

*Center for Advanced Studies in
Measurement and Assessment*

CASMA Research Report

Number 12

**Some Test Theory for the
Reliability of Individual Profiles**

Robert L. Brennan[†]

March 2005

[†]Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Director, Center for Advanced Studies in Measurement and Assessment (CASMA), 210D 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
Fax: 319-384-0505
Web: www.education.uiowa.edu/casma

All rights reserved

Contents

True, Observed, and Error Variances	1
Reliability-like Coefficients	4
Estimators of Reliability-like Coefficients	7
Other Perspectives on Profile Reliability	8
Error-Tolerance Ratios	9
Profile Stability	10
Decision Consistency	11
Concluding Comments	14
Appendix: Proof that $\sigma^2(\delta) = \sigma^2(\Delta) - \sigma^2(\bar{X})$	14
References	15

Abstract

This paper develops the basics of a test theory for characterizing an individual's profile in measurement terms such as true score variance, observed score variance, error variances, reliability-like coefficients, error-tolerance ratios, decision-consistency indices, etc. This is accomplished by making use of well-known statistical results (especially the analysis of variance identity), principles from both univariate and multivariate generalizability theory, and other well-known measurement results. It is shown that some lessons learned from the test theory for single scores or score composites do not always generalize to profiles of scores. This report is a work in progress that will be updated periodically as new results and insights are developed.

Given the wide use of profiles in many testing programs, it is perhaps surprising how little literature is devoted to measurement characteristics of profiles. Cronbach and Gleser (1953) published a seminal paper on the subject, but it is over 50 years old. The primary purpose of this paper is to develop the basics of a test theory for characterizing an individual's profile in measurement terms such as true score variance, observed score variance, error variances, reliability-like coefficients, error-tolerance ratios, decision-consistency indices, etc. This is accomplished by making use of well-known statistical results (especially the analysis of variance identity), principles from both univariate and multivariate generalizability theory, and other well-known measurement results. This report is a work in progress that will be updated periodically as new results are developed.

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 (2001) provides a recent extended treatment of the theory; Shavelson and Webb (1991) provide a brief monograph that describes the basics of the theory. See Brennan (1997) for a detailed discussion of the history of generalizability theory.

Brennan (2001) provides a treatment of the reliability of profiles from the perspective of multivariate generalizability theory. However, his treatment focuses on the profile for a typical (i.e., randomly selected) examinee, as opposed to the profile for a particular individual. This paper focuses on a specific individual. As such, aspects of this paper borrow from the literature on conditional error variance in multivariate generalizability theory (see, for example, Brennan, 2001, pp. 314–317).

The notation used in this paper is closely related to the notation in Brennan (2001), with one very important difference. In this paper, unless otherwise specified, all scores, means, and variances are for an *individual* person (say p), not a population of persons. The p subscript is suppressed to simplify notation. Note also that means and variances are over the variables (e.g., test scores) referenced in the profile, or over replications of the profile in the sense discussed later; means and variances are *not* over persons in some population, unless so specified. To simplify matters somewhat, test items are assumed to be dichotomously scored, although many results do not depend on this assumption.

True, Observed, and Error Variances

Let τ_v be the person's true (or universe) score for variable v ($v = 1, 2, \dots, k$). The variance of the k true scores is

$$\sigma^2(\tau) = \frac{1}{k} \sum_{v=1}^k (\tau_v - \mu)^2, \quad (1)$$

where τ_v is the true score for variable v , and $\mu = \sum \tau_v / k$.

Suppose the variables are locally independent (e.g., they do not share common items or stimuli). Then, the conditional (i.e., person-specific) profile error

variance is defined as is the average of the Δ -type error variances for the k variables:¹

$$\sigma^2(\Delta) \equiv \frac{1}{k} \sum_{v=1}^k \sigma^2(\Delta_v), \quad (2)$$

where the Δ -type error for variable v is

$$\Delta_v \equiv X_v - \tau_v, \quad (3)$$

with X_v being the observed mean score for variable v . That is, $\sigma^2(\Delta)$ quantifies (in a variance sense) the degree of similarity between an observed profile and a true profile.

If observed scores are binomially distributed (i.e., items are dichotomously scored), then Equation 2 becomes

$$\sigma^2(\Delta) = \frac{1}{k} \sum_{v=1}^k \frac{\tau_v(1 - \tau_v)}{n_v}, \quad (4)$$

where n_v is the number of items contributing to an observed score for variable v . Equation 4 is the mean-score metric version of Feldt's (1984) conditional error variance, which is the average over variables (or strata) of Lord's (1957) conditional error variance for each variable.

In practice, of course, we do not have a person's true score profile, but we can get the person's observed score profile for any *replication* of the measurement procedure. A replication for a person is defined as a set of k observed scores, with the observed score for variable v based on a random sample of n_v items. For a particular replication, r , of the measurement procedure, the observed score variance is defined as

$$S^2(X) = \frac{1}{k} \sum_{v=1}^k (X_{vr} - \bar{X}_r)^2, \quad (5)$$

where X_{vr} is the observed score (in the mean score metric) for variable v and replication r , and $\bar{X}_r = \sum X_{vr}/k$ is the observed mean over the k variables. In older literature, $S^2(X)$ or its square root is sometimes called the scatter in the profile and \bar{X}_r is sometimes called the elevation of the profile.

The expected value, \mathbf{E} , (over replications of the measurement procedure) of the observed score variance can be expressed as

$$\mathbf{E}S^2(X) = \mathbf{E} \left[\frac{1}{k} \sum_{v=1}^k (X_{vr} - \bar{X}_r)^2 \right]. \quad (6)$$

That is, "on average," the variance of the observed scores for a profile is given by Equation 6.

¹This is a special case of Equation 10.47 in Brennan, 2001, p. 314.

Table 1: Some Notational Conventions

Replication	Variables				Mean	Variance
	v_1	v_2	\dots	v_k		
1	X_{11}	X_{12}	\dots	X_{1k}	\bar{X}_1	$S^2(X_1)$
2	X_{21}	X_{22}	\dots	X_{2k}	\bar{X}_2	$S^2(X_2)$
\vdots	\vdots	\vdots	\dots	\vdots	\vdots	\vdots
r	X_{r1}	X_{r2}	\dots	X_{rk}	\bar{X}_r	$S^2(X_r)$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
Expected Value	τ_1	τ_2	\dots	τ_k	μ	$\mathbf{E}S^2(X)$
Variance	$\sigma^2(\Delta_1)$	$\sigma^2(\Delta_2)$	\dots	$\sigma^2(\Delta_k)$	$\sigma^2(\bar{X})$	

It is not easy to use Equation 6 directly to obtain an expression for $\mathbf{E}S^2(X)$ in terms of true-score parameters, as in Equations 1 and 4. This can be accomplished, however, using the analysis of variance (ANOVA) identity. Consider a matrix in which the k columns represent the variables in a profile and the rows represent replications of the profile (see Table 1). The ANOVA identity states that the total variance equals the average of the column variances plus the variance of the column means or, equivalently, the average of the row variances plus the variance of the row means. That is, the ANOVA identity implies that

$$\sigma^2(\Delta) + \sigma^2(\tau) = \mathbf{E}S^2(X) + \sigma^2(\bar{X}), \quad (7)$$

where

- $\sigma^2(\Delta)$ is the average of the variances for the variables (i.e., columns),
- $\sigma^2(\tau)$ is the variance of true scores $\tau_1, \tau_2, \dots, \tau_k$ (i.e., column expected values) for the k variables,
- $\mathbf{E}S^2(X)$ is the expected value (over replications) of the observed score variance (for rows), and
- $\sigma^2(\bar{X})$ is the variance (over replications) of the means for the k variables (for rows).

By definition,

$$\sigma^2(\bar{X}) \equiv \sigma_r^2 \left(\frac{1}{k} \sum_{v=1}^k X_{vr} \right), \quad (8)$$

where the r is used as a subscript to σ^2 simply to emphasize that the variance is over replications. When scores for the variables are locally independent and binomially distributed,

$$\sigma^2(\bar{X}) = \frac{1}{k^2} \sigma_r^2 \left(\sum_{v=1}^k X_{vr} \right) = \frac{1}{k^2} \sum_{v=1}^k \sigma_r^2 (X_{vr}) = \frac{1}{k^2} \sum_{v=1}^k \frac{\tau_v(1-\tau_v)}{n_v}.$$

Under these assumption, it follows from Equation 4 that

$$\sigma^2(\bar{X}) = \frac{\sigma^2(\Delta)}{k}. \quad (9)$$

Using the ANOVA identity in Equation 7 along with Equation 9, we can now express $\mathbf{E}S^2(X)$ in Equation 6 as:

$$\mathbf{E}S^2(X) = \sigma^2(\tau) + \sigma^2(\Delta) - \sigma^2(\bar{X}) \quad (10)$$

$$= \sigma^2(\tau) + \left(\frac{k-1}{k}\right) \sigma^2(\Delta). \quad (11)$$

Using Equations 1 and 4, for binomially distributed variables

$$\mathbf{E}S^2(X) = \frac{1}{k} \sum_{v=1}^k (\tau_v - \mu)^2 + \left(\frac{k-1}{k}\right) \sum_{v=1}^k \frac{\tau_v(1-\tau_v)}{n_v}. \quad (12)$$

In generalizability theory it is traditional to denote the decomposition of expected observed score variance as

$$\mathbf{E}S^2(X) = \sigma^2(\tau) + \sigma^2(\delta). \quad (13)$$

Given this convention, it is evident from Equation 11 that

$$\sigma^2(\delta) = \left(\frac{k-1}{k}\right) \sigma^2(\Delta), \quad (14)$$

where $\sigma^2(\delta)$ is the average of the δ -type error variances for the k variables. For any variable v , the δ -type error is

$$\delta_v \equiv (X_v - \bar{X}) - (\tau_v - \mu) = (X_v - \tau_v) - (\bar{X} - \mu), \quad (15)$$

where $\bar{X} = \sum X_v/k$ is the observed mean over the k variables.²

There are at least two important points to emphasize about Equation 14. First, it holds when variables are binomially distributed; it does not necessarily hold when variables have other distributions. Second, $\sigma^2(\delta)$ is functionally related to $\sigma^2(\Delta)$, which does not occur in univariate generalizability theory. This is the first hint that lessons learned from the test theory for single scores or score composites do not always generalize to profiles of scores.

Reliability-like Coefficients

In conventional test theory (classical test theory, generalizability theory, etc.) there are at least three definitions of reliability:

1. the correlation between “parallel” forms, which is the oldest definition;

²It is proven in the appendix that $\sigma^2(\delta) = \sigma^2(\Delta) - \sigma^2(\bar{X})$, which proves the equivalence of Equations 10 and 13.

2. the squared correlation between observed and true scores, which is usually regarded as the “canonical” definition; and
3. the ratio of true score variance to observed score variance, or one of its variants such as 1 minus the ratio of error variance to observed score variance.

Under the strict assumption of classically parallel forms, all three definitions are equivalent (see Feldt & Brennan, 1989). Under weaker assumptions, the three definitions tend to be approximately equivalent. This is also the case for profiles as discussed in this section.

It is particularly important to note, however, that here the definition of true score variance is the variance of true scores (over variables) for a particular person. In conventional treatments of reliability, true score variance is the variance *over persons* of true scores. Hence, the mathematics here will look familiar to readers who are knowledgeable about definitions of reliability, but the interpretations here are quite different from traditional interpretations.

The first definition can be formulated for profiles as

$$\rho(X, X') = \frac{\sigma(X, X')}{\sigma(X)\sigma(X')},$$

for any two randomly parallel profiles X and X' , where each such profile consists of observed mean scores based on random samples of the same set of n_v values. Over a universe of randomly parallel profiles, this definition can be viewed as

$$\mathbf{E} \rho(X, X') \doteq \frac{\mathbf{E} \sigma(X, X')}{\mathbf{E} S^2(X)}, \quad (16)$$

where the expectations for the correlation and covariance are taken over all pairs of randomly parallel profiles (i.e., replications).³ By definition, the expected value of the covariance is

$$\mathbf{E} \sigma(X, X') = \mathbf{E} \left[\frac{1}{k} \sum_{v=1}^k (X_{vr} - \bar{X}_r)(X_{vr'} - \bar{X}_{r'}) \right], \quad (17)$$

where r and r' designate different replications. Since the expected value of a sum is the sum of the expected values, and since profiles are based on different sets of independent samples of items,

$$\begin{aligned} \mathbf{E} \sigma(X, X') &= \frac{1}{k} \sum_{v=1}^k \mathbf{E} [(X_{vr} - \bar{X}_r)(X_{vr'} - \bar{X}_{r'})] \\ &= \frac{1}{k} \sum_{v=1}^k \mathbf{E}(X_{vr} - \bar{X}_r) \mathbf{E}(X_{vr'} - \bar{X}_{r'}) \\ &= \frac{1}{k} \sum_{v=1}^k (\tau_v - \mu)^2 \end{aligned} \quad (18)$$

$$= \sigma^2(\tau). \quad (19)$$

It follows that

$$\mathbf{E} \rho(X, X') \doteq \frac{\sigma^2(\tau)}{\mathbf{E} S^2(X)}, \quad (20)$$

which is equivalent to the third definition of reliability.

The second definition of reliability can be formulated for profiles as

$$\rho^2(X, \tau) = \frac{[\sigma(X, \tau)]^2}{\sigma^2(X) \sigma^2(\tau)}.$$

Over a universe of replicated profiles, this definition can be viewed as³

$$\mathbf{E} \rho^2(X, \tau) \doteq \frac{\mathbf{E} [\sigma(X, \tau)]^2}{\mathbf{E} S^2(X) \sigma^2(\tau)}. \quad (21)$$

A simpler expression for this equation is unknown to the author. However, if it is assumed that $\mathbf{E} [\sigma(X, \tau)]^2$ is approximately equal to $[\mathbf{E} \sigma(X, \tau)]^2$, then⁴

$$\mathbf{E} \rho^2(X, \tau) \doteq \frac{[\sigma^2(\tau)]^2}{\mathbf{E} S^2(X) \sigma^2(\tau)} = \frac{\sigma^2(\tau)}{\mathbf{E} S^2(X)}, \quad (22)$$

which is equivalent to the variance-ratio definition of reliability—the third definition discussed previously.⁵

The remainder of this paper will use the variance-ratio definition of reliability primarily. Specifically, we consider the profile generalizability coefficient

$$\mathcal{G}_p = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)} = \frac{\sigma^2(\tau)}{\mathbf{E} S^2(X)} = 1 - \frac{\sigma^2(\delta)}{\mathbf{E} S^2(X)}, \quad (23)$$

and the profile dependability coefficient

$$\Phi_p = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)} = 1 - \frac{\sigma^2(\Delta)}{\sigma^2(\tau) + \sigma^2(\Delta)}. \quad (24)$$

Roughly speaking, Φ_p quantifies profile consistency in the sense of the similarity of scatter in scores (universe vs. observed), whereas \mathcal{G}_p quantifies profile consistency in the sense of the similarity of scatter in the scores after the average elevation of scores (universe vs. observed) is removed. Note, again, that for both coefficients observed score variance and true score variance are for the individual person, not over persons.

Using Equation 14, an alternative expression for \mathcal{G}_p is

$$\mathcal{G}_p = 1 - \frac{(k-1)\sigma^2(\Delta)/k}{\mathbf{E} S^2(X)}. \quad (25)$$

³This is an approximation because the expected value of a ratio is greater than or equal to the ratio of the expected values. That is, the right side of the equation is a lower limit.

⁴Limited experience suggests to this author that this assumption is often reasonable. Technically, however, using a version of Cauchy's Inequality or Chebyshev's Inequality, it can be shown that $\mathbf{E} [\sigma(X, \tau)]^2 \geq [\mathbf{E} \sigma(X, \tau)]^2$. Since $\mathbf{E} \sigma(X, \tau) = \sigma^2(\tau)$, it follows that the result in Equation 22 is less than or equal to the result in Equation 21.

⁵Strictly speaking, $\mathbf{E} \rho^2(X, \tau) \geq \mathbf{E} \rho(X, X') \geq \sigma^2(\tau)/\mathbf{E} S^2(X)$.

Using Equations 7 and 9, an alternative expression for Φ_p is

$$\Phi_p = 1 - \frac{\sigma^2(\Delta)}{\mathbf{E}S^2(X) + \sigma^2(\Delta)/k}. \quad (26)$$

It can be shown that⁶

$$\Phi_p = \left[1 + \frac{k}{k-1} \left(\frac{1}{\mathcal{G}_p} - 1 \right) \right]^{-1}. \quad (27)$$

Similarly,

$$\mathcal{G}_p = \left[1 + \frac{k-1}{k} \left(\frac{1}{\Phi_p} - 1 \right) \right]^{-1}. \quad (28)$$

Estimators of Reliability-like Coefficients

Clearly, in most real-data situations, for any given person we seldom have more than one profile. That is, we typically have only a single replication of the profile for a person. Even so, \mathcal{G}_p and Φ_p can still be estimated. This is facilitated by noting from Equations 25 and 26 that \mathcal{G}_p and Φ_p depend only upon k and the set of n_v sample sizes (which are specified a priori), as well as $\sigma^2(\Delta)$ and $\mathbf{E}S^2(X)$, which are easily estimated.

An estimator of $\mathbf{E}S^2(X)$ is simply the observed score variance for the profile:

$$\mathit{Est}[\mathbf{E}S^2(X)] = S^2(X) = \frac{1}{k} \sum_{v=1}^k (X_v - \bar{X})^2. \quad (29)$$

For k independent binomially distributed variables, an unbiased estimator of $\sigma^2(\Delta)$ in Equation 4 is (see Feldt, 1984, and Lord, 1957):

$$\hat{\sigma}^2(\Delta) = \frac{1}{k} \sum_{v=1}^k \frac{X_v(1-X_v)}{n_v-1}. \quad (30)$$

Using these estimators in Equation 25, it follows that a computational formula for $\hat{\mathcal{G}}_p$ is

$$\hat{\mathcal{G}}_p = 1 - \frac{\frac{k-1}{k^2} \sum_{v=1}^k \frac{X_v(1-X_v)}{n_v-1}}{\frac{1}{k} \sum_{v=1}^k (X_v - \bar{X})^2}. \quad (31)$$

Similarly, using Equation 26, a computational formula for $\hat{\Phi}_p$ is

$$\hat{\Phi}_p = 1 - \frac{\frac{1}{k} \sum_{v=1}^k \frac{X_v(1-X_v)}{n_v-1}}{\frac{1}{k} \sum_{v=1}^k (X_v - \bar{X})^2 + \frac{1}{k^2} \sum_{v=1}^k \frac{X_v(1-X_v)}{n_v-1}}. \quad (32)$$

⁶First derived by Ping Yin, March, 2005.

As discussed next, however, these estimators should be used with caution.

From Equations 23 and 24, it is evident that $\mathcal{G}_p \geq \Phi_p$, as in univariate generalizability theory. Given this fact, and using the expressions for \mathcal{G}_p and Φ_p in Equations 25 and 26, respectively, it is easy to show that

$$\mathbf{E}S^2(X) \geq \sigma^2(\Delta) \quad (33)$$

for binomially distributed variables. It is important to note, however, that this result is expressed in term or parameters, not estimates.

In terms of estimates, the observed score variance for a given profile could be larger, smaller, or equal to $\hat{\sigma}^2(\Delta)$. (Consider, for example, the trivial case in which the observed profile is flat, resulting in an observed score variance of 0.) If the observed score variance for a profile is less than $\hat{\sigma}^2(\Delta)$, then we will have the paradoxical result that $\hat{\mathcal{G}}_p < \hat{\Phi}_p$ using Equations 31 and 32. Such a result is most likely to occur when an observed score profile is relatively flat with variable mean scores not too far from .5. Note also that it is evident from the form of Equations 31 and 32 that $\hat{\mathcal{G}}_p$ and $\hat{\Phi}_p$ can be negative when the examinee's observed score variance is small.⁷

These characteristics of these particular estimates of \mathcal{G}_p and Φ_p may mean that they have limited utility in some circumstances. In any case, it is always wise to examine the component parts such as \bar{X} , $S^2(X)$, $\hat{\sigma}^2(\Delta)$, and $\hat{\sigma}^2(\delta)$.

Other Perspectives on Profile Reliability

Are there relationships between the person-level (or conditional) coefficients \mathcal{G}_p and Φ_p and the traditional generalizability coefficient ($\mathbf{E}\rho^2$) and Phi coefficient (Φ) that are defined *over a population of persons*? Some, but not many. Suppose we formulate the context as one in which a composite is defined as a simple average of the k variables. Then, for example:

- The overall error variance in Φ (for the composite) is identical to the average over persons of $\sigma^2(\Delta)$ given in Equation 4.
- There is no simply relationship between the overall error variance in $\mathbf{E}\rho^2$ and the average over persons of $\sigma^2(\delta)$ given in Equation 14. The essential reason is that the definition of δ in $\mathbf{E}\rho^2$ involves deviating observed and universe scores from their means *over persons*, but the definition of δ in \mathcal{G}_p is an intraindividual definition.
- Both true score variance and observed score variance are defined within the person for \mathcal{G}_p and Φ_p , but they are defined over persons for $\mathbf{E}\rho^2$ and Φ .

⁷As $S^2(X) \rightarrow 0$, $\hat{\mathcal{G}}_p \rightarrow -\infty$ and $\hat{\Phi}_p \rightarrow (1-k)$. Negative values can also occur for traditional coefficients in which observed score variance is over a population of persons, but negative values are a much more likely occurrence for $\hat{\mathcal{G}}_p$ and $\hat{\Phi}_p$ than for traditional coefficients.

For expository purposes, to make the above comparisons it was assumed that a composite was of interest and it consisted of equally weighed components. These assumptions, however, are not necessary (and not necessarily reasonable) when interest focuses exclusively on profiles. That is, the existence of a profile does not imply that a composite of any particular type must be formed, and profile reliability is conceptually and mathematically different from traditional perspectives on reliability.

Error-Tolerance Ratios

Various indices of measurement precision are discussed by Kane (1996). In particular, he defines an error-tolerance ratio (E/T) as the root mean square of the errors of interest divided by the root mean square of the tolerances of interest. E/T will be small if errors are small relative to the tolerances, suggesting that measurements have substantial precision for the intended use. Suppose, for example, that the error root mean square is $\sigma(\Delta)$ and the tolerance root mean square is the standard deviation of the examinee's true scores, $\sigma(\tau)$. Then,

$$E/T = \frac{\sigma(\Delta)}{\sigma(\tau)}. \quad (34)$$

Under these circumstances, it is easy to show that

$$E/T = \sqrt{\frac{1 - \Phi_p}{\Phi_p}}, \quad (35)$$

and

$$\Phi_p = \frac{1}{1 + (E/T)^2}. \quad (36)$$

Similar relationships hold for E/T and \mathcal{G}_p when the root mean square of the errors of interest is $\sigma(\delta)$ and the tolerance is $\sigma(\tau)$.

The notion of an error-tolerance ratio is not restricted to situations involving square roots of variances. For example, suppose there is some target or criterion profile defined as $\lambda_v = \lambda$ for all v . Then, the error for variable v is

$$(X_v - \lambda) - (\tau_v - \lambda) = X_v - \tau_v = \Delta_v.$$

This means that $E = \sigma(\Delta)$, which is indeed a standard deviation. However, the tolerance for an individual examinee is

$$T = \sqrt{\frac{1}{k} \sum_{v=1}^k (\tau_v - \lambda)^2}, \quad (37)$$

which is not a standard deviation. In this case, it is easy to show that

$$E/T = \frac{\sigma(\Delta)}{\sqrt{\sigma^2(\tau) + (\mu - \lambda)^2}}. \quad (38)$$

Equations 35 and 35 suggest a much tighter linkage between Φ_p and an error-tolerance ratio than is necessarily the case. In particular, the tolerance need not be defined as $\sigma(\tau)$ for the *individual* examinee. For example, tolerance could be defined as the square root of the expected true-score variance over a *population* of persons; i.e.,

$$T = \sqrt{\mathbf{E}_p \sigma_p^2(\tau)}, \quad (39)$$

where $\sigma_p^2(\tau)$ is the true score variance for examinee p , and \mathbf{E}_p designates the expected value over the population of persons. In this case, error variability for a particular examinee is compared to the magnitude of true score variability for a “typical” examinee.

Alternatively, tolerance could be defined as the square root of the average (over variables) of the true score variance (over persons); i.e.,

$$T = \sqrt{\frac{1}{k} \sum_{v=1}^k \mathbf{E}_p [\tau_{pv} - \mathbf{E}_p(\tau_{pv})]^2}, \quad (40)$$

where τ_{pv} is the true score for examinee p on variable v . Under this definition of tolerance, suppose that $E/T = 2$, with E defined as $\sigma(\Delta)$. Roughly speaking, this means that there is twice as much error variability in the particular examinee’s observed scores as there is variability among all examinees’ true scores for the k variables.

As another example, tolerance could be defined as the standard deviation (over examinees in the population) of the examinees’ true composite scores; i.e.,

$$T = \sqrt{\mathbf{E}_p [\mu_p - \mathbf{E}_p(\mu_p)]^2}, \quad (41)$$

where μ_p is the true composite score for examinee p . Under this definition of tolerance, suppose that $E/T = 2$, with E defined as $\sigma(\Delta)$. Roughly speaking, this means that there is twice as much error variability in the particular examinee’s observed scores as there is variability among all examinees’ true composite scores.

These examples illustrate that a measure of tolerance scales error in a manner that the investigator judges useful for some purpose. There is no “right” answer to what the tolerance should be. It is even possible that the tolerance might vary for different examinees or different groups of examinees.

Profile Stability

In general, we define profile stability for an individual examinee is the extent to which observed profiles are stable over replications with respect to shape and/or elevation. We quantify profile stability in the sense of shape (*PSS*) as a root mean squared error for two randomly selected replications; specifically,

$$PSS \equiv \sqrt{\frac{1}{k} \sum_{v=1}^k \mathbf{E}(X_{vr} - X_{vr'})^2}, \quad (42)$$

where the expectation is taken over pairs of replications, r and r' . Since X_{vr} and $X_{vr'}$ share the same true score, it follows that

$$\begin{aligned} PSS &= \sqrt{\frac{1}{k} \sum_{v=1}^k \mathbf{E}(\Delta_{vr} - \Delta_{vr'})^2} \\ &= \sqrt{\frac{1}{k} \sum_{v=1}^k 2\sigma^2(\Delta_v)} \\ &= \sqrt{2} \sigma(\Delta). \end{aligned} \tag{43}$$

It is instructive to compare the definition of PSS in Equation 42 with the definition of $\sigma(\Delta)$ (see Equations 2 and 3):

$$\sigma(\Delta) \equiv \sqrt{\frac{1}{k} \sum_{v=1}^k \mathbf{E}(X_{vr} - \tau_v)^2}, \tag{44}$$

which quantifies the extent to which the profile of observed scores matches the profile of true scores. The additional error involved in comparing two observed profiles (rather than one observed profile and the true profile) gives rise to the $\sqrt{2}$ multiplier in Equation 43.

PSS quantifies the extent to which the shape of an observed profile is stable for two randomly selected replications. We can use the same logic to quantify the extent to which the elevation (E) of an observed profile is stable for two randomly selected replications; specifically,

$$PSE \equiv \sqrt{\mathbf{E}(\bar{X}_r - \bar{X}_{r'})^2}. \tag{45}$$

It follows that

$$PSE = \sqrt{2} \sigma(\bar{X}) = \sqrt{\frac{2}{k}} \sigma(\Delta), \tag{46}$$

where the last equality follows from Equation 9 for locally independent, binomially distributed variables.

Decision Consistency

All of the measures of individual profile reliability discussed so far share one characteristic in common—they all use a squared-error loss definition of error. Threshold-loss error in the context of decision consistency provides an alternative perspective. (See Kane & Brennan, 1980, for a detailed consideration of loss functions and traditional coefficients.) In traditional treatments of reliability, decision consistency is defined as the proportion of examinees classified in the same manner on two administrations of a test or two forms of the same test. When data are available for only one form administered once, the comparison is between

- scores for two hypothetical replications of the measurement procedure, or
- the actual scores and scores for one hypothetical replication.

Here we will focus on comparisons of the second type.

For an individual's profile, we conceptualize decision consistency as the proportion of replicated profiles that share a defined characteristic. For example, suppose an examinee's observed profile of $k = 4$ scores were $\{X_1 = .8, X_2 = .6, X_3 = .9, X_4 = .5\}$ based on $\{20, 10, 20, \text{ and } 8\}$ items. Variable 3 can be viewed as a strength for the examinee since the examinee's score on the third variable is the largest. Consequently, an investigator might be interested in the proportion of profiles that share this characteristic. Similarly, variable 4 can be viewed as a weakness, and an investigator might be interested in the proportion of profiles that share this characteristic. As a more complicated example, an investigator might be interested in the proportion of profiles for which scores for variables 1 and 3 are the largest.

These types decision consistency indices are easily obtained by comparing the actual profile to some number (R) of replicated profiles that can be generated using the binomial distribution or a bootstrap procedure. Let us consider the binomial procedure first.

Binomial procedure. Recall that we are assuming that each of the k variables have independent binomial distributions with parameters n_v and τ_v . Since we do not know the τ_v values, we use the X_v values.⁸ Then, the steps are as follows:

1. For each of the variables, obtain a binomial random variable and divide it by n_v to convert it to the mean score metric. The resulting k scores constitute a replicated profile.
2. Compare the actual profile to the replicated profile to determine if there is a match relative to the defined characteristic.
3. Repeat the first two steps R times.
4. The resulting proportion of matches is the decision consistency statistic for the defined characteristic.

Bootstrap procedure. The bootstrap procedure, or more correctly the *stratified* bootstrap procedure, involves the same four steps as the binomial procedure, except that Step 1 is altered as follows. For each of the k variables, v , take a random sample with replacement of size n_v from the actual n_v 0/1 item scores. The means for these k bootstrap samples constitute a replicated profile. As $R \rightarrow \infty$, the bootstrap procedure and binomial procedure (using X_v in place of τ_v) give results that are progressively closer to each other.

Overlapping confidence intervals. It is relatively common practice for test publishers to produce score reports in which a confidence interval is placed

⁸Of course, if we have better estimates of the τ_v , we can use them. An obvious possibility is to use regressed score estimates, if one is willing to "borrow" information from other examinees.

around each score in a person's profile. Almost always the width of each interval is two standard errors of measurement, and the person is usually told something like, "If any two intervals for your scores overlap, then your level of achievement for the two tests is likely the same." This common practice raises a number of questions about overlapping confidence intervals, some of which are discussed by Brennan (2001, pp. 317–320) under normality assumptions. Here, we outline how answers to such questions might be obtained without making such assumptions.

Consider the example introduced above, and recall that Lord's (1957) estimated standard error of measurement is

$$\hat{\sigma}(\Delta_v) = \sqrt{\frac{X_v(1 - X_v)}{n_v - 1}}.$$

It follows that the profile statistics and confidence interval limits are:

n_v	20	10	20	8
X_v	.8	.6	.9	.5
$\hat{\sigma}(\Delta_v)$.0918	.1633	.0688	.1890
upper limit	.89	.76	.97	.69
lower limit	.71	.44	.83	.31

Therefore, for the actual data, the confidence intervals overlap for variables 1 and 2, 1 and 3, and 2 and 4. Using either the binomial or the bootstrap procedure we can obtain a replicated profile, perform the same computations, and compare results for the actual and replicated profile. We can do this R times to obtain numerous types of results such as:

- the proportion of times that confidence intervals overlap for variables 1 and 2,
- the proportion of times that confidence intervals overlap for variables 1 and 3,
- the proportion of times that confidence intervals overlap for variables 2 and 4.
- the proportion of times that confidence intervals overlap for variables 1 and 2, 1 and 3, *and* 2 and 4.

A limitation. Assessing decision consistency using the binomial or bootstrap procedure in the manner discussed here has considerable practical utility but it has a theoretical limitation. Specifically, for both procedures we are essentially conditioning on the examinee's observed mean scores rather than the true scores. A worthy topic for research is the extent to which this limitation has meaningful consequences for the types of decision consistency indices discussed here.

Concluding Comments

In this paper, a profile replication for a person has been defined as a set of k observed scores, with the observed score for variable v based on a random sample of n_v items. Sometimes this random sampling assumption may not be tenable, but often it seems rather reasonable. For example, in many large-scale testing programs, diagnostic profiles for a particular form of a test are reported based on raw scores in the sense of proportions of items correct in k diagnostic areas. Presumably, for a different form of the test, different items would contribute to each of the diagnostic areas, but the profile would still consist of raw scores. If so, with forms viewed as replications, the random sampling assumption seems reasonable defensible, even though not strictly true.

For the theory discussed thus far in this paper, the sample size pattern (i.e., n_1, n_2, \dots, n_k) must be the same for each replication. However, for diagnostic profiles, it is often the case that the test specifications do not require fixed sample sizes for the diagnostic areas. If so, the results discussed in this paper might be reported for various likely sample size patterns—not just the pattern in the test form under consideration.

As noted in a previous section, if an observed profile is flat or close to flat, $S^2(X) \rightarrow 0$ and Equations 31 and 32 will tend to be negative. In one sense, this can be useful information in that it indicates that the true score profile is also very likely to be flat. Note that the likelihood of a negative-estimate could be mitigated by inducing some variability into $S^2(X)$. One ad hoc approach that might be worthy of study is to use a bootstrap procedure to get multiple values of $S^2(X)$ for the examinee and then average over these values.

Consider the case in which examinee composite scores are provided for a test or battery consisting of k parts. If the reliability of the composite scores in the traditional sense is high, it is very likely that individual profiles will be relatively stable over replications. However, high composite reliability in the traditional sense implies nothing about the shape of the profiles. Indeed, all or most of them could be (nearly) flat. As noted previously, lessons learned from the test theory for single scores or score composites do not always generalize to profiles of scores.

Appendix: Proof that $\sigma^2(\delta) = \sigma^2(\Delta) - \sigma^2(\bar{X})$

By definition, $\sigma^2(\delta)$ is the average of the δ -type error variances for the k variables; i.e.,

$$\sigma^2(\delta) \equiv \frac{1}{k} \sum_{v=1}^k \sigma^2(\delta_v), \quad (47)$$

where the δ -type error is given by Equation 15, namely,

$$\delta_v \equiv (X_v - \bar{X}) - (\tau_v - \mu) = (X_v - \tau_v) - (\bar{X} - \mu).$$

For any variable v the δ -type error variance is

$$\begin{aligned}\sigma^2(\delta_v) &= \mathbf{E}[(X_v - \tau_v) - (\bar{X} - \mu)]^2, \\ &= \sigma^2(\Delta_v) + \sigma^2(\bar{X}) - 2\mathbf{E}[(X_v - \tau_v)(\bar{X} - \mu)],\end{aligned}$$

where the expectation is taken over replications. Now, using the definition of $\sigma^2(\delta)$ in Equation 47 and the fact that $\sigma^2(\Delta) \equiv \sum_{v=1}^k \sigma^2(\Delta_v)/k$,

$$\begin{aligned}\sigma^2(\delta) &= \sigma^2(\Delta) + \sigma^2(\bar{X}) - \frac{2}{k} \sum_{v=1}^k \mathbf{E}[(X_v - \tau_v)(\bar{X} - \mu)] \\ &= \sigma^2(\Delta) + \sigma^2(\bar{X}) - \frac{2}{k} \mathbf{E} \left\{ \sum_{v=1}^k [(X_v - \tau_v)(\bar{X} - \mu)] \right\} \\ &= \sigma^2(\Delta) + \sigma^2(\bar{X}) - 2 \mathbf{E} \left[(\bar{X} - \mu) \sum_{v=1}^k \frac{X_v - \tau_v}{k} \right] \\ &= \sigma^2(\Delta) + \sigma^2(\bar{X}) - 2 \mathbf{E}(\bar{X} - \mu)^2 \\ &= \sigma^2(\Delta) - \sigma^2(\bar{X}).\end{aligned}$$

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. (2001). *Generalizability theory*. Springer-Verlag.
- Cronbach, L. J. & Gleser, G. C. (1953). Assessing similarity between profiles. *The Psychological Bulletin*, 50, 456–473.
- 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. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement*, 44, 883–891.
- 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. (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.

Lord, F. M. (1957). Do tests of the same length have the same standard errors of measurement? *Educational and Psychological Measurement*, *17*, 510–521.

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