Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 42

Utility Indexes for Composite Scores[1]

Robert L. Brennan[2]

May 1, 2015

Revised: September 10, 2016

[1] The research reported here was supported, in part, by a contract with the College Board. It should not be inferred, however, that the College Board supports the methodology reported here. The author gratefully acknowledges the helpful reviews of this report by Hyung Jin Kim, Associate Research Scientist, CASMA.

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

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

Contents

Abstract

Suppose there is a battery of tests, each of which generates a test score for examinees. Suppose, as well, that an additional score denoted X is created based on examinee performance on a subset of the items from each of the tests in the battery, or some of them. This paper extends the basic principles of classical test theory and the approach taken by Brennan (2011) to address the question of whether or not it is preferable, in a certain sense, to estimate true scores on X using X itself or using some composite, Z, that is a weighted combination of scores for the tests in the battery. The theory presented here is quite general in that it applies to both raw scores and scale scores, and it permits the user to choose how the component parts of Z and X are weighted. A special case of this theoretical framework gives the Haberman(2008)/Brennan(2011) results for subscores.

Suppose that S is a subscore of Y, in the sense that S is entirely contained within Y. Brennan (2011) considers a procedure for determining whether or not, in a certain sense, Y is a better/worse estimate of true-score for S (i.e., T_S) than S itself. His results mirror those of Haberman (2008), although the two researchers employ different approaches to arrive at their conclusions. In short, the Haberman/Brennan papers provide a basis for deciding whether or not the subscore S provides "more useful" information about T_S (in a certain sense) than that provided by Y.

In Brennan's (2011) procedure, the operational definition of "more useful" involves a comparison of reliability of S, as defined by

$$\rho_S^2 = \rho^2(T_S, S) = \left[\frac{\sigma(T_S, S)}{\sigma(T_S)\,\sigma(S)}\right]^2, \tag{1}$$

with

$$U = \rho^2(T_S, Y) = \left[\frac{\sigma(T_S, Y)}{\sigma(T_S)\,\sigma(Y)}\right]^2, \tag{2}$$

where U is called a utility index. The basic notion is that S is preferred if $\rho^2(T_S, S) > \rho^2(T_S, Y)$, and Y is preferred if $\rho^2(T_S, Y) > \rho^2(T_S, S)$.

This paper applies the same type of logic, but in a more complex context. Specifically, suppose there is a battery of tests, each of which generates a test score for examinees. Suppose, as well, that an additional score denoted X is created based on examinee performance on a subset of the items from each of the tests in the battery, or some of them. Since the X score is not associated with an additional test that is distinct from the tests in the battery, X might be called a pseudo test or cross-test.

This paper extends the basic principles of classical test theory (see Feldt & Brennan, 1989, or Haertel, 2006), and the approach taken by Brennan (2011), to address the question of whether or not it is preferable, in a certain sense, to use scores on X as an estimate of T_X or to use some composite Z that is a weighted combination of scores for the actual tests in the battery. As discussed later, a special case of the framework provided in this paper gives the Haberman/Brennan results for subscores.

It should be noted that the disattenuated correlation $\rho(T_X, T_Z)$ does *not* address the issues discussed in this paper. The magnitude of the disattenuated correlation tells us the extent to which the constructs measured by X and Z are linearly related through their true scores. It is entirely possible for X and/or Z to be quite unreliable even if $\rho(T_X, T_Z) = 1$. In particular, the disattenuated correlation tells us nothing about whether the observed score for X or Z is a better estimate of T_X (or T_Z, for that matter). By contrast, the goal in this paper is specifically to identify whether X or Z is a better estimate of T_X.

1 Components of Reliability and U

Suppose there exist k tests, each of which generates a score Z_i. Suppose, as well, that an additional test score is created that consists of examinee performance on a *subset* of items from at least one, and possibly all, of the k tests. Each of these subsets of items is designated X_i. We let Z_i denote not only test i but also scores for test i; similarly, X_i stands for both a subset of items on test i and scores for the subset. For now, it is easiest to think of Z_i and X_i as raw scores, although the theory presented here makes no such assumption. Later, we consider special issues that can arise with scale scores.

Let the composite of full-length tests be

$$Z = w_1\,Z_1 + w_2\,Z_2 + \cdots + w_k\,Z_k, \tag{3}$$

where the w_i are nominal weights ($w_i \geq 0$ for all i) specified by the user. Similarly, let the composite of subsets of items from the Z_i be

$$X = v_1\,X_1 + v_2\,X_2 + \cdots + v_k\,X_k, \tag{4}$$

where the true and error scores associated with X are

$$T = v_1\,T_1 + v_2\,T_2 + \cdots + v_k\,T_k, \tag{5}$$

and

$$E = v_1\,E_1 + v_2\,E_2 + \cdots + v_k\,E_k, \tag{6}$$

and the v_i are nominal weights ($v_i \geq 0$ for all i) specified by the user. Note that the theory presented here does not require that there be any relationship between the w_i and v_i weights. Also, if $v_i = 1$, this does *not* mean that $X_i = Z_i$; rather $v_i = 1$ means that the score for the X_i subset of items from Z_i is weighted 1 in obtaining the composite X.

Formulating the theory in terms of the weights w_i and v_i leads to complicated equations at times, but the weights add considerable flexibility to the theory. Clearly, for example, Z can be any weighted combination of any (or all) of the Z_i. In most practical contexts, it is likely that all the v_i would be set to 1 but, again, the theory presented here makes no such assumption.

T_i is the true score associated with X_i; i.e., T_i is the expected value of X_i. It is important to note that the true score for the full set of items associated with Z_i is *not* necessarily the same as the true score for the subset of items associated with X_i. Note also that, for the utility indexes considered in this paper, we do *not* need to make explicit reference to true scores or error scores for the Z_i. Therefore, there is no ambiguity in letting T_i be the true score associated with X_i, with E_i being the error associated with X_i. In particular, to simplify notation, in this paper E_i will be used as a shorthand version of E_{X_i}. Similarly, E means E_X.

The numerator of the utility index in Equation 2 is $[\sigma(T_X, Z)]^2$, where

$$\sigma(T_X, Z) = \sigma[(v_1\,T_1 + v_2\,T_2 + \cdots + v_k\,T_k), (w_1\,Z_1 + w_2\,Z_2 + \cdots + w_k\,Z_k)], \tag{7}$$

which involves k^2 terms that could be displayed in a $k \times k$ matrix in which the off-diagonal elements are

$$\begin{aligned}
\sigma(v_i\,T_i, w_j\,Z_j) &= v_i\,w_j\,\sigma[(X_i - E_i), Z_j] \\
&= v_i\,w_j\,[\sigma(X_i, Z_j) - \sigma(E_i, Z_j)] \\
&= v_i\,w_j\,\sigma(X_i, Z_j), \qquad (8)
\end{aligned}$$

and the diagonal elements are

$$\begin{aligned}
\sigma(v_i\,T_i, w_i\,Z_i) &= v_i\,w_i\,\sigma(T_i, Z_i) \\
&= v_i\,w_i\,\sigma[(X_i - E_i), Z_i] \\
&= v_i\,w_i\,[\sigma(X_i, Z_i) - \sigma(E_i, Z_i)] \qquad (9) \\
&= v_i\,w_i\,[\sigma(X_i, Z_i) - \sigma^2(E_i)]. \qquad (10)
\end{aligned}$$

It is important to remember that X_i is the score for a subset of the items in Z_i, and E_i represents the errors in X_i, not Z_i. Equation 10 follows from Equation 9 because: (a) there are two sets of errors in Z_i—errors associated with the items in X_i and errors associated with the remaining items in Z_i; (b) the covariance of E_i and the errors associated with the remaining items in Z_i is 0; and (c) the covariance of E_i and the errors in Z_i associated with X_i is simply the variance of E_i. In essence, then, X_i and Z_i share some of the same error; stated differently, X_i and Z_i have correlated error, because the items in X_i are a subset of the items in Z_i. This fact is central to the theoretical framework of this paper.

Using Equations 8 and 10 in Equation 7, we obtain

$$\sigma(T_X, Z) = \sum_{i=1}^{k} v_i\,w_i\,[\sigma(X_i, Z_i) - \sigma^2(E_i)] + \sum_{i \neq j}\sum v_i\,w_j\,\sigma(X_i, Z_j), \qquad (11)$$

and the square of this equation gives the numerator of U in Equation 2. The denominator is $\sigma^2(T_X)\,\sigma^2(Z)$, where

$$\begin{aligned}
\sigma^2(T_X) &= \sigma^2(v_1\,T_1 + v_2\,T_2 + \cdots + v_k\,T_k) \\
&= \sum_{i=1}^{k} v_i^2 \sigma^2(T_i) + \sum_{i \neq j}\sum v_i\,v_j\,\sigma(T_i, T_j) \\
&= \sum_{i=1}^{k} v_i^2 [\sigma^2(X_i) - \sigma^2(E_i)] + \sum_{i \neq j}\sum v_i\,v_j\,\sigma(X_i, X_j), \qquad (12)
\end{aligned}$$

and

$$\begin{aligned}
\sigma^2(Z) &= \sigma^2(w_1\,Z_1 + w_2\,Z_2 + \cdots + w_k\,Z_k) \\
&= \sum_{i=1}^{k} w_i^2 \sigma^2(Z_i) + \sum_{i \neq j}\sum w_i\,w_j\,\sigma(Z_i, Z_j). \qquad (13)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\sigma^2(X) &= \sigma^2(v_1\,X_1 + v_2\,X_2 + \cdots + v_k\,X_k) \\
&= \sum_{i=1}^{k} v_i^2 \sigma^2(X_i) + \sum_{i \neq j}\sum v_i\,v_j\,\sigma(X_i, X_j). \qquad (14)
\end{aligned}$$

Equations 11–14 are expressed in terms of variance and covariances for the X_i and Z_i, the v_i and w_i weights, and the error variances associated with the X_i. As such, these are very general expressions. If X and Z are determined directly for each examinee, then it is straightforward to obtain $\sigma^2(X)$ and $\sigma^2(Z)$, and simplified versions of Equations 11 and 12 are

$$\sigma(T_X, Z) = \sigma(X, Z) - \sum_{i=1}^{k} v_i \, w_i \, \sigma^2(E_i), \tag{15}$$

and

$$\sigma^2(T_X) = \sigma^2(X) - \sum_{i=1}^{k} v_i^2 \, \sigma^2(E_i). \tag{16}$$

Note that each of the $\sigma^2(E_i)$ is associated with items from a different Z_i. It follows that the $\sigma^2(E_i)$ are associated with different constructs (or fixed strata) which may be measured by different types of stimuli and/or items. Therefore, it may be reasonable to use different procedures to estimate the different $\sigma^2(E_i)$. This matter is discussed more fully later.

2 Relative Utility Index U_r for Composite Scores

The magnitude of U alone does not tell us much about the merits of using Z rather than X. It seems clear that we need to compare $U = \rho^2(T_X, Z)$ to some other statistic that involves X. An obvious comparative statistic is the reliability of X, $\rho_X^2 = \rho^2(T_X, X)$. A convenient form for this comparison of U and ρ_X^2 is[3]

$$U_r = \frac{U}{\rho_X^2} = \frac{\rho^2(T_X, Z)}{\rho^2(T_X, X)}. \tag{17}$$

If $U_r > 1$, then Z is preferable to X as an estimate of T_X; if $U_r < 1$, then X is preferable to Z as an estimate of T_X, given the theory presented here.

U_r in Equation 17 can also be expressed as

$$\begin{aligned} U_r &= \left[\frac{\sigma^2(T_X, Z)}{\sigma^2(T_X)\,\sigma^2(Z)}\right] \left[\frac{\sigma^2(T_X)\,\sigma^2(X)}{\sigma^2(T_X, X)}\right] \\ &= \left[\frac{\sigma^2(X)}{\sigma^2(Z)}\right] \left[\frac{\sigma^2(T_X, Z)}{\sigma^2(T_X, X)}\right] \\ &= \left[\frac{\sigma^2(X)}{\sigma^2(Z)}\right] \left[\frac{\sigma(T_X, Z)}{\sigma^2(T_X)}\right]^2, \end{aligned} \tag{18}$$

since

$$\sigma(T_X, X) = \sigma(T_X, T_X + E_X) = \sigma^2(T_X) + \sigma(T_X, E_X) = \sigma^2(T_X),$$

because true and error scores are uncorrelated.

[3] Note that the form of the relative utility index U_r in Equation 17 is different from that in Brennan (2010), in which X is a subscore—i.e., one of the Z_i.

3 Estimating U_r for Raw Scores and $\widetilde{U}_r$ for Scale Scores

The next two subsections focus primarily on estimation of the $\sigma^2(E_i)$ (the error variances associated with the subsets of items in X) in the equations for $\sigma(T_X, Z)$ and $\sigma^2(T_X)$. Once these estimates are available, Equations 11–14 can be used to obtain U_r in Equation 18. The first subsection considers raw scores. The second considers scale scores—particularly, scale scores that are non-linear transformations of raw scores.

In this and subsequent sections, notational distinctions are not made between parameters and estimates, because doing so renders too many equations much more complicated than necessary. The context of the discussions makes the intended meaning clear.

3.1 Raw Scores

If the Z_i and X_i are raw scores in the traditional sense of number (or proportion) of items correct for dichotomously-scored items, or number (or proportion) of points for polytomously-scored items, then the estimation of $\sigma(T_X, Z)$ and $\sigma^2(T_X)$ is not too complicated. Doing so, however, requires careful attention to estimating the k values of $\sigma^2(E_i)$.

Recall that $\sigma^2(E_i)$ is the error variance associated with X_i, not Z_i (although the items come from Z_i). Therefore, it is *not* theoretically sensible to use an estimate of error variance for Z_i (nor, in most cases, a linear transformation of it) as an estimate of $\sigma^2(E_i)$. This is a non-trivial matter. The items chosen from Z_i for inclusion in X_i are presumably systematically different in some content-relevant sense from the rest of the items in Z_i. If that were not true, then there would be no particular reason to consider using X_i. That is, forming a composite, X, by selecting subsets of items from the Z_i implicitly assumes that there are at least two fixed strata for each Z_i. This means that:

- X_i and the remaining items in Z_i are different in some content or construct-relevant sense, and
- X_i and X_j are different in some content or construct-relevant sense.

The phrase “content or construct-relevant” is a bit fuzzy. To be more theoretically correct, we could say that the true scores for X_i and Z_i are not assumed to be linearly related, and the true scores for X_i and X_j are not assumed to be linearly related.

If we assume that the items in X_i satisfy the assumption of essential tau-equivalence, then

$$\sigma^2(E_i) = \sigma^2(X_i)(1 - \rho^2_{\alpha i}), \qquad (19)$$

can be used, where $\rho^2_{\alpha i}$ is coefficient α associated with X_i. This estimate of $\sigma^2(E_i)$ can be used in Equations 11 and 12 to obtain estimates of $\sigma(T_X, Z)$ and $\sigma^2(T_X)$, respectively.

Sometimes it is unreasonable to assume essential tau-equivalence for one (or more) of the component parts of X. In such cases, a reliability coefficient based on congeneric assumptions might be considered instead of coefficient α (see Feldt & Brennan, 1989, or Haertel, 2006).

Alternatively, generalizability theory might be considered (see Brennan, 1998, 2001). Suppose, for example, that: (a) Z_i consists of five passages, and each item in Z_i is associated with a single passage; (b) the first two passages test content area c_1, the last three passages test content area c_2, and there are (possibly) different numbers of items per passage; and (c) X_i consists of the c_1 passages. If so, letting p stand for persons, m stand for items, and h stand for passages, the multivariate G study design for Z_i is $p^\bullet \times (m^\circ \!:\! h^\circ)$, the D study design is $p^\bullet \times (M^\circ \!:\! H^\circ)$, and the computer program mGENOVA (Brennan, 2001b) can be used to estimate $\sigma^2(E_i)$, which is the error variance associated with the c_1 passages (see, also, Brennan, 2001a, pp 231–233, 283–284).[4]

3.2 Scale Scores

Assume we have raw-to-scale-score transformations for each of the Z_i and for X, but we do *not* have scale-score transformations for the X_i. In a large-scale testing program, having scale score transformations for each of the X_i would be extraordinarily unusual, since they would not likely be reported scores.

Let $\widetilde{Z}_i$ and $\widetilde{X}$ designate the scale-score versions of Z_i and X. The scale score variance $\sigma^2(\widetilde{X})$ is easily obtained by direct use of scale scores for each examinee. Similarly, the scale score variances $\sigma^2(\widetilde{Z}_i)$ are easily obtained and can be used in Equation 13 to obtain $\sigma^2(\widetilde{Z})$. The scale-score relative utility index, $\widetilde{U}_r$, has the same form as U_r, namely,

$$\widetilde{U}_r = \frac{\widetilde{U}}{\rho^2_{\widetilde{X}}} = \frac{\rho^2(T_{\widetilde{X}}, \widetilde{Z})}{\rho^2(T_{\widetilde{X}}, \widetilde{X})} = \left[\frac{\sigma^2(\widetilde{X})}{\sigma^2(\widetilde{Z})}\right] \left[\frac{\sigma(T_{\widetilde{X}}, \widetilde{Z})}{\sigma^2(T_{\widetilde{X}})}\right]^2 . \tag{20}$$

3.2.1 Variance of $\boldsymbol{T_{\widetilde{X}}}$

The scale-score analogue of $\sigma^2(T_X)$ is

$$\sigma^2(T_{\widetilde{X}}) \quad = \quad \sigma^2(\widetilde{X}) - \sigma^2(\widetilde{E}), \tag{21}$$

where $\sigma^2(\widetilde{E})$ is the error variance associated with the scale-score transformation of X. This error variance can be estimated using many procedures, some of which are discussed in Subsection 3.3.

[4] mGENOVA and Brennan (2001a) usually treat raw scores in the proportion-correct, not number-correct metric, which influences the equations for variance components, covariance components, and error variances. This metric difference can be circumvented by using $\sigma^2(E_i) = \sigma^2(X_i)(1 - \boldsymbol{E}\rho^2)$, where $\sigma^2(X_i)$ is expressed in the intended metric, and $\boldsymbol{E}\rho^2$ is the generalizability coefficient for X_i — i.e., the set of two passages associated with c_1.

3.2.2 Covariance of $T_{\widetilde{X}}$ and $\widetilde{Z}$

Because the $\widetilde{X}_i$ are not likely to be available, it is awkward (although possible) to make direct use of the scale score analogue of $\sigma(T_X, Z)$ in Equation 11 to obtain $\sigma(T_{\widetilde{X}}, \widetilde{Z})$. Often, a more useful approach is to use the following derivation:

$$
\begin{aligned}
\sigma(T_{\widetilde{X}}, \widetilde{Z}) &= \sigma\left[\left(\widetilde{X} - \sum_{i=1}^{k} v_i\, \widetilde{E}_i\right),\ \sum_{i=1}^{k} w_i\, \widetilde{Z}_i\right] \\
&= \sum_{i=1}^{k} w_i\, \sigma(\widetilde{X}, \widetilde{Z}_i) - \sigma\left[\left(\sum_{i=1}^{k} v_i\, \widetilde{E}_i\right), \left(\sum_{i=1}^{k} w_i\, \widetilde{Z}_i\right)\right] \\
&= \sum_{i=1}^{k} w_i\, \sigma(\widetilde{X}, \widetilde{Z}_i) - \sum_{i=1}^{k} v_i\, w_i\, \sigma(\widetilde{E}_i, \widetilde{Z}_i) \qquad (22) \\
&= \sum_{i=1}^{k} w_i\, \sigma(\widetilde{X}, \widetilde{Z}_i) - \sum_{i=1}^{k} v_i\, w_i\, \sigma^2(\widetilde{E}_i). \qquad (23)
\end{aligned}
$$

Equation 23 follows from Equation 22 by the same logic discussed in conjunction with the derivation of Equation 10.

If $w_i = v_i = 1$ for $i = 1, 2, \ldots, k$, then the summation term in Equation 23 is $\sigma^2(\widetilde{E})$, which can be estimated using procedures such as those discussed in Subsection 3.3. Otherwise, however, there is no obvious way to estimate the individual $\sigma^2(\widetilde{E}_i)$ terms, since it is not likely that there will be a raw-to-scale-score conversion for the X_i. There is, however, a relatively simple ad hoc procedure discussed next (see also Section 5.3).

Suppose we assume that error variances for raw and scale scores associated with X_i and $\widetilde{X}_i$, respectively, are proportional in the sense that

$$\frac{\sigma^2(\widetilde{E}_i)}{\sigma^2(\widetilde{E})} = \frac{\sigma^2(E_i)}{\sigma^2(E)}. \qquad (24)$$

Then,

$$\sigma^2(\widetilde{E}_i) = \frac{\sigma^2(\widetilde{E})}{\sigma^2(E)}\, \sigma^2(E_i) \qquad (25)$$

It follows from Equation 23 that

$$\sigma(T_{\widetilde{X}}, \widetilde{Z}) = \sum_{i=1}^{k} w_i\, \sigma(\widetilde{X}, \widetilde{Z}_i) - \frac{\sigma^2(\widetilde{E})}{\sigma^2(E)} \sum_{i=1}^{k} w_i\, v_i\, \sigma^2(E_i), \qquad (26)$$

where each of the terms is estimable in a fairly direct way, except perhaps for $\sigma^2(\widetilde{E})$, which is discussed next.

3.3 Estimating $\sigma^2(\widetilde{E})$

Given the theory outlined above, the only error variance for scale scores that is required is $\sigma^2(\widetilde{E})$ in Equations 21 and 26. This is the error variance for the scale-score transformation of X.

3.3.1 Linear Transformations

Suppose $\widetilde{X} = a + b\,X$, which means that $\widetilde{X}$ is a linear transformation of X. In this case, $\sigma^2(\widetilde{E}) = b^2\,\sigma^2(E)$. It follows from Equations 21 and 26 that

$$\sigma^2(T_{\widetilde{X}}) = \sigma^2(\widetilde{X}) - b^2\,\sigma^2(E)$$

and

$$\sigma(T_{\widetilde{X}}, \widetilde{Z}) = \sigma(\widetilde{X}, \widetilde{Z}) - b^2 \sum_{i=1}^{k} w_i\, v_i\, \sigma^2(E_i).$$

3.3.2 Non-linear Transformations

There are a number of procedures for obtaining $\sigma^2(\widetilde{E})$ for scale scores that are non-linear transformations of raw scores. Most of these procedures are based on obtaining conditional scale-score error variance for persons, and then integrating (or summing) over persons to obtain the overall scale-score error variance. Kolen, Hanson, and Brennan (1992) as well as Lee, Brennan, and Kolen (2000) provide descriptions or summaries of most of these procedures.

Kolen and Brennan (2014, pp. 405–407) discuss a procedure for constructing scale scores that involves a linear transformation of arcsine transformed raw scores. Since the arcsine transformation is non-linear, the resulting raw-to-scale score transformation is also non-linear. In addition, since the arcsine transformed raw scores have approximately equal conditional standard errors of measurement, so do the linear-transformed scale scores. The process of obtaining scale scores in this manner involves prespecifying the overall scale-score standard error of measurement, which can be used as $\sigma(\widetilde{E})$ for these scale scores.

4 Examples

The new SAT© Suite of Assessments consists of substantially revised versions of the SAT©, PSAT© and PSAT8/9© (which, in sense, replaces Readisteps©). These three new assessments were vertically scaled based on results from a large scaling study conducted in late 2014 and early 2015. All assessments include three reported Test scores in Reading (R), Language (L)[5], and Math (M), as well as the two Cross-Test scores in History (H) and Science (S). Importantly, H and S are *not* scores for distinct separately-timed tests. Rather, the items that contribute to H and S are subsets of items in Reading, Language, and Math. The H and S scores for the SAT© and the PSAT© are the focus of the discussion in this section. Note that in this section, italicized letters usually refer to scores on tests, but sometimes italicized letters refer to the tests themselves.

[5] This is often called the Writing and Language Test, or the Evidence-based Writing and Language Test. Here it is called the Language (L) Test merely to simplify terminology and notation in formulas.

4.1 SAT© Background

For the SAT©, the numbers of items that contribute to the three Test scores are: 52 (R), 44 (L), and 58 (M); the number of items that contribute to the two Cross-Test scores are 35 (H) and 35 (S). All items are multiple-choice, they are dichotomously scored, and the raw scores are number-correct scores (no correction for guessing). The following is a mapping of the notation for this example onto the notation used elsewhere in this paper: $Z_1 = R$, $Z_2 = L$, $Z_3 = M$, and $X = H$ or $X = S$.

For the SAT© form considered here, H consists of the sum of scores for 21 items from Reading, six items from Language, and eight items from Math. A similar statement holds for S, but the items are *different* from those for H.[6] Clearly H and S can be viewed as composites, as defined in this paper.

Number-correct raw scores on H are the simple sum of the subsets of raw scores from R, L, and M. It follows that $v_1 = v_2 = v_3 = 1$ and $H = X = X_1 + X_2 + X_3$. Similar statements apply to S.

In considering this example, it is useful to know that the Reading and Language Tests are passage-based, with each item associated with one and only one passage. The items in two history passages in the Reading Test contribute to H, and the items in two science passages in the Reading Test contribute to S.

For the Tests and Cross-Tests, raw scores are not reported to examinees. Rather, the reported scores are scale scores that range from 10 to 40. The raw to scale-score conversions were obtained from the vertical scaling study mentioned above. In that study, an arcsine transformation of raw scores was used that ultimately resulted in scale scores with approximately equal conditional standard errors of measurement, CSEMs (see Kolen & Brennan, 2014, pp. 405–410).

A seemingly obvious question to ask about H is, "Are the observed scale scores for H better estimates of true scores for H than the observed scale scores for R, L, or M?" A corresponding question applies to S. The values of the relative utility indexes that answer these questions are reported in Table 1 based on the statistics in Tables 2 and 3 which were obtained from over 5000 scaling-study examinees in grades 11 and 12.

4.2 Computation of $\widetilde{U}_r$ for H vs. R in the SAT©

Consider, for example, $\widetilde{U}_r$ for H when the Z composite is simply R [i.e., $Z = 1(R) + 0(L) + 0(M) = R$]. In Table 1, the reported value is $\widetilde{U}_r = .917$. Recall from Equation 20 that the four statistics required to estimate $\widetilde{U}_r$ are $\sigma^2(\widetilde{X})$, $\sigma^2(\widetilde{Z})$, $\sigma(T_{\widetilde{X}}, \widetilde{Z})$, and $\sigma^2(T_{\widetilde{X}})$. From Table 2, we obtain

$$\sigma^2(\widetilde{X}) = \sigma^2(\widetilde{H}) = 26.786 \qquad \text{and} \tag{27}$$

$$\sigma^2(\widetilde{Z}) = \sigma^2(\widetilde{R}) = 24.671. \tag{28}$$

[6] For more information on SAT© scores see https://collegereadiness.collegeboard.org/about/scores/structure and https://collegereadiness.collegeboard.org/pdf/scoring-sat-practice-test-1.pdf.

Table 1: $\widetilde{U}$ and $\widetilde{U}_r$ (compared to R, L, and M) for H and S in the SAT©

	w_i			History (H)		Science (S)	
	R	L	M	$\widetilde{U}$	$\widetilde{U}_r$	$\widetilde{U}$	$\widetilde{U}_r$
$Z = R$	1	0	0	.751	.917	.795	.917
$Z = L$	0	1	0	.779	.951	.793	.915
$Z = M$	0	0	1	.688	.840	.685	.790

Table 2: Scale-Score Variance-Covariance Matrix, $\sigma(\widetilde{E})$, and Reliability (ρ^2) for SAT© Tests and Cross-Tests

	R	L	M	H	S
R	24.671	22.223	16.931	23.228	23.747
L	22.223	30.064	19.448	23.508	24.212
M	16.931	19.448	21.914	19.138	19.561
H	23.228	23.508	19.138	26.786	22.173
S	23.747	24.212	19.561	22.173	27.103
$\sigma(\widetilde{E})$	1.7	1.8	1.5	2.2	1.9
ρ^2	.883	.892	.897	.819	.867

Note. For the five tests, $\sigma(\widetilde{E})$ is the "constant" conditional standard error of measurement that resulted from the scaling study, and ρ^2 is reliability using the formula $1 - \sigma^2(\widetilde{E})/\sigma^2(\bullet)$, where $\sigma^2(\bullet)$ stands for the scale-score variance for any one of the three tests or two cross-tests (diagonal elements in table).

Both of these statistics can be computed directly from examinees' reported scale scores.

Using Equation 21 and results reported in Table 2,

$$\sigma^2(T_{\widetilde{X}}) = \sigma^2(\widetilde{H}) - \sigma^2(\widetilde{E}) = 26.786 - (2.2)^2 = 21.946. \tag{29}$$

For this example, $w_i = 1$ for a single test, R, and $w_i = 0$ for the other two tests. This simplifies Equation 23 for $\sigma(T_{\widetilde{X}}, \widetilde{Z})$. Specifically,

$$\sigma(T_{\widetilde{X}}, \widetilde{Z}) = \sigma(\widetilde{X}, \widetilde{Z}) - \sigma^2(\widetilde{E}_1), \tag{30}$$

where $\widetilde{Z}$ is the scale scores for R, $\widetilde{X}$ is the scale scores for H, $\sigma(\widetilde{X}, \widetilde{Z}) = 23.228$ from Table 2, and $\widetilde{E}_1$ means the scale score error variance for the Reading items that contribute to H. From Equation 30, it is clear that $\sigma(T_{\widetilde{X}}, \widetilde{Z})$ gets reduced by the correlated error $\sigma^2(\widetilde{E}_1)$ that arises because Z and X (R and H, respectively in this example) share two reading passages and their associated items.

The computation of $\sigma(T_{\widetilde{X}}, \widetilde{Z})$ is challenging, primarily because we do not have the raw-to-scale score conversion tables for the sets of Reading, Language,

Table 3: Raw-score Statistics and $\sigma^2(\widetilde{E_i})$ for R, L, and M Items in SAT© Cross-Tests

	Cross-Test H			Cross-Test S		
	R	L	M	R	L	M
$\sigma^2(X_i)$	14.386	2.047	4.263	22.754	2.950	2.817
$\rho^2_{\alpha i}$	.720	.465	.709	.833	.693	.540
$\sigma^2(E_i)$	4.033	1.096	1.242	3.797	.906	1.294
$\sigma^2(\widetilde{E_i})$	3.064	.832	.944	2.285	.546	.779

Note. $\sigma^2(E_i) = \sigma^2(X_i)(1-\rho^2_{\alpha i})$ with computations performed with six decimal digits. $\sigma^2(\widetilde{E_i})$ was computed using Equation 25 with $\sigma^2(\widetilde{E}) = (2.2)^2 = 4.840$ for H and $\sigma^2(\widetilde{E}) = (1.9)^2 = 3.610$ for S.

and Math items that are in H. To circumvent this problem, the approach outlined in Section 3.2.2 was used. As discussed above $\sigma^2(\widetilde{E}) = (2.2)^2 = 4.840$, and from Table 3, $\sigma^2(E_1) = 4.033$. Since the three parts of H do not share any items or passages, we can assume that error variances for raw scores are uncorrelated. Using results in Table 3, it follows that for number-correct raw scores

$$\sigma^2(E) = \sigma^2(E_1) + \sigma^2(E_2) + \sigma^2(E_3) = 4.033 + 1.096 + 1.242 = 6.371, \qquad (31)$$

where the subscripts 1, 2, and 3 refer to R, L, and M, respectively.[7] Using Equation 25,

$$\sigma^2(\widetilde{E}_1) = \left(\frac{4.840}{6.371}\right) 4.033 = 3.064, \qquad (32)$$

and using Equation 30

$$\sigma(T_{\widetilde{X}}, \widetilde{Z}) = 23.228 - 3.064 = 20.164. \qquad (33)$$

Finally, replacing the results in Equations 27, 28, 29, and 33 in Equation 20, we obtain :

$$\widetilde{U}_r = \left[\frac{\sigma^2(\widetilde{X})}{\sigma^2(\widetilde{Z})}\right] \left[\frac{\sigma(T_{\widetilde{X}}, \widetilde{Z})}{\sigma^2(T_{\widetilde{X}})}\right]^2 = \left[\frac{26.786}{24.671}\right] \left[\frac{20.164}{21.946}\right]^2 = .917. \qquad (34)$$

Since $\widetilde{U}_r < 1$, use of H is favored over use of R—or, more specifically, it is preferable to use observed scale scores for H as estimates of true scale scores for H than to use observed scale scores for R as estimates of true scale scores for H.

Kim and Brennan (2015) provide R code for performing these computations. The required input is a subset of the statistics in Tables 2 and 3.

[7] Raw-score error variances for the X_i were obtained using coefficient α. Other procedures based on G theory were examined, as well, but the differences in results were trivial.

Table 4: $\widetilde{U}$ and $\widetilde{U}_r$ for H and S in the PSAT©

	w_i			History (H)		Science (S)	
	R	L	M	$\widetilde{U}$	$\widetilde{U}_r$	$\widetilde{U}$	$\widetilde{U}_r$
$Z = R$	1	0	0	.741	.891	.730	.911
$Z = L$	0	1	0	.758	.911	.755	.944
$Z = M$	0	0	1	.578	.695	.572	.715

4.3 PSAT©

The structure of the PSAT© is essentially the same as that for the SAT© discussed at the beginning of Section 4.1. The primary differences are: (a) the PSAT© is somewhat easier than the SAT©; (b) the PSAT© Tests are shorter — the three Test lengths are 47, 44, and 48 for R, L, and M, respectively; the two Cross-Test lengths are 32 for both H and S; and (c) the scale scores for each PSAT© Test and Cross-Test range from 8 to 38 rather than 10 to 40. These differences between the SAT© and PSAT© are consistent with the fact that the PSAT© is vertically scaled to the SAT©.Note that both H and S consist of 19, 6, and 7 items from R, L, and M, respectively, with H and S each including number-right scores for two passages from R.

Table 4 provides the relative utility indices for H and S compared to R, L, and M. As is the case for the SAT©, $\widetilde{U}_r < 1$ suggesting that both H and S are better estimates of their respective true scores than are R, L, and M, even though R, L, and M are longer than H and S. It comparing Tables 4 and 1, it is evident that H and S are somewhat more highly favored for the PSAT© than for the SAT©.

5 Concluding Comments

In this section, the discussion is usually couched in terms of the raw-score variables Z, Z_i, X, and X_i—as well as T_X. Distinctions are drawn between raw-score and scale-score variables only when doing so seems necessary to make a particular point (especially in Section 5.3).

The theory presented in this paper is very general, although computations are relatively simple. They require only $\sigma^2(X)$, $\sigma^2(Z)$, $\sigma(T_X, Z)$, and $\sigma^2(T_X)$—or, at a finer level of detail, the w_i weights, the v_i weights, the variance-covariance matrix for the Z_i, the variance-covariance matrix for the X_i, and the $\sigma^2(E_i)$. For scale scores, the $\sigma^2(\widetilde{E}_i)$ are required, also.

There are virtually no constraints on how Z_i, X_i, w_i, and v_i are defined, with two primary exceptions: (a) the items in X_i must come from Z_i; and (b) the w_i, and v_i cannot be negative. Neither error variances nor reliabilities for the Z_i (or the Z composite) are required. Also, numbers of items contributing to variables are not required. The only required indicators of uncertainty are the

$\sigma^2(E_i)$—and the $\sigma^2(\widetilde{E}_i)$, if scale scores are considered. The theory, then, is very flexible, but flexibility comes with a price–namely, some degree of complexity because of: (a) correlated error that arises from the overlapping items in Z_i and X_i; and (b) the need to be careful about definitions and transformations of variables.

5.1 Definitions of Z_i, X_i, w_i, and v_i

Although there are few theoretical constraints on the Z_i, X_i, w_i, and v_i, there are conceptual differences. For example, the Z_i are associated with actual tests in a battery, whereas the X_i are sets of items from the Z_i that combine (via the v_i weights) to give X scores. Also, there may be various different sets of w_i weights used to define different Z composites, but almost certainly there will be only one set of v_i, since there is typically only one X-type score. Both the w_i and v_i weights are viewed here as nominal weights specified by an investigator based on substantive considerations. Different investigators might legitimately specify different sets of weights based on different definitions/criteria for Z and/or X.

In this subsection, the Z_i and X_i are usually considered to be number-correct scores for tests of dichotomously-scored items (or number-of-points scores for polytomously-scored items). Sometimes, however, Z_i and X_i are specified as proportion-correct scores (or proportion-of-points scores). The scale-score analogues of Z_i and X_i are not considered explicitly in this section, although the basic principles discussed here apply to scale scores, as well.

Probably the most likely scenario for raw scores is: (i) Z_i and X_i are number-correct scores; (ii) the v_i are all 1; and (iii) the w_i are permitted to take on different values depending upon various criteria that may be of interest. In general, U_r will vary if different values are used for the individual w_i and/or v_i.

5.1.1 The w_i Weights

Suppose there are three full-length tests—Z_1, Z_2, and Z_3. For convenience, let us assume that $\sum w_i = 1$.

Table 5 provides six possible sets of weights for the w_i. For sets a, b, and c, $w_i = 1$ for one i and $w_i = 0$ for the other two i. These are the weights used in the examples in Section 4. They are the most likely choices in many practical situations, but other sets of w_i weights are certainly possible. (Note that $w_i = 0$ has nothing to do with whether or not Z_i contributes items to X, because contributions to X are reflected by the v_i, not the w_i.)

For set d, the Z_i are weighted equally. For set e, Z_1 and Z_2 are each weighted half as much as Z_3. Set f gives equal weight to Z_1 and Z_2 and no weight to Z_3.

There is no single linear transformation of one set of w_i weights in Table 5 that gives any one of the other sets of w_i weights. Therefore, except in trivial circumstances the relative utility indexes for these sets of weights will be different.

Assuming that $\sum w_i = 1$ is convenient for illustrative purposes, and for most practical purposes, but not theoretically necessary. Suppose, for example, that

Table 5: Illustrative Sets of w weights for a Z Composite of $k = 3$ Tests

Set	w_1	w_2	w_3	Comment
a	1	0	0	all weight on test 1
b	0	1	0	all weight on test 2
c	0	0	1	all weight on test 3
d	.33	.33	.33	equal weights for all tests
e	.25	.25	.50	test 3 weighted twice as much as tests 1 and 2
f	.50	.50	0	test 3 unweighted; tests 1 and 2 equally weighted

the set of weights is $w_1 = w_2 = 10$ and $w_3 = 0$. This is analogous to set f in Table 5. Indeed the relative utility indexes for two sets of w_i weights will be equal (assuming the v_i are unchanged), because the two sets of w_i weights are linear transformations of each other. However, the scale of Z will be much different for the two sets of weights. Consequently, care must be taken in interpreting results.

In the above discussion, the w_i weights have been discussed as if they were nominal weights determined a priori, presumably based on some context of contextual considerations. That is probably the most likely scenario. In theory, however, the w_i weights could be determined based on some statistical criterion or methodology. The most obvious possibility would be regression weights of some kind. The resulting scale for Z, however, is not likely to correspond with a reported score, which undermines the usefulness of such weights for users with access to reported scores, only.

5.1.2 The v_i Weights

Prespecifying the v_i requires careful thought, because they substantially influence the interpretation and psychometric properties of X. For example, assume $k = 3$ and suppose X_1, X_2, and X_3 contain 25, 20, and 10 dichotomously-scored items from Z_1, Z_2, and Z_3, respectively. Then, if the X_i are total scores and X is intended to be a total score ranging from 0 to 55, it follows that the three v_i should be set to 1. Of course, other v_i weights are possible. For example the weights could be $v_1 = v_2 = 1$ and $v_3 = 2$; if so, total scores on X would range from 0 to 65.

Continuing with the example in the previous paragraph, suppose the X_i are specified in terms of proportion-correct scores between 0 and 1. Then setting the three v_i to 1 will cause X to have a range of 0 to 3, which is not necessarily wrong per se, but it is not likely to be the investigator's intent. It is more likely that the investigator wants scores on X to range from 0 to 1, which can be achieved in numerous ways, including: (i) setting $v_1 = v_2 = v_3 = 1/3$; or (ii) setting $v_1 = 25/55$, $v_2 = 20/55$, and $v_3 = 10/55$. These different sets of weights will lead to different values for U_r.

When the X_i are proportion-correct scores, the v_i could be defined as

$$v_1 = 25,\ v_2 = 20,\ \text{ and }\ v_3 = 10,$$

which would transform the proportion-correct X_i scores such that the X composite has a range of 0 to 55. This is the same range as for the X composite obtained using total scores for X_i with

$$v_1 = v_2 = v_3 = 1.$$

There is no single linear transformation of the first set of v_i weights (25, 20, and 10) that gives the second set of v_i weights (all 1). Still, U_r will be the same for both sets of v_i weights *provided* the $\sigma^2(E_i)$ are for the intended X_i metric (i.e., number-correct or proportion-correct scores).

Strictly speaking, v_i can be set to 0 even when Z_i contributes items to the test X. Doing so is seldom sensible, however, because test X will include Z_i-type items, but the score X will not include the X_i score.

In short, there is no right answer to what the nominal v_i weights should be, and care should be taken in specifying them. In the author's experience, when the v_i are nominal weights, it is usually appropriate that they be chosen to obtain the intended range of scores on X once the X_i scores are defined.

5.2 Transformations of Component Parts of U_r

Since the raw-score relative utility index U_r is defined as the ratio of two squared correlations, $\rho^2(T_X, Z)$ and $\rho^2(T_X, X)$ (see Equation 17), U_r is unaffected by linear transformations of Z, X, and/or T_X. Note that a linear transformation of any composite variable affects its component parts. For example, if Z is transformed to $Z' = a + b\,Z$, then $Z' = a + b\sum_i w_i\,Z_i$.

Obviously, if Z, X, and/or T_X is/are transformed non-linearly, as they often are for scale scores, then it is almost certain that the U_r will change. Also, as discussed next, if the component parts of Z and/or X are linearly transformed, it is not necessarily true that U_r will be unchanged.

Suppose that $k = 3$ and: (a) there are different numbers of items (n_1, n_2, and n_3) that contribute to the Z_i; (b) there are different numbers of items (m_1, m_2, and m_3) that contribute to the X_i; and (c) $m_i < n_i$ for all i. If both the Z_i and the X_i are scored number-correct, then there will be some specific value for U_r that is dependent on the prespecified w_i and v_i. By contrast, if the Z_i and/or the X_i are scored proportion correct, then the U_r value will be different, except in trivial cases. In short, using *different* linear transformations for the individual Z_i and/or X_i does not necessarily mean that Z and/or X are linearly transformed, because Z and X depend upon the prespecified w_i and v_i, respectively.

5.3 Proportionality Assumption for Scale-score Error Variances

For scale scores, the $\sigma^2(\widetilde{E}_i)$ are required. Obtaining them is not always straightforward, however, since raw-to-scale score transformations for the X_i are seldom available. This matter is addressed here through adopting the proportionality assumption in Equation 24, which leads to the formula for $\sigma^2(\widetilde{E}_i)$ in Equation 25.

Equation 25 holds when $\widetilde{X}$ is a linear transformation of X. Logically, then, this assumption should hold approximately for non-linear transformations that do not depart too much from linearity. Furthermore, if most of the non-linearity is attributable to the relationship between true scores T_X and $T_{\widetilde{X}}$, the assumption seems reasonably plausible.

If doubts persist about the appropriateness of the proportionality assumption, relative utility might be estimated using different, plausible choices for $\sigma^2(\widetilde{E}_i)$—subject, of course, to the constraint that $\sum v_i^2 \sigma^2(\widetilde{E}_i) = \sigma^2(\widetilde{E})$. If the various choices give similar values for relative utility, concerns about the proportionality assumption diminish.

Suppose $w_i = 1$ for one of the Z_i and $w_i = 0$ for the other $k-1$ Z_i. This is the case considered in the computational example in Section 4.2 where $w_1 = 1$, and we examined whether H or R was a better estimate of true scores on H. For this example, one approach to examining the reasonableness of the proportionality assumption is to find the value of $\sigma^2(\widetilde{E}_1)$ that gives $\widetilde{U}_r = 1$ and compare this value of $\sigma^2(\widetilde{E}_1)$ to the proportionality-based $\sigma^2(\widetilde{E}_1)$.

Replacing Equation 30 in Equation 20 gives

$$\widetilde{U}_r = \left[\frac{\sigma^2(\widetilde{X})}{\sigma^2(\widetilde{Z})}\right] \left[\frac{\sigma(\widetilde{X},\widetilde{Z}) - \sigma^2(\widetilde{E}_1)}{\sigma^2(T_{\widetilde{X}})}\right]^2 . \tag{35}$$

Solving for $\sigma^2(\widetilde{E}_1)$ gives

$$\sigma^2(\widetilde{E}_1) = \sigma(\widetilde{X},\widetilde{Z}) - \sigma^2(T_X) \left[\frac{\sigma(\widetilde{Z})}{\sigma(\widetilde{X})}\right] \sqrt{\widetilde{U}_r},$$

and if $\widetilde{U}_r = 1$, then

$$\sigma^2(\widetilde{E}_1) = \sigma(\widetilde{X},\widetilde{Z}) - \sigma^2(T_X) \left[\frac{\sigma(\widetilde{Z})}{\sigma(\widetilde{X})}\right].$$

For the SAT© computational example in Section 4.2, the value of $\sigma^2(\widetilde{E}_1)$ that gives $\widetilde{U}_r = 1$ is

$$\sigma^2(\widetilde{E}_1) = 23.228 - 21.946\sqrt{\frac{24.671}{26.786}} = 2.166.$$

Recall that the actual value of $\widetilde{U}_r = .917$ resulted from $\sigma^2(\widetilde{E}_1) = 3.064$, which is over 40% larger that 2.166. Equation 35 clearly indicates that larger values of $\sigma^2(\widetilde{E}_1)$ lead to smaller values of $\widetilde{U}_r$. It follows that the rather large disparity (2.166 vs 3.064) lends credence to the conclusion the H is a better estimate of true scores on H than is R. If the disparity were quite small, an investigator might conclude that there is only weak support for the the use of H over R.

5.4 Special Case: Decisions About Subscores

Suppose that: (i) $w_i = 1$ for all k values of i; (ii) $v_i = 1$ for a single i, say $i*$; (iii) $v_i = 0$ for the remaining $k-1$ values of i; and (iv) the subset of items that contribute to $X_{i*} = X$ is the entire full-length test Z_{i*}. That is, rather than being a pseudo test, X is actually one of the tests in the battery. This is the case treated by Haberman (2008) and Brennan (2011) in their consideration of decisions about subscores. As Brennan (2011) shows, the Haberman/Brennan procedures lead to the same decision about whether X or Z is preferable as an estimate of T_X. The above special case also leads to the same decision, since U_r in this paper and the relative utility index in Brennan (2011) always lead to the same decision in this special case.

5.5 Multidimensionality

The theoretical framework in this paper is inherently multidimensional in at least two senses. First, if the Z_i do not conform to a multidimensional model, it might be argued that there is no psychometric reason for distinguishing among them, and we might as well concatenate all of them rather than pretend they measure different content or constructs. Second, the very fact that we create test X by selecting specific subsets of items, X_i, from the Z_i, suggests that there is an intended difference between the X_i-type items in each Z_i and the non-X_i-type items in each Z_i. If that is not true for at least some X_i, then the meaningfulness of X is undermined.

5.6 Caveats

The theoretical framework in this paper supports a decision to report X if $U_r < 1$; otherwise, Z is supported. The theory, however, does not directly address all validation issues involving the uses and interpretations of X scores (see, for example, Kane, 2013, and Brennan, 2013). Next, we consider two situations in which using only U_r to make a decision about reporting X may be at variance with another legitimate concern.

First, suppose $U_r > 1$ implying that reporting Z is preferred over X. Still, there may be other substantive reasons to prefer X over Z. One such reason could be that Z includes too many items that muddle the interpretation of scores for the intended X-type construct. Second, suppose $U_r < 1$ implying that reporting X is preferred over Z. Still, the X scores could be judged to be too unreliable (or contain too much error variance) for reporting purposes.

6 References

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

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

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

Brennan, R. L. (2011, November). *Utility Indexes for Decisions about Subscores.* (CASMA Research Report No. 33). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma)

Brennan, R. L. (2013). Commentary on "Validating the interpretations and Uses of Test Scores." *Journal of Educational Measurement, 50*, 74-83.

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. (Currently published by Oryx).

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*, pp. 204–229.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*, 1–73.

Kim, H. J., & Brennan, R. L. (2015, June). *Utility Indexes for Composite Scores: Matrix Formulation Including R Code and Examples* (CASMA Research Report No. 44). (Revised July 2016.) Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma)

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer-Verlag.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29*, 285–307.

Lee, W., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement, 37*, 1–20.

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