Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 42

Utility Indexes for Composite Scores

Robert L. Brennan

May 1, 2015

1The 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, Research Scientist at CASMA.

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

Abstract iv

1 Components of Reliability and $U$ 1

2 Relative Utility Index $U_r$ for Composite Scores 4

3 Estimating $U_r$ for Raw Scores and $\tilde{U}_r$ for Scale Scores 5
   3.1 Raw Scores ........................................... 5
   3.2 Scale Scores ........................................... 6
      3.2.1 Variance of $T_{X_i}$ ................................ 6
      3.2.2 Covariance of $T_{X_i}$ and $\tilde{Z}$ ............... 7
   3.3 Estimating $\sigma^2(\tilde{E})$ .............................. 7
      3.3.1 Linear Transformations .............................. 8
      3.3.2 Non-linear Transformations .......................... 8

4 Summary and Concluding Comments 8
   4.1 Definitions of $Z_i$, $X_i$, $w_i$, and $v_i$ .................. 9
      4.1.1 The $w_i$ Weights .................................. 9
      4.1.2 The $v_i$ Weights .................................. 10
   4.2 Transformations of Component Parts of $U_r$ ............... 11
   4.3 More than one Pseudo Test .............................. 11
   4.4 Special Case: Decisions About Subscores .................... 12
   4.5 Proportionality Assumption ............................. 12
   4.6 Multidimensionality ................................... 12
   4.7 Caveats ................................................. 13

5 References 13
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 $X$ is a subscore of $Z$, in the sense that $X$ is entirely contained within $Z$. Brennan (2011) considers a procedure for determining whether or not, in a certain sense, $Z$ is a better/worse estimate of true-score for $X$ (i.e., $T_X$) than $X$ 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 $X$ provides “more useful” information (in a certain sense) than that provided by $Z$.

In Brennan’s (2011) procedure, the operational definition of “more useful” involves a comparison of reliability of $X$, as defined by

$$
\rho_X^2 = \rho^2(T_X, X) = \left[ \frac{\sigma(T_X, X)}{\sigma(T_X) \sigma(X)} \right]^2, \tag{1}
$$

with

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

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

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.

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 tests in the battery. 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 focus 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

\[
Z = w_1 Z_1 + w_2 Z_2 + \cdots + w_k Z_k,
\]

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

\[
X = v_1 X_1 + v_2 X_2 + \cdots + v_k X_k,
\]

where the true and error scores associated with \( X \) are

\[
T = v_1 T_1 + v_2 T_2 + \cdots + v_k T_k,
\]

and

\[
E = v_1 E_1 + v_2 E_2 + \cdots + v_k E_k,
\]

and the \( v_i \) are nominal weights (\( v_i \geq 0 \) for all \( i \)) specified by the user.

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. This is especially true for \( Z \), which essentially serves as a potential replacement for \( X \). Clearly, \( Z \) 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 short, here \( X_i = T_i + E_i \) can be viewed as a shorthand version of \( X_i = T_{X_i} + E_{X_i} \).

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

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

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

\[
\sigma(v_i T_i, w_j Z_j) = v_i w_j \sigma((X_i + E_i), Z_j) = v_i w_j \sigma((X_i, Z_j) + (E_i, Z_j)) = v_i w_j \sigma(X_i, Z_j) = v_i w_j \sigma(X_i, Z_j).
\]
and the diagonal elements are

$$
\sigma(v_i T_i, w_i Z_i) = v_i w_i \sigma(T_i, Z_i)
$$

$$
= v_i w_i \sigma([X_i - E_i], Z_i)
$$

$$
= v_i w_i \sigma([X_i, Z_i] - (E_i, Z_i))
$$

$$
= v_i w_i \sigma(X_i, Z_i) - \sigma(E_i, Z_i)]
$$

$$
= v_i w_i \sigma(X_i, Z_i) - \sigma^2(E_i)\) \tag{9}
$$

$$
= v_i w_i \sigma(X_i, Z_i) - \sigma^2(E_i)\). \tag{10}
$$

It is important to remember that $E_i$ 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 not associated; (b) the covariance of $E_i$ and the errors for the unassociated 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 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

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

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

$$
\sigma^2(T_X) = \sigma^2(v_1 T_1 + v_2 T_2 + \cdots + v_k T_k)
$$

$$
= \sum_{i=1}^{k} v_i^2 \sigma^2(T_i) + \sum_{i \neq j} v_i v_j \sigma(T_i, T_j)
$$

$$
= \sum_{i=1}^{k} v_i^2 [\sigma^2(X_i) - \sigma^2(E_i)] + \sum_{i \neq j} v_i v_j \sigma(X_i, X_j),
$$

and

$$
\sigma^2(Z) = \sigma^2(w_1 Z_1 + w_2 Z_2 + \cdots + w_k Z_k)
$$

$$
= \sum_{i=1}^{k} w_i^2 \sigma^2(Z_i) + \sum_{i \neq j} w_i w_j \sigma(Z_i, Z_j).
$$

Similarly,

$$
\sigma^2(X) = \sigma^2(v_1 X_1 + v_2 X_2 + \cdots + v_k X_k)
$$

$$
= \sum_{i=1}^{k} v_i^2 \sigma^2(X_i) + \sum_{i \neq j} v_i v_j \sigma(X_i, X_j). \tag{14}
$$

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 $\sigma^2(E_i)$ 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

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

and

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

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, the estimation procedures for the different $\sigma^2(E_i)$ may be (and in many cases probably should be) different. 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 $\rho_X^2$ is

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

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 present here.

$U_r$ in Equation 17 can also be expressed as

$$U_r = \left[ \frac{\sigma^2(T_X, Z)}{\sigma^2(T_X)} \right] \left[ \frac{\sigma^2(T_X, Z)}{\sigma^2(T_X)} \right] = \left[ \frac{\sigma^2(T_X)}{\sigma^2(T_X)} \right], \quad (18)$$

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.

---

3Note 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 $\bar{U}_r$ 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.4

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

$$\sigma^2(E_i) = \sigma^2(X_i)(1 - \rho^2_{ai}),$$

can be used, where $\rho^2_{ai}$ is coefficient $\alpha$ 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.

4In 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.
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 \( \alpha \) (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 \( p \times (m \circ h) \), the D study design is \( p \times (M \circ H) \), 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).

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

\[
\bar{U}_r = \frac{\bar{U}}{\sigma^2(\bar{X})} = \frac{\sigma^2(T_{\bar{X}}, \bar{Z})}{\sigma^2(T_{\bar{X}}, \bar{X})} = \left[ \frac{\sigma^2(\bar{X})}{\sigma^2(\bar{Z})} \right] \left[ \frac{\sigma(\bar{T}_{\bar{X}}, \bar{Z})}{\sigma(\bar{T}_{\bar{X}})} \right]^2.
\]

#### 3.2.1 Variance of \( T_{\bar{X}} \)

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

\[
\sigma^2(T_{\bar{X}}) = \sigma^2(\bar{X}) - \sigma^2(\bar{E}),
\]

where \( \sigma^2(\bar{E}) \) 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.

---

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 \( \sigma^2(E_i) = \sigma^2(X_i)(1 - E_i \rho^2) \), where \( \sigma^2(X_i) \) is expressed in the intended metric, and \( E_i \rho^2 \) is the generalizability coefficient for \( X_i \) — i.e., the set of two passages associated with \( c_1 \).
3.2.2 Covariance of $T_{\tilde{X}}$ and $\tilde{Z}$

Because the $\tilde{X}_i$ 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 $\sigma(T_{\tilde{X}}, \tilde{Z})$. Often, a more useful approach is to use the following derivation:

$$\sigma(T_{\tilde{X}}, \tilde{Z}) = \sigma\left[\frac{\tilde{X} - \sum_{i=1}^{k} v_i \tilde{E}_i}{\sum_{i=1}^{k} w_i \tilde{Z}_i}\right]$$

$$= \sum_{i=1}^{k} w_i \sigma(\tilde{X}, \tilde{Z}_i) - \sigma\left[\frac{\sum_{i=1}^{k} v_i \tilde{E}_i}{\sum_{i=1}^{k} w_i \tilde{Z}_i}\right]$$

$$= \sum_{i=1}^{k} w_i \sigma(\tilde{X}, \tilde{Z}_i) - \sum_{i=1}^{k} v_i w_i \sigma(\tilde{E}_i, \tilde{Z}_i)$$

$$= \sum_{i=1}^{k} w_i \sigma(\tilde{X}, \tilde{Z}_i) - \sum_{i=1}^{k} v_i w_i \sigma^2(\tilde{E}_i).$$

Equation 22 follows from Equation 21 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 22 is $\sigma^2(\tilde{E})$, 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 $\sigma^2(\tilde{E}_i)$ 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.

Suppose we assume that error variances for raw and scale scores are proportional in the sense that

$$\frac{\sigma^2(\tilde{E}_i)}{\sigma^2(\tilde{E})} = \frac{\sigma^2(E_i)}{\sigma^2(E)}.$$  \hfill (23)

Then,

$$\sigma^2(\tilde{E}_i) = \frac{\sigma^2(\tilde{E})}{\sigma^2(E)} \sigma^2(E_i)$$  \hfill (24)

It follows from Equation 22 that

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

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

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

Given the theory outlined above, the only error variance for scale scores that is required is $\sigma^2(\tilde{E})$ in Equations 20 and 25. This is the error variance for the scale-score transformation of $X$. 

7
3.3.1 Linear Transformations

Suppose $\tilde{X} = a + bX$, which means that $\tilde{X}$ is a linear transformation of $X$. In this case, $\sigma^2(E) = b^2 \sigma^2(E)$. It follows from Equations 20 and 25 that

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

and

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

3.3.2 Non-linear Transformations

There are a number of procedures for obtaining $\sigma^2(E)$ 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 $\sigma(E)$ for these scale scores.

4 Summary and Concluding Comments

In this section, the discussion is generally couched in terms of the raw-score variables $Z_i, Z_i, X_i$, 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.

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 $\sigma^2(E_i)$ 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 $\sigma^2(E_i)$, 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.

4.1 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 \( v_i \).

4.1.1 The \( w_i \) Weights

Suppose there are three full-length tests—\( Z_1, Z_2, \) and \( Z_3 \). Table 1 provides six possible sets of weights for the \( w_i \). For set \( a \) the \( Z_i \) are weighted equally, which may seem like the most obvious possibility, but it is actually only one possibility. For set \( b \) the first two tests are weighted half as much as the third. Clearly, sets \( a \) and \( b \) give some weight to each of the three tests. The theory, however, does not require doing so. For example, set \( c \) gives no weight to the third test, and sets \( d-f \) place all weight on a single test. None of the \( Z \) composites associated with each set of weights is a linear transformation of any of the other \( Z \) composites. Therefore, the relative utility indexes for the sets of weights will be different.

The fact that \( w_3 = 0 \) in set \( c \) has nothing to do with whether or not \( Z_3 \) contributes items to \( X \), because contributions to \( X \) are reflected by the \( v_i \); \( Z_3 \) contributes items to \( X \) when \( v_3 > 0 \). In short, there is nothing inconsistent in using, say, \( v_1 = v_2 = v_3 = 1 \) with set \( c \). The same type of logic applies to sets \( d-f \), where \( Z \) is \( Z_1, Z_2, \) or \( Z_3 \), respectively.

For convenience, each set of \( w_i \) weights in Table 1 has been scaled so that \( \sum w_i = 1 \). Since \( U_r \) is invariant with respect to linear transformations of \( Z \), each of the three \( w \) weights for any set could be multiplied by the same constant.
without changing $U_r$. A similar statement holds for the $v_i$ weights used to define $X$.

4.1.2 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

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

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 that gives the second set of $v_i$ weights. Still, $U_r$ will be the same for both sets of $v_i$ weights provided that the $\sigma^2(E_i)$ are for the intended $X_i$ metric (i.e., number-correct or proportion-correct scores).

Table 1: Illustrative Sets of $w$ weights for a $Z$ Composite of $k = 3$ Tests

<table>
<thead>
<tr>
<th>Set</th>
<th>$w_1$</th>
<th>$w_2$</th>
<th>$w_3$</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>.33</td>
<td>.33</td>
<td>.33</td>
<td>equal weights for all tests</td>
</tr>
<tr>
<td>b</td>
<td>.25</td>
<td>.25</td>
<td>.50</td>
<td>test 3 weighted twice as much as tests 1 and 2</td>
</tr>
<tr>
<td>c</td>
<td>.50</td>
<td>.50</td>
<td>0</td>
<td>tests 1 and 2 equally weighted</td>
</tr>
<tr>
<td>d</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>all weight on test 1</td>
</tr>
<tr>
<td>e</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>all weight on test 2</td>
</tr>
<tr>
<td>f</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>all weight on test 3</td>
</tr>
</tbody>
</table>
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 the 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.

### 4.2 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 \( Z' = a + bZ \), 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 relative utility index will change. Non-linearity can sometimes occur in subtle ways, as discussed next.

Suppose that \( k = 3 \) and: (a) there are different numbers of items \((n_1, n_2, \text{ and } n_3)\) that contribute to the \( Z_i \); (b) there are different numbers of items \((m_1, m_2, \text{ and } m_3)\) that contribute to the \( X_i \); and (c) \( m_i < n_i \) for \( i = 1, 2, 3 \). If both the \( Z_i \) and the \( X_i \) are scored number-correct, then there will be some specific value for \( U_r \). By contrast, if the \( Z_i \) and/or the \( X_i \) are scored proportion correct, then the \( U_r \) value will be different (unless, of course, the \( w_i \) and/or the \( v_i \) are chosen to convert proportion-correct scores to number-correct scores.) The reason for the difference is that 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.

### 4.3 More than one Pseudo Test

Suppose \( Y \) is another pseudo test in the sense that it consists of sets of items from each of the \( Z_i \), but the sets for \( Y \) are different from the sets for \( X \). In addition to comparing both \( X \) and \( Y \) to various possible \( Z \) composites, there may be interest in comparing \( X \) and \( Y \) to each other. So, two additional questions may be of interest:

1. is \( X \) a better estimate of \( T_X \) than \( Y \), and
2. is \( Y \) a better estimate of \( T_Y \) than \( X \)?

For the first question, \( Y \) plays the role of \( Z \); for the second, \( X \) plays the role of \( Z \).

Consider \( U_r \) in Equation 18 with \( Z \) replaced by \( Y \) (the first question, above):

\[
U_r = \frac{\sigma^2(X)}{\sigma^2(Y)} \left[ \frac{\sigma(T_X, Y)}{\sigma^2(T_X)} \right]^2
= \frac{\sigma^2(X)}{\sigma^2(Y)} \left[ \frac{\sigma(X, Y)}{\sigma^2(X) - \sigma^2(E_X)} \right]^2,
\]

(26)
since $\sigma(T_X, Y) = \sigma(X - E_X, Y) = \sigma(X, Y)$ because there are no overlapping items in $X$ and $Y$. $U_r$ for the second question is obtained by interchanging $X$ and $Y$ in Equation 26.

### 4.4 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.

### 4.5 Proportionality Assumption

For scale scores, the $\sigma^2(E_i)$ 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 through adopting the proportionality assumption in Equation 23, which leads to the formula for $\sigma^2(E_i)$ in Equation 24.

Equation 24 holds when $\bar{X}$ 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 $T_{X'}$, 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 $\sigma^2(E_i)$—subject, of course, to the constraint that $\sum v_i^2 \sigma^2(E_i) = \sigma^2(E)$. If the various choices give similar values for relative utility, concerns about the proportionality assumption diminish.

### 4.6 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, 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.
4.7 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 \( X \) rather than \( Z \) is not supported. 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 \( X \) rather than \( Z \) is supported. Still, the \( X \) scores could be judged to be too unreliable (or contain too much error variance) for reporting purposes.

5 References


