were asked to conduct a series of tests on samples of lake water to discover why fish are dying (e.g., is the water too acidic?). At each station, students were provided with the necessary materials and asked to respond to a series of questions in a specified format (e.g., fill in a table).
>
> A predetermined scoring rubric developed by teams of teachers in California was used to evaluate the quality of students' written responses ... to each of the tasks. Each rubric was used to score performance on a scale from 0 to 4 (0 = no attempt, 1 = serious flaws, 2 = satisfactory, 3 = competent, 4 = outstanding). All tasks were scored by three raters.

For the CAP, the G study design is $p \times t \times r$ with $n_t = 5$ tasks and $n_r = 3$ raters. We treat tasks and raters are both random for the G study. Table 5 reports the G study estimated variance components, along with D studies for various sample sizes. D study results in parentheses at the bottom of the table are for a mixed model in which tasks are fixed. Otherwise, results are for a random model. Note that the values of $\hat{\Phi}(\lambda)$ and $Est[E_\Delta/T(\lambda)]$ are based on the hypothetical assumption that $\overline{X} = 2.5$ and $\lambda = 3$.

Table 5: D Study $p \times T \times R$ Designs For CAP Data

		D Studies					
	n'_t	5	5	5	10	10	10
$\hat{\sigma}^2(\alpha)$	n'_r	1	2	3	1	2	3
$\hat{\sigma}^2(p) = .298$	$\hat{\sigma}^2(p)$	.298	.298	.298	.298	.298	.298
$\hat{\sigma}^2(t) = .092$	$\hat{\sigma}^2(T)$	.018	.018	.018	.009	.009	.009
$\hat{\sigma}^2(r) = .003$	$\hat{\sigma}^2(R)$	.003	.002	.001	.003	.002	.001
$\hat{\sigma}^2(pt) = .493$	$\hat{\sigma}^2(pT)$	.099	.099	.099	.049	.049	.049
$\hat{\sigma}^2(pr) = .000$	$\hat{\sigma}^2(pR)$	.000	.000	.000	.000	.000	.000
$\hat{\sigma}^2(tr) = .002$	$\hat{\sigma}^2(TR)$	.000	.000	.000	.000	.000	.000
$\hat{\sigma}^2(ptr) = .148$	$\hat{\sigma}^2(pTR)$	.030	.015	.010	.015	.007	.005
	$\hat{\sigma}^2(\tau)$	.30(.40)	.30(.40)	.30(.40)	.30	.30	.30
	$\hat{\sigma}^2(\delta)$	.13(.03)	.11(.01)	.11(.01)	.06	.06	.05
	$\hat{\sigma}^2(\Delta)$	.15(.03)	.13(.02)	.13(.01)	.08	.07	.06
	$\hat{\sigma}^2(\overline{X})$	.03(.01)	.02(.01)	.02(.01)	.02	.01	.01
	$\boldsymbol{E}\hat{\rho}^2$	.70(.93)	.72(.96)	.73(.98)	.82	.84	.85
	$\widehat{\Phi}$	.67(.92)	.69(.96)	.70(.97)	.80	.82	.82
	$\hat{\Phi}(\lambda)$	.78(.95)	.80(.97)	.80(.98)	.87	.89	.89
	$Est(E_\delta/T)$	.66(.95)	.62(.95)	.60(.95)	.46	.44	.43
	$Est(E_\Delta/T)$	.71(.29)	.67(.20)	.66(.17)	.51	.48	.47
	$Est[E_\Delta/T(\lambda)]$	.54(.23)	.50(.16)	.49(.13)	.38	.36	.35

Notes. G study estimated variance components were provided by Xiaohong Gao. Results for $\hat{\Phi}(\lambda)$ and $Est[E_\Delta/T(\lambda)]$ assume $\overline{X} = 2.5$ and $\lambda = 3$. In computing both of these statistics, $Est(\mu - \lambda)^2 = (\overline{X} - \lambda)^2 - \hat{\sigma}^2(\overline{X})$, where $\hat{\sigma}^2(\overline{X}) = [\sigma^2(\tau) + \sigma^2(\delta)]/n'_p + [\sigma^2(\Delta) - \sigma^2(\delta)]$ is an estimate of $\boldsymbol{E}(\overline{X} - \mu)^2$. For these analyses, $n'_p = 100$.

It is evident from Table 5 that, for the random model, increasing the number of raters and/or tasks

- leaves universe score variance unchanged,
- decreases error variances,
- increased coefficients, and
- decreases error-tolerance ratios.

All of these results are predictable from the formulas for these quantities. The same results apply for the mixed model with tasks fixed.

It is also evident from Table 5 that, compared to the random model, the mixed model leads to

- an increase in universe score variance,
- decreases in error variances,

- increases in coefficients, and
- decreases in error-tolerance ratios.

Again, these results are predictable from the formulas for these quantities.

One of the most famous results in classical test theory is the Spearman-Brown formula, the simplest version of which states that the reliability of a double-length test (say r') relative to the reliability of the original test (say r) is $r' = 2\,r/(1+r)$. For the CAP example, "double-length" can be interpreted as twice as many observations contributing to an examinee's score. Note from Table 5 that with $n'_t = 5$ and $n'_r = 1$, the random model $\boldsymbol{E}\hat{\rho}^2$ is .70. Applying the Spearman-Brown formula gives a predicted reliability of $r' = 2(.70)/(1+.70) = .82$, which is dramatically different from the random model $\boldsymbol{E}\hat{\rho}^2$ is .72 for $n'_t = 5$ and $n'_r = 2$. Why? The reason is that, for the $\boldsymbol{E}\hat{\rho}^2$ result, $\hat{\sigma}^2(pT) = .099$ (which contributes to error variance) is unaffected by doubling the number of raters; whereas, the Spearman-Brown formula effectively divides this variance component by two. Hence, the Spearman-Brown result is positively biased. Indeed, the the Spearman-Brown formula generally does not apply when there are two or more random facets.

Another tenet of classical test theory is that longer tests are more reliable than shorter ones. Stated more generally, scores based on more observations are more reliable than scores based on fewer observations. This tenet is clearly not confirmed by the random model results in Table 5. For example, with $n'_t = 5$ and $n'_r = 3$ (i.e., 15 observations) $\widehat{\Phi} = .70$, but with $n'_t = 10$ and $n'_r = 1$ (i.e., 10 observations) $\widehat{\Phi} = .80$. Why? The reason is that the task and rater facets make different contributions to error variance. So doubling the number of tasks leads to a greater reduction in error variance than does tripling the number of raters.

3.4 Rater Reliability Issues

In many areas of scientific endeavor, certain types of reliability issues are addressed using some statistic that reflects the similarity of ratings. Here, this matter is considered from the perspective of correlation coefficients and their generalizability theory counterparts.

3.4.1 Interrater Reliability

Interrater coefficients are a frequently discussed type of reliability. Actually, there are at least two interrater coefficients, both of which are often misunderstood:

- *Standardized:* correlation between the scores assigned by the same two raters to student responses to the *same* task; and
- *Nonstandardized:* correlation between the scores assigned by the same two raters to student responses to *different* tasks.

The magnitudes of standardized coefficients are often quite high, while nonstandardized coefficients tend to be small. As discussed by Brennan (2000), these results are predictable from a careful consideration of the D study designs, sample sizes, and universes of generalization that are implicit in these two coefficients.

Letting t stand for tasks and r stand for raters, the standardized coefficient uses the G study $p \times t \times r$ design with $n_r = 2$ raters evaluating the same $n_t = 1$ task; or, since there is only one task, the design could be specified as $p \times (r{:}t)$. For the D study, the design is $p \times t \times r$ [or $p \times (r{:}t)$]with a single *random* rater and a single *fixed* task. It follows that the standardized coefficient is

$$\boldsymbol{E}\rho^2 = \frac{\sigma^2(p) + \sigma^2(pt)}{\sigma^2(p) + \sigma^2(pt) + [\sigma^2(pr) + \sigma^2(ptr)]}. \tag{40}$$

For this coefficient, $n'_t = 1$ because only one task is involved in the correlation, and $n'_r = 1$ because a correlation between two raters gives an estimate of reliability for a single rater.

The nonstandardized coefficient has tasks nested within persons, which means that it effectively uses the G study $(t{:}p) \times r$ design with $n_r = 2$ raters evaluating a different task for each person. For the D study, the design is $(t{:}p) \times r$ with a single *random* rater and a single *random* task. It follows that the nonstandardized coefficient is

$$\boldsymbol{E}\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p) + [\sigma^2(t{:}p) + \sigma^2(pr) + \sigma^2(tr{:}p)]}, \tag{41}$$

where $\sigma^2(t{:}p) = \sigma^2(t) + \sigma^2(pt)$ and $\sigma^2(tr{:}p) = \sigma^2(tr) + \sigma^2(ptr)$.[7] The nonstandardized coefficient is smaller than the standardized coefficient in Equation 40 for two reasons: (a) universe score variance for the nonstandardized coefficient is smaller because it does not contain $\sigma^2(pt)$; and (b) relative error variance for the nonstandardized coefficient is larger by $\sigma^2(t) + \sigma^2(pt) + \sigma^2(tr)$.

3.4.2 Intrarater Reliability

Another frequently discussed type of reliability is intrarater reliability, which is usually obtained by correlating the scores for the same rater obtained on two different occasions o (or replications). From the perspective of generalizability theory, the G study design is $p \times r \times o$; or, since there is only one rater, the design could be specified as $p \times (o{:}r)$.

For the D study, the design is $p \times r \times o$ [or $p \times (o{:}r)$] with a single *random* occasion and a single *fixed* rater. It follows that intrarater reliability is

$$\boldsymbol{E}\rho^2 = \frac{\sigma^2(p) + \sigma^2(pr)}{\sigma^2(p) + \sigma^2(pr) + [\sigma^2(po) + \sigma^2(pro)]}. \tag{42}$$

[7] That is, $\sigma^2(t{:}p)$ in the $(t{:}p) \times r$ design is the sum of $\sigma^2(t)$ and $\sigma^2(pt)$ in the $p \times t \times r$ design; similarly $\sigma^2(tr{:}p)$ is the sum of $\sigma^2(tr)$ and $\sigma^2(ptr)$.

For this coefficient, $n'_r = 1$ because only one rater is involved in the correlation, and $n'_o = 1$ because a correlation between two occasions gives an estimate of reliability for a single occasion.

3.4.3 Comparing Interrater (Standardized) and Intrarater Reliability

Clearly, in one sense, Equation 42 directly parallels Equation 40 in that the former can be obtained from the latter by replacing t and r in Equation 40 with r and o, respectively. This similarity in form, however, does not permit us to judge which type of rater reliability is larger. The basic problem is that Equation 40 for interrater reliability (standardized) does not explicitly represent the role of occasions on which the ratings were obtained, and Equation 42 for intrarater reliability does not explicitly represent the role of the tasks that were evaluated by the raters. So, to compare the two, we need to conceptualize a universe that involves all three facets—tasks, raters, and occasions of rating.[8] Also, we need to know more about the role of occasions for the interrater reliability analysis, and we need to know more about the role of tasks for the intrarater reliability analysis.

For example, suppose that, for the interrater analysis, each rating is collected on a *different* occasion; and for the intrarater analysis, the rater evaluates persons' responses to the *same* task. If so, it can be shown that the denominators of the generalizability coefficients will be the same, but the numerators (universe score variances) will be different. Specifically,

$$\text{Interrater: } \sigma^2(\tau) = \sigma^2(p) + \sigma^2(pt),$$

which is the same as the numerator of Equation 40; but

$$\text{Intrarater: } \sigma^2(\tau) = \sigma^2(p) + \sigma^2(pt) + \sigma^2(pr) + \sigma^2(ptr),$$

is clearly bigger than the numerator of Equation 42. In follows immediately that the intrarater coefficient will be larger. The reason for this is that:

- for the (standardized) interrater reliability coefficient with each rating collected on a different occasion, occasions are completely confounded with the ratings, which are effectively treated as a random facet when the correlation is computed; and
- for the intrarater reliability coefficient in which the rater evaluates persons' responses to the same task, task is completely confounded with rater, which is treated as fixed when the correlation is computed.

These are not the only possibilities, however. For example, for the interrater analysis, if each rating is collected on the *same* occasion, then it can be shown

[8] An even richer analysis would take into account not only variability attributable to the occasions on which the ratings were obtained, but also variability attributable to the occasions on which the tasks were administered.

that the correlation coefficient effectively treats universe score variance as:

$$\text{Interrater: } \sigma^2(\tau) = \sigma^2(p) + \sigma^2(pt) + \sigma^2(po) + \sigma^2(pto).$$

Also, for the intrarater analysis, if each person responds to a *different* task, then the correlation coefficient effectively treats universe score variance as:

$$\text{Intrarater: } \sigma^2(\tau) = \sigma^2(p) + \sigma^2(pr).$$

It is evident that correlation coefficients provide a deceptively simple (and sometimes quite misleading) basis for making statements about whether interrater reliability is bigger or smaller than intrarater reliability. Furthermore, the manner in which data are collected can "cause" a computed correlation coefficient to misrepresent the intended universe of generalization. For example, if an investigator estimates interrater reliability by computing a correlation coefficient using a data collection design in which each rating is collected on the same occasion, then occasion will be treated as if it were fixed in the universe of generalization, which may or may not be what the investigator intends. If not, then the investigator is well advised to use a generalizability coefficient in which $\sigma^2(\tau) = \sigma^2(p) + \sigma^2(pt)$. Indeed, a generalizability theory is almost always more informative than a correlation coefficient for quantifying interrater and/or intrarater reliability.

3.4.4 Important Caveats

The foregoing discussion of rater reliability issues may be complicated, but it is still somewhat idealized and incomplete. In real situations, almost inevitably additional complexities will arise. However, under any circumstances, it is important to identify clearly:

- the facets in the intended universe of generalization, with particular attention to which of them are fixed, and which of them are random; and
- the data collection design, with particular attention to which facets are effectively fixed (i.e., constant for all observations) and which are effectively random (i.e., allowed to vary).

Any reported indices or coefficients should reflect the intended universe of generalization; if not, bias will be introduced. In this regard, using correlation coefficients as indices of interrater and intrarater reliability can be misleading if they are not interpreted very carefully.

Carefully considered rater reliability coefficients, in the sense of correlation coefficients or their generalizability counterparts, can be of value in evaluating the extent to which raters are functioning as intended or desired. However, except in trivial cases, such coefficients usually do *not* characterize the reliability of student mean or total scores. In particular, in most cases, investigators want to generalize student scores over tasks, which means that $\sigma^2(pt)$ should be part of error variance, not universe score variance. Furthermore, for generalizability coefficients for student scores, variance components containing r should be

divided by the actual number of raters used in a D study (n'_r), and variance components containing t should be divided by the actual number of tasks used in a D study (n'_t). It is quite common for rater reliability coefficients (especially interrater coefficients) to be relatively high, when generalizability coefficients for student scores are relatively low. One perspective on this is discussed next.

3.5 Reliability-Validity Paradox

In an extensive consideration of validity from the perspective of generalizability theory, Kane (1982) argues that a restricted universe of generalization (what he calls a universe of allowable observations) for a standardized measurement procedure can be conceptualized as a reliability-defining universe, while the broader universe of generalization can be considered a validity-defining universe. Doing so provides an elegant explanation of the reliability–validity paradox, whereby attempts to increase reliability through standardization (i.e., fixing facets) can actually lead to a decrease in some measures of validity (Lord & Novick, 1968, p. 334). (See also Brennan (2001b, pp. 132–135.)

As an example, let us return to the issue of interrater reliability. Let t (i.e., task) be a fixed facet in the *restricted* universe, and let r (i.e., rater) be a random facet. Universe scores for the restricted universe are

$$\mu_{pt} = \underset{r}{\boldsymbol{E}} X_{ptr},$$

whereas universe scores for the unrestricted universe are

$$\mu_p = \underset{t}{\boldsymbol{E}} \underset{r}{\boldsymbol{E}} X_{ptr}.$$

That is, in the unrestricted universe, the t facet is treated as random. In Kane's terminology, inferences from X_{ptr} to μ_{pt} are in the realm of reliability, while inferences from X_{ptr} to μ_p relate to validity.

From this point of view, the interrater reliability coefficient discussed previously is

$$\boldsymbol{E}\rho^2(X_{ptr}, \mu_{pt}) = \frac{\sigma^2(p) + \sigma^2(pt)}{\sigma^2(p) + \sigma^2(pt) + [\sigma^2(pr) + \sigma^2(ptr)]}, \tag{43}$$

the squared validity coefficient is

$$\boldsymbol{E}\rho^2(X_{ptr}, \mu_p) = \frac{\sigma^2(p)}{\sigma^2(p) + [\sigma^2(pr) + \sigma^2(pt) + \sigma^2(ptr)]}, \tag{44}$$

and the squared disattenuated validity coefficient is

$$\boldsymbol{E}\rho^2(\mu_{pt}, \mu_p) = \frac{\boldsymbol{E}\rho^2(X_{ptr}, \mu_p)}{\boldsymbol{E}\rho^2(X_{ptr}, \mu_{pt})} = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(pt)}. \tag{45}$$

The three coefficients in Equations 43–45 have a simple relationship:

$$\boldsymbol{E}\rho^2(X_{ptr}, \mu_p) = \boldsymbol{E}\rho^2(X_{ptr}, \mu_{pt}) \times \boldsymbol{E}\rho^2(\mu_{pt}, \mu_p). \tag{46}$$

The reliability–validity paradox arises because of the differential role played by $\sigma^2(pt)$ in the coefficients in Equation 46. Specifically, as $\sigma^2(pt)$ gets larger,

- reliability *increases* because $\sigma^2(pt)$ contributes to universe score variance for the restricted universe, and
- validity *decreases* because $\sigma^2(pt)$ contributes to relative error variance for the broader universe of generalization.

Note also that, as $\sigma^2(pt)$ gets larger, the squared disattenuated validity coefficient decreases, which weakens inferences from the restricted universe to the more broadly defined universe of generalization.

From Kane's perspective on the reliability–validity paradox, it makes sense to call Equation 44 a squared validity coefficient. Clearly, however, Equation 44 is a generalizability coefficient for a universe of generalization in which both raters and tasks are random. This terminological ambiguity illustrates that generalizability theory "blurs" arbitrary distinctions between reliability and validity (Cronbach et al., 1972, p. 380; Brennan, 2001a) and forces an investigator to concentrate on the intended inferences, whatever terms are used to characterize them.

4 Multivariate Designs

Multivariate generalizability theory is treated extensively by Brennan (2001b, chaps. 9–12). It is very powerful and flexible, but more complicated then univariate generalizability theory. The fundamental difference between multivariate and univariate generalizability theory is that in multivariate generalizability theory each of the levels of one or more fixed facets is treated as a distinct variable. Here, we simply illustrate how multivariate generalizability theory might be used, primarily in the context of difference scores using a synthetic data example. The ultimate focus here is on indices and coefficients that might be reported.

Consider the synthetic data at the top of Table 6 for a multivariate $p^\bullet \times i^\bullet$ design.[9] These data might be viewed as the responses of ten persons (p) to each of six items or tasks (i) with each such response evaluated according to two fixed criteria (v). For example, in job analyses often experts (i.e., persons) are asked to evaluate certain tasks with respect to their frequency of occurrence (v_1) and their criticality (v_2).

The bottom left-hand part of Table 6 reports the estimated variance and covariance components in three matrices, $\widehat{\mathbf{\Sigma}}_p$, $\widehat{\mathbf{\Sigma}}_i$, and $\widehat{\mathbf{\Sigma}}_{pi}$. The estimated variance components were obtained using Equations 8–10. The estimated covariance components were obtained using similar equations, with mean-products replacing mean squares (see Brennan, 2001b, pp. 289–293).

The bottom right-hand part of Table 6 reports variance-covariance matrices for universe scores ($\widehat{\mathbf{\Sigma}}_p$), δ-type errors ($\widehat{\mathbf{\Sigma}}_\delta$), Δ-type errors ($\widehat{\mathbf{\Sigma}}_\Delta$), and (expected) observed variances and covariances ($\mathbf{S}$) for a D study sample size of $n_i' = 6$. The

[9] Filled circles indicate "linked facets" in the sense that every person has scores on both levels of v, and every item is scored with respect to both levels of v.

Table 6: Synthetic Data Example for Balanced $p^\bullet \times i^\bullet$ Design

	v_1						v_2								
p	i_1	i_2	i_3	i_4	i_5	i_6	i_1	i_2	i_3	i_4	i_5	i_6	$\overline{X}_{p1}$	$\overline{X}_{p2}$	$\overline{X}_{p2} - \overline{X}_{p1}$
1	6	4	3	5	4	4	6	4	5	6	4	5	4.3333	5.0000	.6667
2	3	2	2	4	5	5	6	4	5	6	2	5	3.5000	4.6667	1.1667
3	6	5	7	5	4	3	6	5	8	4	6	3	5.0000	4.8333	−.1667
4	4	2	2	3	3	5	5	4	2	4	3	5	3.1667	3.8333	.6667
5	4	4	3	5	4	6	4	5	5	3	3	7	4.3333	4.5000	.1667
6	8	5	4	7	5	4	9	6	6	5	7	7	5.5000	6.6667	1.1667
7	5	4	5	7	4	4	4	5	6	7	4	5	4.8333	5.1667	.3333
8	4	5	3	4	5	6	6	7	6	6	5	6	4.5000	6.0000	1.5000
9	7	5	4	6	6	5	6	7	3	5	5	6	5.5000	5.3333	−.1667
10	5	3	3	7	4	5	6	5	4	6	2	6	4.5000	4.8333	.3333
Mean													4.5167	5.0833	.5667

$$\widehat{\boldsymbol{\Sigma}}_p = \begin{bmatrix} .3682 & .3193 \\ .3193 & .3689 \end{bmatrix} \qquad \widehat{\boldsymbol{\Sigma}}_p = \begin{bmatrix} .3682 & \textit{.8663} \\ .3193 & .3689 \end{bmatrix}$$

$$\widehat{\boldsymbol{\Sigma}}_i = \begin{bmatrix} .3444 & .0919 \\ .0919 & .3200 \end{bmatrix} \qquad \widehat{\boldsymbol{\Sigma}}_\delta = \begin{bmatrix} .2087 & \textit{.5081} \\ .1175 & .2561 \end{bmatrix}$$

$$\widehat{\boldsymbol{\Sigma}}_{pi} = \begin{bmatrix} 1.2522 & .7048 \\ .7048 & 1.5367 \end{bmatrix} \qquad \widehat{\boldsymbol{\Sigma}}_\Delta = \begin{bmatrix} .2661 & \textit{.4627} \\ .1328 & .3094 \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} .5769 & .4367 \\ .4367 & .6250 \end{bmatrix}$$

last three matices are easily obtained from $\widehat{\boldsymbol{\Sigma}}_p$, $\widehat{\boldsymbol{\Sigma}}_i$, and $\widehat{\boldsymbol{\Sigma}}_{pi}$ as follows:

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_\delta &= \frac{1}{6}\widehat{\boldsymbol{\Sigma}}_i \\ \widehat{\boldsymbol{\Sigma}}_\Delta &= \frac{1}{6}(\widehat{\boldsymbol{\Sigma}}_i + \widehat{\boldsymbol{\Sigma}}_{pi}) \\ \mathbf{S} &= \widehat{\boldsymbol{\Sigma}}_p + \widehat{\boldsymbol{\Sigma}}_\delta \end{aligned}$$

The upper diagonal elements in italics are disattenuated correlations. So, for example, the estimated correlation between the universe scores for the two levels of v is .8663; and the estimated correlation between the δ-type errors is .5081, which is a direct indication of correlated error.[10]

[10]The correlation for the $\mathbf{S}$ matrix is .7272. This is the expected value of the correlation between *observed* scores for the two levels of v assuming $n'_i = 6$.

4.1 Composites

For difference scores, the composite universe score of interest is

$$\mu_{pC} = \mu_{p2} - \mu_{p1}, \tag{47}$$

and the simplest and most obvious estimator of a universe difference score is

$$\overline{X}_{pC} = \overline{X}_{p2} - \overline{X}_{p1}. \tag{48}$$

It follows that the composite universe score variances and error variances are:

$$\begin{aligned} \sigma^2_C(p) &= \sigma^2_1(p) + \sigma^2_2(p) - 2\,\sigma_{12}(p) \\ \sigma^2_C(\delta) &= \sigma^2_1(\delta) + \sigma^2_2(\delta) - 2\,\sigma_{12}(\delta) \\ \sigma^2_C(\Delta) &= \sigma^2_1(\Delta) + \sigma^2_2(\Delta) - 2\,\sigma_{12}(\Delta) \end{aligned}$$

For the synthetic data example, it is easily verified that

$$\hat{\sigma}^2_C(p) = .0985, \quad \hat{\sigma}^2_C(\delta) = .2298, \quad \text{and} \quad \hat{\sigma}^2_C(\Delta) = .3099.$$

The equations in Table 2 on page 21 can now be used to obtain estimates of the various coefficients and indices. For example, the estimated generalizability coefficient for the difference score composite is

$$\boldsymbol{E}\hat{\rho}^2 = \frac{\hat{\sigma}^2_C(p)}{\hat{\sigma}^2_C(p) + \hat{\sigma}^2_C(\delta)} = \frac{.0985}{.0985 + .2298} = .30$$

Note that positively correlated errors [$\sigma_{12}(\delta) = .1175$ and $\rho_{12}(\delta) = .5081$ for this example] decrease composite error variance relative to what it would be if errors were uncorrelated, which means that $\boldsymbol{E}\hat{\rho}^2$ will be higher when errors are positively correlated.

Still, $\boldsymbol{E}\hat{\rho}^2 = .30$ for the difference scores is considerably smaller than the generalizability coefficients for the component parts ($\boldsymbol{E}\hat{\rho}^2 = .64$ for v_1 and $\boldsymbol{E}\hat{\rho}^2 = .59$ for v_1). This type of result is an often-cited criticism of difference scores. This criticism, however, is frequently unwarranted, in part because it fails to take into account the magnitude of the average difference score, which could be huge without having any impact on $\boldsymbol{E}\rho^2$. (See Yin & Brennan, 2002 for a more detailed discussion and list of additional references.) For example, suppose our interest focuses on difference scores that are different from 0. Setting $\lambda = 0$, we might compute the error-tolerance ratio:

$$Est[E_\Delta/T(\lambda = 0)] = \sqrt{\frac{\hat{\sigma}^2(\Delta)}{\hat{\sigma}^2(p) + \widehat{\mu^2}}} = \sqrt{\frac{.3099}{.0985 + .5667^2 - .1129}} = 1.0052,$$

This error-tolerance ratio can be transformed to

$$\widehat{\Phi} = \frac{1}{1 + \{Est[E_\Delta/T(\lambda = 0)]\}^2} = \frac{1}{1 + 1.0052^2} = .50,$$

which is substantially higher than $\boldsymbol{E}\hat{\rho}^2 = .30$ for the difference scores. In this sense, we can much more reliably detect differences from 0 than we can rank order examinees on the basis of difference scores.

Table 7: Observed Profile Variability for Synthetic Data Example of $p^\bullet \times I^\bullet$ Design with $n'_i = 6$

p	$\overline{X}_{p1}$	$\overline{X}_{p2}$	Mean	Var^a
1	4.3333	5.0000	4.6667	.1111
2	3.5000	4.6667	4.0833	.3403
3	5.0000	4.8333	4.9167	.0069
4	3.1667	3.8333	3.5000	.1111
5	4.3333	4.5000	4.4167	.0069
6	5.5000	6.6667	6.0833	.3403
7	4.8333	5.1667	5.0000	.0278
8	4.5000	6.0000	5.2500	.5625
9	5.5000	5.3333	5.4167	.0069
10	4.5000	4.8333	4.6667	.0278
Mean	4.5167	5.0833	4.8000	.1542
Var^a	.5192	.5625	.4669	.0337

$$\widehat{\boldsymbol{\Sigma}}_p = \begin{bmatrix} .3682 & .8663 \\ .3193 & .3689 \end{bmatrix}$$

$$\widehat{\boldsymbol{\Sigma}}_\delta = \begin{bmatrix} .2087 & .5081 \\ .1175 & .2561 \end{bmatrix}$$

$$\widehat{\boldsymbol{\Sigma}}_\Delta = \begin{bmatrix} .2661 & .4627 \\ .1328 & .3094 \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} .5769 & .4367 \\ .4367 & .6250 \end{bmatrix}$$

^{a}Biased estimates.

4.2 Profiles

Whether or not composite scores are of interest, an investigator may be concerned about profiles of the n_v scores for the objects of measurement. For the synthetic data example in Table 6, the observed score profile for a person is simply the mean scores for the two variables, which are reported in Table 7 along with the mean and variance of the two mean scores for each person. Obviously, for each person, the observed profile contains some unknown amount of error, and the true profile for a person is unknown. However, we can consider expected (over persons) within-person profile variability, which we denote generically as $\mathcal{V}(*)$. For example, $\mathcal{V}(\overline{\mathbf{X}}_p)$ is the expected within-person profile variability for persons' observed mean scores, and $\mathcal{V}(\boldsymbol{\mu}_p)$ is the expected within-person profile variability for universe scores.

The $\mathcal{V}(*)$ formulas discussed in this section result from the well-known analysis of variance identity; namely, that total variance equals within variance plus between variance. For raw scores, $\overline{X}_{pv}$, Brennan (2001b, p. 321) shows that

$$\mathcal{V}(\overline{\mathbf{X}}_p) = \boldsymbol{E}_p \left[\underset{v}{\text{var}}(\overline{X}_{pv}) \right] = [\overline{S}^2_v(p) - \overline{S}_{vv'}(p)] + \text{var}(\overline{X}_v), \tag{49}$$

where $\overline{S}^2_v(p)$ is the average over the n_v observed variances, and $\overline{S}_{vv'}(p)$ is the average over *all* n_v^2 elements of the observed variance-covariance matrix, which includes those with $v = v'$. If the term in square brackets is multiplied by $(n_p - 1)/n_p$, the value obtained using Equation 49 is identical to the value obtained through directly computing the average of the variances of the n_p profiles for a finite number of persons. Of course, this multiplicative factor approaches unity as $n_p \to \infty$.

For the synthetic data, the average of the two observed variances in the $\mathbf{S}$ matrix given in Table 6 is $\overline{S}_v^2(p) = .6010$, the average of all elements in $\mathbf{S}$ is $\overline{S}_{vv'}(p) = .5188$, and the variance (biased estimate) of the two means (4.5167 and 5.0833) is .0803. Therefore,

$$\hat{\mathcal{V}}(\overline{\mathbf{X}}_p) = (.6010 - .5188) + .0803 = .1625.$$

If the term in parentheses is multiplied by $(n_p - 1)/n_p$, $\hat{\mathcal{V}}(\overline{\mathbf{X}}_p) = .1542$, which is identical to the value reported in Table 7 based on direct computation of the average of the observed profile variances for the 10 persons.

Using the same logic, universe-score profile variability is

$$\mathcal{V}(\boldsymbol{\mu}_p) = \boldsymbol{E}_p \left[\underset{v}{\text{var}}(\mu_{pv})\right] = [\overline{\sigma}_v^2(p) - \overline{\sigma}_{vv'}(p)] + \text{var}(\mu_v). \tag{50}$$

For the synthetic data example, the average of the two universe score variances in $\widehat{\boldsymbol{\Sigma}}_p$ is .3686, the average of all four variance-covariance elements is .3439, an estimate of $\text{var}(\mu_v)$ is $\text{var}(\overline{X}_v) = .0803$, and the universe-score profile variability is

$$\hat{\mathcal{V}}(\boldsymbol{\mu}_p) = (.3686 - .3439) + .0803 = .1050.$$

Obviously, since universe scores are unknown, this result cannot be obtained through direct computation using persons' universe scores.

The variance of the δ-type errors for a randomly selected person is

$$\mathcal{V}(\boldsymbol{\delta}_p) = \boldsymbol{E}_p \left[\underset{v}{\text{var}}(\delta_{pv})\right] = \overline{\sigma}_v^2(\delta) - \overline{\sigma}_{vv'}(\delta). \tag{51}$$

For the synthetic data example, $\hat{\mathcal{V}}(\boldsymbol{\delta}_p) = .0575$, with a corresponding standard error of measurement of $\sqrt{.0575} = .24$. Similarly, we can express the variance of the Δ-type errors for a typical person as

$$\mathcal{V}(\boldsymbol{\Delta}_p) = \boldsymbol{E}_p \left[\underset{v}{\text{var}}(\Delta_{pv})\right] = \overline{\sigma}_v^2(\Delta) - \overline{\sigma}_{vv'}(\Delta), \tag{52}$$

which is $\hat{\mathcal{V}}(\boldsymbol{\Delta}_p) = .0775$, with a corresponding standard error of measurement of $\sqrt{.0775} = .28$.

As in univariate theory, it is natural to consider functions of variabilities, particularly ratios. One such ratio is

$$\mathcal{G} = \frac{\mathcal{V}(\boldsymbol{\mu}_p)}{\mathcal{V}(\overline{\mathbf{X}}_p)} = \frac{\mathcal{V}(\boldsymbol{\mu}_p)}{\mathcal{V}(\boldsymbol{\mu}_p) + \mathcal{V}(\boldsymbol{\delta}_p)}, \tag{53}$$

which is the proportion of the variance in the profile of observed scores for a typical person that is attributable to the variance in the profile of universe scores for such a person. If $\mathcal{V}(\boldsymbol{\mu}_p)$ is viewed as a measure of the flatness of the profile of universe scores, and $\mathcal{V}(\overline{\mathbf{X}}_p)$ is viewed as a measure of the flatness of the profile of observed scores, then $\mathcal{G}$ is a measure of the relative flatness of these profiles for a typical person.

$\mathcal{G}$ is also interpretable approximately as a type of generalizability coefficient for a randomly selected person p. For a specified person, we can define a person-level generalizability coefficient as the ratio of $\text{var}(\mu_{pv})$ to $\text{var}(\overline{X}_{pv})$, where the variance is taken over levels of v. Obviously, this ratio is not estimable for a given person, because universe scores are unknown. The expected value, over persons, of this ratio would be the expected generalizability coefficient for a randomly selected person. We approximate this expected value over persons with the ratio of the expected values in Equation 53. For the synthetic data with $n_i' = 6$, $\hat{\mathcal{G}} = .1050/.1625 = .65$, which suggests that, on average (i.e., for a typical person), 65% of the variance in observed mean scores for the n_v variables is attributable to variance in universe scores.

We might also consider an error-tolerance ratio. Suppose our tolerance for error is relative to the profile of population mean scores for the two variables. Then, we might define tolerance as $\sqrt{\text{var}_v(\mu_v)}$; i.e., tolerance is the standard deviation of the mean scores for the variables. Then,

$$E_\Delta/T = \frac{\mathcal{V}(\boldsymbol{\Delta}_p)}{\sqrt{\text{var}_v(\mu_v)}}$$

4.3 Regressed Estimates

A well-known result from classical test theory is the formula for Kelley's (1947) regressed score estimates of true scores:

$$\hat{\mu}_p = (1 - \boldsymbol{E}\rho^2)\mu + \boldsymbol{E}\rho^2\overline{X}_p. \tag{54}$$

Feldt and Brennan (1989, pp. 120–121) provide a derivation using simple linear regression in conjunction with the classical test theory results that the mean of true scores equals the mean of observed scores, and true score variance is reliability times observed score variance. Brennan (2001b) discusses in detail how multivariate generalizability theory can be used to extend further these basic ideas to the estimation of universe score profiles and universe score composites. Here we focus on the relatively simple two-variable case in the synthetic data in Table 6.

4.3.1 Estimating Profiles Through Regression

When $n_v = 2$, we need to obtain two prediction equations: one for v_1 and one for v_2. For raw scores, these two equations can be represented as

$$\hat{Y}_v = b_{0v} + b_{1v}X_{1v} + b_{2v}X_{2v}. \tag{55}$$

Note that we use information from the second/first variable to help predict universe scores for the first/second variable.

In a typical multiple regression, observed values are available for both the dependent and independent variables. That is *not* true for the application of multiple regression to the estimation of universe scores, because universe scores

play the role of Y and are obviously unknown. At first blush, this may seem to present an insurmountable problem, because without having values for Y we cannot directly compute the covariances involving Y in the normal equations. With the models we are using, however, we can determine the covariances even though we do not know the individual scores. For example, when $n_v = 2$, the two sets of normal equations can be represented as:

$$\begin{array}{rcrcl} S_1^2(p)\, b_{1v} & + & S_{12}(p)\, b_{2v} & = & \sigma_{1v}(p) \\ S_{12}(p)\, b_{1v} & + & S_2^2(p)\, b_{2v} & = & \sigma_{2v}(p), \end{array} \tag{56}$$

where $b_{0v} = \mu_v - b_{1v}\,\overline{X}_{1v} - b_{2v}\,\overline{X}_{2v}$. If $v = 1$, then $\sigma_{1v}(p) = \sigma_1^2(p)$ and $\sigma_{2v}(p) = \sigma_{12}(p)$. Similarly, if $v = 2$, $\sigma_{1v}(p) = \sigma_{12}(p)$ and $\sigma_{2v}(p) = \sigma_2^2(p)$. In short, the right-hand side of the normal equations are elements of $\mathbf{\Sigma}_p$, which are estimable using previously discussed procedures; and the multipliers of the regression coefficients on the left-hand side are elements of $\mathbf{S}$.

For variable v, the proportion of the universe score variance that is explained by the regression is

$$R_v^2 = \frac{\sigma_v^2(\hat{\mu}_p)}{\sigma_v^2(p)}, \tag{57}$$

where

$$\sigma_v^2(\hat{\mu}_p) = \sum_{j=1}^{n_v} b_{jv}\sigma_{vj}(p); \tag{58}$$

and letting $\mathcal{E} = \hat{\mu}_p - \mu_p$, the standard error of estimate is

$$\sigma_v(\mathcal{E}) = \sigma_v\sqrt{1 - R_v^2}. \tag{59}$$

Further, the covariance of the regressed score estimates for v and v' is

$$\sigma_{vv'}(\hat{\mu}_p) = \sum_{j=1}^{n_v} b_{jv'}\sigma_{vj}(p) = \sum_{j=1}^{n_v} b_{jv}\sigma_{v'j}(p). \tag{60}$$

The expressions for variances and covariances of regressed scores in equations 58 and 60, respectively, do not apply in multiple regression generally. Rather, these simplified expressions are a consequence of the fact that, for regressed score estimates, the dependent variable is the universe score for one of the independent variables. A unique and useful feature of these expressions is that they depend only on the regression weights and the elements of $\mathbf{\Sigma}_p$.

For this synthetic data example, the regressed-score profile equations can be shown to be

$$\begin{array}{rcl} \hat{\mu}_{p1} & = & 1.4050 + .5339\,\overline{X}_{p1} + .1377\,\overline{X}_{p2} \\ \hat{\mu}_{p2} & = & 1.8647 + .2263\,\overline{X}_{p1} + .4321\,\overline{X}_{p2}. \end{array}$$

The actual values are provided in Table 8. The estimated variances and covariance of the regressed score estimates are

$$\hat{\sigma}_1^2(\hat{\mu}_p) = .2405, \quad \hat{\sigma}_2^2(\hat{\mu}_p) = .2317, \quad \text{and} \quad \hat{\sigma}_{12}(\hat{\mu}_p) = .2213;$$

Table 8: Regressed Score Estimates for Synthetic Data Example

	Raw Scores				Regressed-Score Estimates			
p	$\overline{X}_{p1}$	$\overline{X}_{p2}$	Mean	Var[a]	$\hat{\mu}_{p1}$	$\hat{\mu}_{p2}$	Mean	Var[a]
1	4.3333	5.0000	4.6667	.1111	4.4073	5.0058	4.7066	.0896
2	3.5000	4.6667	4.0833	.3403	3.9165	4.6732	4.2948	.1432
3	5.0000	4.8333	4.9167	.0069	4.7403	5.0847	4.9125	.0297
4	3.1667	3.8333	3.5000	.1111	3.6237	4.2377	3.9307	.0943
5	4.3333	4.5000	4.4167	.0069	4.3384	4.7898	4.5641	.0509
6	5.5000	6.6667	6.0833	.3403	5.2598	5.9900	5.6249	.1333
7	4.8333	5.1667	5.0000	.0278	4.6972	5.1910	4.9441	.0610
8	4.5000	6.0000	5.2500	.5625	4.6340	5.4756	5.0548	.1771
9	5.5000	5.3333	5.4167	.0069	5.0761	5.4139	5.2450	.0285
10	4.5000	4.8333	4.6667	.0278	4.4733	4.9716	4.7224	.0621
Mean	4.5167	5.0833	4.8000	.1542	4.5167	5.0833	4.8000	.0870
Var[a]	.5192	.5625	.4669	.0337	.2165	.2085	.2058	.0023

[a]Biased estimates.

the estimated R^2 values are

$$\hat{R}_1^2 = \frac{.2404}{.3682} = .6534, \qquad \text{and} \qquad \hat{R}_2^2 = \frac{.2317}{.3689} = .6280;$$

and the estimated standard errors of estimate are

$$\hat{\sigma}_1(\mathcal{E}) = .3572, \qquad \text{and} \qquad \hat{\sigma}_2(\mathcal{E}) = .3705.$$

It is evident from Table 8 that regression occurs in the sense that the profiles of predicted universe scores are generally flatter than the profiles of observed (raw) scores. Specifically, the average within-person variance for the regressed scores (.0870) is less than the average within-person variance for the observed scores (.1542), which means that the regressed scores are about 44% less variable than the observed scores.

To characterize the entire measurement procedure for a population we can consider expected within-person profile variability using $\mathcal{V}(*)$ formulas such as $\mathcal{V}(\overline{\mathbf{X}}_p)$ in Equation 49 and $\mathcal{V}(\boldsymbol{\mu}_p)$ in Equation 50. A corresponding formula for $\hat{\mu}_p$ is

$$\mathcal{V}(\hat{\boldsymbol{\mu}}_p) = [\overline{\sigma}_v^2(\hat{\mu}_p) - \overline{\sigma}_{vv'}(\hat{\mu}_p)] + \text{var}(\hat{\mu}_v), \tag{61}$$

which is the expected regressed score profile variability in the sense of the variability of regressed score estimates for a "typical" person. We take $\mathcal{V}(\hat{\boldsymbol{\mu}}_p)$ as a definition of expected profile flatness or, more specifically, lack of flatness for regressed score estimates. For the synthetic data,

$$\mathcal{V}(\hat{\boldsymbol{\mu}}_p) = (.2361 - .2287) + .0803 = .0877.$$

Table 8 reports that the average (over persons) of the variances of the regressed score estimates is .0870, which is the value obtained by multiplying the term (.2361 − .2287) by the "correction factor" $(n_p - 1)/n_p = .9$.

It can be shown that

$$\mathcal{V}(\hat{\boldsymbol{\mu}}_p) \leq \mathcal{V}(\boldsymbol{\mu}_p) \leq \mathcal{V}(\overline{\mathbf{X}}_p),$$

which mirrors the inequality relationships in univariate theory[11]; that is,

$$\sigma^2(\hat{\mu}_p) \leq \sigma^2(\mu_p) \leq \sigma^2(\overline{X}_p).$$

As in univariate theory, it is natural to consider various functions of variabilities, particularly ratios. One such ratio is $\mathcal{G}$, which was discussed previously. With regressed score estimates another obvious ratio to consider is

$$\mathcal{R}^2 = \frac{\mathcal{V}(\hat{\boldsymbol{\mu}}_p)}{\mathcal{V}(\boldsymbol{\mu}_p)}, \tag{62}$$

which we designate $\mathcal{R}^2$ because of its obvious similarity to the squared correlation in a multiple regression.[12] This ratio can be interpreted approximately as the proportion of expected variability in universe scores for a "typical" person that is explained by the regressions. For the synthetic data example, $\mathcal{R}^2 = .0877/.1050 = .835$. The corresponding approximation to the variance of the errors of estimate for a typical person is

$$\mathcal{V}(\boldsymbol{\mu}_p) - \mathcal{V}(\hat{\boldsymbol{\mu}}_p) = \mathcal{V}(\boldsymbol{\mu}_p)(1 - \mathcal{R}^2), \tag{63}$$

which is .015 for the synthetic data. This means that the standard error of the regressed score estimates for a typical person is roughly $\sqrt{.015} = .12$. Recall that the standard deviation of the δ-type errors of measurement for a typical person is $\sqrt{.058} = .24$, which is twice as large as the standard error of estimate. Such statements are sometimes used as an argument in favor of using regressed score estimates.

Finally, the proportional reduction in profile variability attributable to using regressed score estimates is

$$\mathcal{RV} = 1 - \frac{\mathcal{V}(\hat{\boldsymbol{\mu}}_p)}{\mathcal{V}(\overline{\mathbf{X}}_p)}, \tag{64}$$

which is $1 - .0877/.1625 = .460$ for the synthetic data. That is, for a typical person, the use of regressed score estimates reduces profile variability by about 46%.

[11] $\sigma^2(\overline{X}_p)$ is the variance of observed persons' mean scores for the population, which was denoted $\boldsymbol{E}S^2(p)$ earlier.

[12] This use of $\mathcal{R}$ should not be confused with the previous use of $\mathcal{R}$ to designate random facets.

4.3.2 Predicted Composites

Regressed-score estimation procedures can also be used for predicting composites of universe scores. When there are two independent variables, an obvious example of a composite is a difference score. In our notation, this composite is $\mu_{pC} = \mu_{p2} - \mu_{p1}$, and we wish to obtain the prediction equation

$$\hat{\mu}_{pC} = b_0 + b_1 \overline{X}_{p1} + b_2 \overline{X}_{p2}.$$

Letting $Y = \mu_{pC}$, X_1 be $\overline{X}_{p1}$, and X_2 be $\overline{X}_{p2}$, the b weights can be obtained by solving the following normal equations:

$$\begin{array}{rcrcl} S_1^2\, b_1 & + & S_{12}\, b_2 & = & S_{YX_1} \\ S_{12}\, b_1 & + & S_2^2\, b_2 & = & S_{YX_2}, \end{array} \tag{65}$$

where, for simplicity, "(p)" has been dropped from the S terms, and the terms on the right are covariances of $Y = \mu_{pC}$ and the two independent variables.

Without loss of generality, we can simplify matters further by using the classical test theory notation $Y = T_2 - T_1$, where $X_1 = T_1 + E_1$, $X_2 = T_2 + E_2$, and true scores (T) and error scores (E) are uncorrelated. It follows that the covariance of universe difference scores with X_1 is

$$S_{YX_1} = S_{(T_2-T_1)(T_1+E_1)} = S_{T_1T_2} - S_{T_1}^2 = \sigma_{12} - \sigma_1^2.$$

Similarly,

$$S_{YX_2} = S_{(T_2-T_1)(T_2+E_2)} = S_{T_2}^2 - S_{T_1T_2} = \sigma_2^2 - \sigma_{12}.$$

Using these results on the right-hand side of Equation Set 65 and solving for the bs gives

$$b_1 = \frac{1}{1 - r_{12}^2}\left[\frac{\sigma_{12}}{S_1^2} - \rho_1^2 - r_{12}\left(\frac{\sigma_2^2 - \sigma_{12}}{S_1 S_2}\right)\right] \tag{66}$$

and

$$b_2 = \frac{1}{1 - r_{12}^2}\left[\rho_2^2 - \frac{\sigma_{12}}{S_2^2} - r_{12}\left(\frac{\sigma_{12} - \sigma_1^2}{S_1 S_2}\right)\right], \tag{67}$$

where, again, "(p)" has been dropped from all terms.

The squared correlation between universe difference scores and their predicted values is

$$R_C^2 = \frac{b_1(\sigma_{12} - \sigma_1^2) + b_2(\sigma_2^2 - \sigma_{12})}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}, \tag{68}$$

where the denominator is the variance of the universe difference scores, $\sigma_C^2(\mu_p)$, and the numerator is the variance of their regressed score estimates. The standard error of estimate based on using the two observed scores to predict the universe difference scores is

$$\sigma_C(\mathcal{E}) = \sqrt{(1 - b_2)(\sigma_2^2 - \sigma_{12}) - (1 + b_1)(\sigma_{12} - \sigma_1^2)}. \tag{69}$$

For the synthetic data example, the predicted difference in universe scores is

$$\hat{\mu}_{pC} = .4597 - .3076\, \overline{X}_{p1} + .2944\, \overline{X}_{p2},$$

the estimated variance of the universe difference scores is

$$\hat{\sigma}^2_C(\mu_p) = .3682 + .3689 - 2(.3193) = .0985,$$

the estimated proportion of composite universe score variance explained by the regression is

$$\hat{R}^2_C = \frac{-.3076(.3193 - .3682) + .2944(.3689 - .3193)}{.0985} = .3009,$$

and the standard error of estimate is

$$\hat{\sigma}_C(\mathcal{E}) = \sqrt{.0985}\sqrt{1 - .3009} = .2624.$$

Recall that in the synthetic data example in Table 6, two scores are obtained from each examinee's response to each item. This "linking" of scores with items gives rise to correlated error, which is quantified by the covariance terms in $\widehat{\mathbf{\Sigma}}_\delta$ and $\widehat{\mathbf{\Sigma}}_\Delta$. Under the traditional assumptions of classical test theory, such covariances are assumed to be zero, which leads to

$$\sigma_{12} = S_{12}.$$

In this case, Equations 66 and 67 are the Lord–McNemar regression coefficients (see Lord, 1956, 1958; McNemar, 1958; Feldt & Brennan, 1989). For the derivation here, however, the classical test theory assumption of uncorrelated errors was not made, which leads to

$$\sigma_{12} = S_{12} - \sigma_{12}(\delta).$$

It follows that the prediction of universe difference scores based on Equations 66 to 67 is more general than the Lord–McNemar prediction. That is, the derivation provided here permits the consideration of designs in which correlated δ-type error may affect the prediction.

5 Concluding Comments

The primary purpose of this paper is to provide discussions and interpretations of the various coefficients and indices that have been proposed for use in generalizability theory. Generalizability coefficients, dependability coefficients, and signal-noise ratios have been in the literature for some time; the error-tolerance ratios and multivariate indices tht have been discussed here are much newer. Each of these coefficients and indices has its own interpretation, or set of interpretations. But, in a sense, they all share a common feature—they all provide a perspective on the relative size of some particular measure of error (usually an error variance). This is indicative of the fact that error variances, and the variance components that enter them, are of fundamental importance in generalizability theory. For this reason, picking one coefficient or index and simply reporting its value is seldom an adequate generalizability analysis. At

a minimum, estimated variance components and error variances should be reported, as well. Coefficients and indices can be very useful if they are interpreted correctly, but almost always they need to be buttressed with additional information—especially information about error variances.

6 References

Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice*, *16*(4), 14–20.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, *24*, 339–353.

Brennan, R. L. (2001a). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*, 295–317.

Brennan, R. L. (2001b). *Generalizability theory*. Springer-Verlag.

Brennan, R. L. & Kane, M. T. (1977a). An index of dependability for mastery tests. *Journal of Educational Measurement*, *14*, 277–289.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

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

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

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

Feldt, L. S. & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 105–146). New York: American Council on Education and MacMillan.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement*, *6*, 125–160.

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

Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement, 4*, 105–126.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151–160.

Llabre, M. M., Ironson, G. H., Spitzer, S. B., Gellman, M. D., Weidler, D. J.,

& Schneiderman, N. (1988). How many blood pressure measurements are enough?: An application of generalizability theory to the study of blood pressure reliability. *Psychophysiology*, *25*, 97–106.

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16*, 421–437.

Lord, F. M. (1958). Further problems in the measurement of growth. *Educational and Psychological Measurement, 18*, 437–451.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McNemar, Q. (1958). On growth measurement. *Educational and Psychological Measurement*, *18*, 47–55.

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

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

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

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

Yin, P., & Brennan, R. L. (2002). An investigation of difference scores for a grade-level testing program. *International Journal of Testing, 2*(2), 83–105.