Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 33

Utility Indexes for Decisions about Subscores[1]

Robert L. Brennan[2]

November 4, 2011
Revised December 28, 2012

[1]The research reported here was supported, in part, by a contract with the College Board. The author expresses his gratitude to Won-Chan Lee, Michael T. Kane, and Kevin Sweeney for their thoughtful comments about previous versions of this report.

[2]Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Director, Center for Advanced Studies in Measurement and Assessment (CASMA).

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

Contents

Dedication

This paper is dedicated to the memory of Leonard S. Feldt, who died only a few days before the initial version of this report was finalized. In 1994 Leonard retired as E. F. Lindquist Professor in the College of Edcucation at the University of Iowa. At the time of his death, Leonard was professor emeritus, but he still came into work practically every day.

Throughout his career, Leonard was actively involved in research on classical test theory. It would be difficult to identify anyone in his cohort who made more contributions to classical test theory, and perhaps even more difficult to identify anyone who was a better teacher of classical test theory.

I had planned to ask Leonard to review this paper, and I would have greatly valued his input. That was not to be. Leonard will be missed, but I feel sure that his contributions to the field of measurement will endure through the many students and researchers he influenced, including me.

Abstract

For many testing programs there is considerable pressure to report subscores for diagnostic purposes, especially to identify areas of strength and weakness. Very frequently, however, subscores contain relatively few items, which usually leads to low reliability. Since subscores are often at least moderately correlated with other subscores, when a particular subscore has low reliability it is rather natural to consider using the total score as a proxy. Doing so involves "borrowing" information from other subscores in some sense. Most of the current approaches in the literature address this matter, in theory or practice, by using Kelley's regressed-score estimates.

By contrast, in this paper, an approach is suggested that is based entirely on classical test theory and conventional notions of reliability, without resort to Kelley's regressed score estimates. This approach yields statistics that provide explicit guidance about whether or not to report a subscore, and what, if anything, might be done if these statistics do not support reporting a particular subscore.

It is often the case that users of test scores want (indeed, demand) that subscores be reported for diagnostic purposes, along with total test scores. Often, however, the subscores are an afterthought in the sense that the test developers did not originally plan for them and, consequently, the specifications for test development do not control for subscores nearly as well as for the total score. Most importantly, very frequently subscores contain relatively few items. These circumstances raise a number of psychometric concerns.

For example, subscores with relatively few items usually have relatively low reliability, but practical constraints such as cost and testing time may preclude increasing the number of items for subscores. Since subscores are often at least moderately correlated, for any particular subscore with low reliability it is rather natural to consider the consequences of using the total score as a proxy for the particular subscore, which involves "borrowing" information from other subscores.

How to operationalize this general notion of "borrowing" information is not so obvious, however. Haberman (2008) and Wainer et al, (2001) provide two approaches that are based on, or motivated by, Kelley's regressed-score estimates (RSEs) (Kelley, 1947).[3] RSEs have a long tradition in psychometrics, but they have some potentially problematic characteristics when they are applied to individual examinees. For example, high-scoring examinees get scores lowered toward the mean, and low scoring examinees get scores increased toward the mean. Also, RSEs are biased estimates; that is, for a given examinee, the expected value of the RSEs over replications of the measurement procedure does not equal the examinee's true score, as it does in classical theory. In addition, RSEs assume a linear regression; no such assumption is required in classical theory. These and other matters concerning RSEs are discussed by Brennan (2012).

The comments in the previous paragraph do not mean that the Haberman (2008) and Wainer et al. (2001) approaches are wrong per se. In this paper, however, a different approach is suggested that is based entirely on classical test theory and conventional notions of reliability, and does not rely on RSEs. This approach yields statistics that provide explicit guidance about whether or not to report a subscore, and what, if anything, might be done (along with appropriate caveats) if these statistics do not support reporting a particular subscore.

It is important to note at the outset that there is no psychometric justification for reporting subscores if the total scores satisfy unidimensionality assumptions, especially the very strict assumptions of unidimensional item response theory. If all items measure the same trait or proficiency, no subset of items provides a measure of anything other than that trait or proficiency (or

[3]The Wainer et al. (2001) approach directly considers the use of RSEs in conjunction with subscore reporting. By contrast, Haberman (2008) mainly focuses on using the logic behind RSEs (particularly the percentage reduction in mean-square error) as a basis for judgments about reporting subscores. For Kelley (1947) the notion of RSEs was restricted to the linear regression of true scores on the corresponding observed scores. In this paper, in Haberman (2008), and in Wainer et al. (2001), RSEs refer generically to any linear regression in which true score is the independent variable.

random noise). Nothing in this paper circumvents this restriction.

1 Basic Equations from Classical Test Theory

This paper considers three observed-score random variables: the focal variable X, the "total" variable Z, and Y, which is the non-X component of Z.[4] It is assumed here that each of these observed-score variables follows the dictates of classical test theory as described extensively by Feldt and Brennan (1989), Haertel (2006), and Lord and Novick (1968).[5]

Each of these variables can be decomposed into true-score (T) and error (E) random variables as follows:

$$X = T_X + E_X, \tag{1}$$

$$Y = T_Y + E_Y, \tag{2}$$

and

$$Z = X + Y = (T_X + E_X) + (T_Y + E_Y) = (T_X + T_Y) + (E_X + E_Y). \tag{3}$$

The canonical definition of reliability for X is

$$\rho^2(T_X, X) = \left[\frac{\sigma(T_X, X)}{\sigma(T_X)\,\sigma(X)}\right]^2, \tag{4}$$

where

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

since $\sigma(T_X, E_X) = 0$ in classical test theory. It follows that

$$\rho^2(T_X, X) = \left[\frac{\sigma^2(T_X)}{\sigma(T_X)\,\sigma(X)}\right]^2 = \frac{\sigma^2(T_X)}{\sigma^2(X)}. \tag{5}$$

We will occasionally abbreviate $\rho^2(T_X, X)$ as ρ_X^2 when doing so does not create any ambiguity. Replacing X with Z in Equations 4 and 5 gives equations for the reliability of Z, which we sometimes abbreviate ρ_Z^2.

In effect, in Equations 4 and 5, X serves as an estimator of T_X. This paper addresses the following questions, among others:

1. Is Z a better/worse estimator of T_X than X, from the perspective of reliability; and

2. How can we quantify the extent to which (1) is true?

[4] These are the same variables that Haberman (2008) considers.

[5] While all of these references draw careful distinctions between assumptions and results derived from assumptions, here we sometimes overlook these distinctions.

2 Utility Index

Rather than using X as an estimator of T_X, suppose we use Z to estimate T_X. Furthermore, to quantify how well Z serves as an estimator of T_X, suppose we use an index that has the same form as Equation 4, namely,

$$\rho^2(T_X, Z) = \left[\frac{\sigma(T_X, Z)}{\sigma(T_X)\,\sigma(Z)}\right]^2, \tag{6}$$

with a range of 0 to 1. We call $\rho^2(T_X, Z)$ an index of utility (U) — specifically, an index that quantifies the utility of using Z as an estimator of T_X. We now consider alternative formulas for $U = \rho^2(T_X, Z)$.

Note that

$$\begin{aligned}
\sigma(T_X, Z) &= \sigma[T_X, (T_X + T_Y) + (E_X + E_Y)] \\
&= \sigma(T_X, T_X) + \sigma(T_X, T_Y) + \sigma(T_X, E_X) + \sigma(T_X, E_Y) \\
&= \sigma^2(T_X) + \sigma(T_X, T_Y).
\end{aligned}$$

It follows that the utility index can be expressed as:

$$\begin{aligned}
U &= \rho^2(T_X, Z) && (7)\\
&= \left[\frac{\sigma^2(T_X) + \sigma(T_X, T_Y)}{\sigma(T_X)\,\sigma(Z)}\right]^2 && (8)\\
&= \left[\frac{\sigma(T_X)}{\sigma(Z)} + \rho(T_X, T_Y)\,\frac{\sigma(T_Y)}{\sigma(Z)}\right]^2. && (9)
\end{aligned}$$

Clearly, all other things being equal, U gets larger as the disattenuated correlation $\rho(T_X, T_Y)$ gets larger; this is discussed in more depth in Section 3.3.

2.1 Expressing U Without True-Score Parameters

Equations 7–9 involve true-score parameters. Here we derive a formula that is usually simpler to use for estimation. First, note that

$$\begin{aligned}
\sigma(X, Y) &= \sigma(T_X + E_X, T_Y + E_Y) \\
&= \sigma(T_X, T_Y) + \sigma(T_X, E_Y) + \sigma(E_X, T_Y) + \sigma(E_X, E_Y) \\
&= \sigma(T_X, T_Y),
\end{aligned}$$

and

$$\sigma(X, Y) = \sigma(X, Z - X) = \sigma(X, Z) - \sigma^2(X).$$

It follows that

$$\sigma(T_X, T_Y) = \sigma(X, Y) = \sigma(X, Z) - \sigma^2(X),$$

and from Equation 5, $\sigma^2(T_X) = \rho_X^2\,\sigma^2(X)$. Therefore, using Equation 8

$$\begin{aligned} U &= \frac{[\rho_X^2\,\sigma^2(X) + \sigma(X,Z) - \sigma^2(X)]^2}{\rho_X^2\,\sigma^2(X)\,\sigma^2(Z)} \\ &= \frac{[\sigma(X,Z) - \sigma^2(X)(1-\rho_X^2)]^2}{\rho_X^2\,\sigma^2(X)\,\sigma^2(Z)} \\ &= \frac{[\sigma(X,Z) - \sigma^2(E_X)]^2}{\rho_X^2\,\sigma^2(X)\,\sigma^2(Z)}, \end{aligned} \tag{10}$$

which is expressed in terms of parameters that are typically straightforward to estimate, or even readily available in technical manuals.

The measurement error variance, $\sigma^2(E_X)$, cannot be observed directly. It is usually estimated as $\hat{\sigma}^2(E_X) = \hat{\sigma}^2(X)(1-\hat{\rho}_X^2)$, which raises questions about how the reliability of X is estimated. In the majority of cases probably coefficient alpha would be used because it is so readily available, but the theory here is silent about which reliability coefficient should be used. This is discussed further in Section 5.4.

2.2 U as a Disattenuated Correlation

The expression $U = \rho^2(T_X, Z)$ is a type of squared disattenuated correlation; i.e., the square of a correlation corrected for the unreliability in X. However, $\rho(T_X, Z)$ is *not* a disattenuated correlation in the usual sense, because X is included in Z. Specifically, if $\rho(T_X, Z)$ were a disattenuated correlation in the usual sense, then it would be true that

$$\rho(T_X, Z) = \frac{\sigma(T_X, Z)}{\sigma(T_X)\sigma(Z)} = \frac{\sigma(X, Z)}{\sigma(T_X)\sigma(Z)} = \frac{\rho(X, Z)}{\rho_X}.$$

Here, however, the above sequence of equalities is *not* true because the numerators of the second and third terms are not equal. That is,

$$\sigma(T_X, Z) = \sigma(X - E_X, Z) = \sigma(X, Z) - \sigma(E_X, Z) \neq \sigma(X, Z) \tag{11}$$

since

$$\sigma(E_X, Z) = \sigma(E_X, X + Y) = \sigma(E_X, X) \neq 0$$

by the assumptions of classical test theory.[6]

2.3 Invariance of U with Respect to Linear Transformations

Because U is the squared correlation of T_X and Z, it is invariant with respect to linear transformations of T_X and/or Z. So, for example, U is unchanged if Z is

[6]The usual assumption that $\sigma(T_X, E_X) = 0$ is equivalent to assuming that the expected value of E_X over persons is 0, *provided* examinees are not selected based on X (see Feldt & Brennan, 1989, p. 109, and Haertel, 2006, p. 69). Conversely, if examinees are selected on the basis of X, then (except in trivial cases) the expected value of E_X is non-zero, which means that the regression of E_X on X is non-zero, and, hence $\sigma(E_X, X)$ is non-zero.

the average of X and Y, rather than the sum of X and Y. In previous sections we have explicitly assumed that $Z = X + Y$ simply for convenience, and we will generally continue to do so in subsequent sections. In some testing programs, however, Z is the mean of the component parts (e.g., the ACT Assessment).

Although U is invariant with respect linear transformations of Z, Equation 10 is expressed assuming that $Z = X + Y$, which means that computations should be done using $Z = X + Y$. In the Appendix (Section 6.1) an equation is provided, corresponding to Equation 10, in which linearly transformed values of Z are used directly.

2.4 Minimum and Maximum Values Allowing only $\rho(T_X, T_Y)$ to Vary

Obviously, the minimum and maximum values of U depend on what is allowed to vary in U. Here we will assume that only the disattenuated correlation $\rho(T_X, T_Y)$ is allowed to vary, and we derive minimum and maximum values using Equation 9.

Assuming $\rho(T_X, T_Y)$ is non-negative, U achieves its minimum value when $\rho(T_X, T_Y) = 0$ or, equivalently, $\sigma(T_X, T_Y) = 0$. In this case, $\sigma(X, Y) = 0$, which means that $\sigma^2(Z) = \sigma^2(X) + \sigma^2(Y)$. It follows that

$$\begin{aligned} \min(U) &= \left[\frac{\sigma(T_X)}{\sigma(Z)}\right]^2 \\ &= \frac{\sigma^2(T_X)}{\sigma^2(X) + \sigma^2(Y)} \qquad (12) \\ &= \frac{\sigma^2(T_X)}{\sigma^2(T_X) + [\sigma^2(E_X) + \sigma^2(Y)]}, \qquad (13) \end{aligned}$$

which is less than ρ_X^2 whenever $\sigma^2(Y) \geq 0$. In short, $\min(U)$ occurs when $\rho(T_X, T_Y) = 0$, and, in this case, the use of Z rather than X adds nothing but additional noise to X, with the additional noise being $\sigma^2(Y)$.

Equation 9 for U achieves its maximum value when $\rho(T_X, T_Y) = 1$, which means that $\sigma(T_X, T_Y) = \sigma(T_X)\,\sigma(T_Y)$. Therefore,

$$\begin{aligned} \max(U) &= \left[\frac{\sigma(T_X) + \sigma(T_Y)}{\sigma(Z)}\right]^2 \qquad (14) \\ &= \frac{\sigma^2(T_X) + \sigma^2(T_Y) + 2\,\sigma(T_X)\,\sigma(T_Y)}{\sigma^2(Z)} \\ &= \frac{\sigma^2(T_X) + \sigma^2(T_Y) + 2\,\sigma(T_X, T_Y)}{\sigma^2(Z)} \\ &= \frac{\sigma^2(T_Z)}{\sigma^2(Z)} \\ &= \rho_Z^2, \qquad (15) \end{aligned}$$

which is the reliability of Z, $\rho^2(T_Z, Z)$. This maximum value is larger than ρ_X^2 since Z includes X, and we have obtained the maximum value under the assumption that $\rho(T_X, T_Y) = 1$.

Assuming that $\rho(T_X, T_Y) = 1$ is equivalent to assuming that T_X and T_Y are congeneric (see Feldt & Brennan, 1989, p. 110–111; Haertel, 2006, p. 71); i.e., $T_X = a + b\,T_Y$, where $b > 0$ is required.[7]

Setting $a = 0$ and $b = 1$ is consistent with the stricter assumption of classically-parallel forms. When X and Y are classically parallel and of equal length, the double-length Spearman-Brown formula applies and

$$\max(U) = \frac{2\,\rho_X^2}{1 + \rho_X^2}.$$

When X and Y are classically parallel except for length, the general form of the Spearman-Brown formula holds:

$$\max(U) = \frac{g\,\rho_X^2}{1 + (g-1)\rho_X^2}, \tag{16}$$

where g is the proportional length of $Z = X + Y$ relative to X.

3 Relative Utility Index

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 to some statistic that quantifies the merits of using X, alone. An obvious comparative statistic is ρ_X^2, the reliability of X. We could use a direct comparison of U and ρ_X^2, but a potentially more useful approach involves an extension of a relatively unknown result from classical test theory.

Let $\rho^2(T_X, X_n)$ and $\rho^2(T_X, X_m)$ be reliabilities for test forms of length n and m, respectively. Lord and Novick (1968, p. 119) show that, under classical test theory assumptions,

$$m = n\left[\frac{\rho^2(T_X, X_m)/(1 - \rho^2(T_X, X_m))}{\rho^2(T_X, X_n)/(1 - \rho^2(T_X, X_n))}\right], \tag{17}$$

which results from an algebraic manipulation of two applications of the Spearman-Brown formula. Note that n and m need not be integers. Dividing both sides of Equation 17 by n gives

$$f = \frac{m}{n} = \frac{\rho^2(T_X, X_m)/(1 - \rho^2(T_X, X_m))}{\rho^2(T_X, X_n)/(1 - \rho^2(T_X, X_n))}, \tag{18}$$

[7] With the additional assumption that $\sigma^2(E_X)$ and $\sigma^2(E_Y)$ follow the dictates of classical theory, as discussed by Feldt and Brennan (1989, p. 112) and Haertel (2006, p.72), ρ_Z^2 is given by the Angoff-Feldt Coeffcient $\sigma(X,Y)/[\lambda_X\lambda_Y\,\sigma^2(Z)]$ where $\lambda_X = [\sigma^2(X) + \sigma(X,Y)]/\sigma^2(Z)$ and $\lambda_Y = 1 - \lambda_X$.

where $|f-1|$ is the proportional increase/decrease in the length of X associated with an increase/decrease in reliability from $\rho^2(T_X, X_n)$ to $\rho^2(T_X, X_m)$.

A natural extension of Equation 18 is to replace X_m with Z in order to quantify the fractional increase/decrease in test length attributable to using Z as the observed score rather than X_n. This logic leads to

$$\widetilde{U} = \frac{\rho^2(T_X, Z)/(1-\rho^2(T_X, Z))}{\rho^2(T_X, X_n)/(1-\rho^2(T_X, X_n))}, \tag{19}$$

where $\widetilde{U}$ stands for the relative utility of using Z instead of X_n. Here X_n plays the role of X in previous sections, and $\rho^2(T_X, Z)$ is the utility index U. Therefore, a notationally simplified version of Equation 19 is

$$\widetilde{U} = \frac{U/(1-U)}{\rho_X^2/(1-\rho_X^2)}. \tag{20}$$

It follows that $100|1-\widetilde{U}|\%$ is the percentage change in the length of X needed to obtain a reliability equal to U. Note that

1. $\widetilde{U}$ can range from 0 to ∞;
2. $\widetilde{U} \leq 1$ implies that the use of X is supported with respect to reliability issues, as discussed in this paper; and
3. $\widetilde{U} > 1$ when $U > \rho_X^2$, which implies that the use of X is not supported with respect to reliability issues, and the use of Z rather than X *may* be justified.

Clearly $\widetilde{U} = 1$ is a benchmark for deciding whether of not there is potential merit (i.e., utility) in using X, alone.

When $\widetilde{U} > 1$, the use of Z rather than X might be justified in that, with respect to reliability, the use of Z effectively increases the length of X by $100(\widetilde{U}-1)\%$. Specifically, $100(\widetilde{U}-1)\%$ is the percentage increase in test length for X needed to obtain a reliability consistent with the X-type information in Z. Using Z as a proxy for increasing the length of X does not mean that the items are actual items in X, or that they necessarily look like items in X. Rather, the notion is that there is X-type information in Y that is psychometrically equivalent (in a reliability sense) to some number of X-type items.

Of course, even if using Z rather than X can be justified in terms of reliability issues (i.e., $\widetilde{U} > 1$), it does not necessarily follow that other practical considerations and validation concerns would support using Z. These matters are discussed further in Section 5.

3.1 Minimum and Maximum Values of $\widetilde{U}$ Allowing only $\rho(T_X, T_Y)$ to Vary

When $\rho(T_X, T_Y) = 0$, using Equation 13 it can be shown that

$$\min\left(\frac{U}{1-U}\right) = \frac{\sigma^2(T_X)}{\sigma^2(E_X)+\sigma^2(Y)}.$$

Also, it is easy to show that

$$\frac{\rho_X^2}{1-\rho_X^2} = \frac{\sigma^2(T_X)}{\sigma^2(E_X)}.$$

It follows from Equation 20 that

$$\min(\widetilde{U}) = \frac{\sigma^2(E_X)}{\sigma^2(E_X)+\sigma^2(Y)}, \tag{21}$$

which is positive in all but trivial cases.

When $\rho(T_X, T_Y) = 1$, the congeneric assumption holds, and replacing Equation 15 in Equation 20 gives

$$\max(\widetilde{U}) = \frac{\rho_Z^2/(1-\rho_Z^2)}{\rho_X^2/(1-\rho_X^2)}, \tag{22}$$

which is the ratio of two signal/noise ratios (the numerator for Z and the denominator for X).

Under the stricter classically parallel assumptions, using Equation 16 it is easy to show that

$$\frac{\rho_Z^2}{1-\rho_Z^2} = \frac{g\,\rho_X^2}{1-\rho_X^2}.$$

It follows that

$$\max(\widetilde{U}) = \frac{g\,\rho_X^2/(1-\rho_X^2)}{\rho_X^2/(1-\rho_X^2)} = g, \tag{23}$$

which equals f in Equation 18 under classically parallel assumptions.

3.2 Practical Advantages of Using $\widetilde{U}$

The use of $\widetilde{U}$ has several practical advantages. For example, $\widetilde{U} \leq 1$ provides a clear basis for arguing the merits of reporting X as a separate score (provided, of course, that X itself is sufficiently reliable for the intended interpretations). Also, when $\widetilde{U} > 1$, $100(\widetilde{U}-1)\%$ provides the percentage increase in test length for X needed to obtain a reliability consistent with the X-type information in Z. Those responsible for testing programs are generally well-informed about the testing time and personnel/administrative costs involved in lengthening a test. Consequently, when $\widetilde{U} > 1$, $100(\widetilde{U}-1)\%$ can provide a practical basis for answering "Is it worth it?" questions.

3.3 $\widetilde{U}$ as a Function of ρ_X^2 and $\rho(T_X, T_Y)$

For interpretative purposes, it is of interest to isolate the role of reliability and $\rho(T_X, T_Y)$ in determining $\widetilde{U}$. This permits us to answer the question, "For a specified value of $\rho(T_X, T_Y)$, how large must reliability be in order that $\widetilde{U} \leq 1$, in which case, the use of X is justified, given the perspective adopted in this paper"?

It is clear from Equation 9 and 20 that, in general, $\widetilde{U}$ is a function of $\sigma(T_X)$, $\sigma(T_Y)$, $\sigma(Z)$, $\rho(T_X, T_Y)$, and ρ_X^2. Alternatively, $\widetilde{U}$ is a function of $\sigma(X)$, $\sigma(Y)$, $\sigma(Z)$, ρ_X^2, ρ_Y^2, and $\rho(T_X, T_Y)$. Given the large number of variables involved in $\widetilde{U}$, the question in the previous paragraph does not have a simple, general answer. The answer is tractable, however, if X and Y have equal standard deviations and equal reliabilities.[8] In this case, it is shown in the Appendix (Section 6.2) that

$$U = \frac{\rho_X^2[1 + \rho(T_X, T_Y)]^2}{2[1 + \rho_X^2\,\rho(T_X, T_Y)]}$$

and

$$\widetilde{U} = \frac{[1 + \rho(T_X, T_Y)]^2(1 - \rho_X^2)}{2 - \rho_X^2\,[1 + \rho(T_X, T_Y)^2]}.$$

Furthermore, it is shown in the Appendix (Section 6.2) that, for a specified value of $\rho(T_X, T_Y)$, the use of X is justified (i.e. $\widetilde{U} \leq 1$) if

$$\rho_X^2 \geq 1 - \left[\frac{1 - \rho(T_X, T_Y)^2}{2\,\rho(T_X, T_Y)}\right]. \tag{24}$$

Also, it is shown in the Appendix (Section 6.2) that, for a specified value of ρ_X^2, the use of X is justified when

$$\rho(T_X, T_Y) \leq (\rho_X^2 - 1) + \sqrt{(1 - \rho_X^2)^2 + 1}. \tag{25}$$

Using Equation 24, Figure 1 provides plots of relative utility as a function of reliability for six different values of $\rho(T_X, T_Y)$ ranging from .50 (at the bottom) to .95 (at the top). A solid horizontal line at $\widetilde{U} = 1$ provides a reference line for deciding whether or not to report X. Except for qualifications treated in Section 5, for values of $\widetilde{U}$ above the line, reporting X is not justified given the perspective adopted in this paper. By contrast, for values of $\widetilde{U}$ on or below the line, reporting X is justified. It is clear from Figure 1 that:

- as $\rho(T_X, T_Y)$ gets larger, ρ_X^2 must get larger to justify reporting X;
- when $\rho(T_X, T_Y)$ is very large, it is unlikely that that reporting X will be justified; and
- as reliability gets smaller, it becomes less likely that reporting X will be justified.

The assumptions leading to Equation 24 (equal standard deviations and equal reliabilities) are stringent, but not too unreasonable in some testing programs. For example, prior to March 2005, for the College Board's SAT examination, there were two major subscores (Verbal and Mathematics), the standard deviations were set equal for a particular population at a particular point in time, and the reliabilities were similar.

[8] This does not mean that X and Y are classically parallel; for that to be true $\rho(T_X, T_Y)$ would need to be 1.

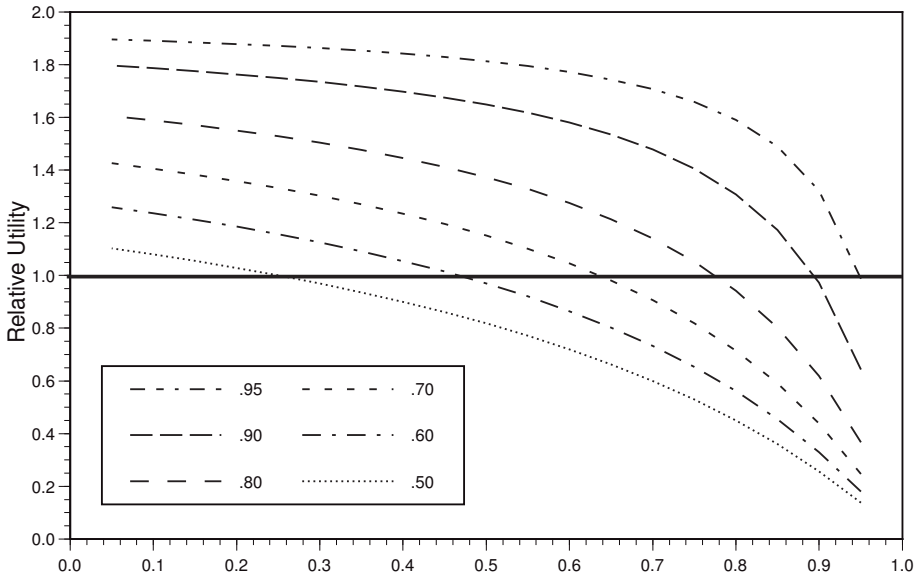

Figure 1: Relative utility ($\widetilde{U}$) as a function of reliability for X and the disattenuated correlation $\rho(T_X, T_Y)$. Reporting X is supported if $\widetilde{U} \leq 1$.

3.4 Using Z when $\widetilde{U} > 1$

When $\widetilde{U} > 1$, Z appears to provide more X-type information than X alone. That does not mean, however, that reporting Z as a substitute for X is necessarily the most reasonable choice to make. For example, if the reliability of X is judged to be sufficiently large, then X might be reported even if $\widetilde{U} > 1$, because X is not confounded by the non X-type information in Z.

When $\widetilde{U} > 1$ and X is judged to be relatively unreliable, there are two possible courses of action: (a) do nothing; i.e., do not report X and do not claim that Z should be used as a proxy for T_X; or (b) use Z as a proxy for T_X with some appropriate transformation of Z. The do-nothing alternative may be the best course of action, especially for tests that are known to be essentially unidimensional.

If, however, the Z-proxy solution is selected, a transformation of Z is likely to be necessary because, since $Z = X + Y$, the statistical characteristics of Z can be quite different from those of X. For example, suppose X and Z are raw scores, and the mean of Z is twice the mean of X. Claiming that untransformed Z scores can be used as a proxy for T_X scores likely will lead to inflated perceptions of examinees' proficiency on the construct measured by X. As discussed next, a transformation of Z might be considered to mitigate this problem. This discussion is intended to be illustrative, only; in practice, many issues not treated here would need to be studied to ensure that the reported

scores are satisfactory from reliability and validity perspectives.

3.4.1 Raw-score Linear Linking Transformation

Consider the raw-score linear transformation $X' = a + b\,Z$, and suppose $b = \sigma(X)/\sigma(Z)$ and $a = \overline{X} - b\,\overline{Z}$; i.e.,

$$X' = \left\{\overline{X} - \left[\frac{\sigma(X)}{\sigma(Z)}\right]\,\overline{Z}\right\} + \left[\frac{\sigma(X)}{\sigma(Z)}\right]\,Z. \tag{26}$$

Since the same examinees take both X and Z, this is a linking transformation using a single-group design (see Kolen & Brennan, 2004, pp. 15ff). The variables X' and X have the same mean and variance, but the rank ordering of persons in terms of X' and X is at least somewhat different in all but trivial cases. (By construction, the rank ordering of persons based on X' is the same as the rank ordering of persons based on Z.)

The error associated with using X' as an estimator of T_X is $E_{X'} = X' - T_X$. As shown in the Appendix (Section 6.4) , $\boldsymbol{E}(X' - T_X) = 0$, and

$$\text{RMSE}(E_{X'}) = \sqrt{2\,\sigma^2(X)[1 - \rho_X\,\rho(T_X, Z)] - \sigma^2(E_X)}, \tag{27}$$

which is less than $\sigma(E_X)$ whenever $\widetilde{U} > 1$. This is one sense in which using Z might be preferable to using X when $\widetilde{U} > 1$.

The scaling in Equation 26 causes $\sigma(E_X)$ and $\text{RMSE}(E_{X'})$ to be comparable in the sense that X and X' have the same mean and standard deviation. The fact that $\text{RMSE}(E_{X'}) < \sigma(E_X)$ can be viewed as a consequence of the fact that using Z effectively increases the length of X by $100(\widetilde{U} - 1)\%$.

3.4.2 Scale-score Linear Linking Transformation

Suppose that:

- $\widetilde{U} > 1$ for X, suggesting that reporting X scores is not supported;
- $\widetilde{U} \leq 1$ for Y suggesting that reporting Y scores is supported, and
- Y has considerably more items than X.

Even if the raw-score linear linking procedure in Section 3.4.1 is applied to X scores, almost certainly it will be difficult to compare the resulting X' and Y scores to assess strengths and weaknesses for individual examinees. The essential problem is that X' and Y will not be on the same scale. They can be put on a common scale, however, using the same linear transformation. For example,

$$SS(X') = 50 + 10\left[\frac{X' - \overline{X'}}{\sigma(X')}\right] \tag{28}$$

and

$$SS(Y) = 50 + 10\left[\frac{Y - \overline{Y}}{\sigma(Y)}\right]$$

provide scale scores that have the same mean (50) and standard deviation (10) for both X' and Y. For $SS(X')$ the RMSE can be obtained by multiplying Equation 6.4 by 10; for $SS(Y)$, the standard error of measurement is $10\,\sigma(E_Y)$.

When $X' = a + b\,Z$, and X' is then transformed using some specific linear scale-score transformation, such as $SS(\star)$, the resulting scale scores are the same for any values of a and b. For example, if X' were RSEs, and the scale-score transformation for RSEs were $SS(\star)$, the resulting RSE scale scores would be the same as those obtained above using X' in Section 3.4.1 in conjunction with Equation 28. So, when the intent of reporting subscores is to assess individual's strengths and weaknesses relative to some population, there is no particular advantage of using RSEs rather than the above linking transformation, or vice-versa.

3.4.3 Cautions

The above discussion of linking transformations should not be interpreted as a recommendation about what should be done when $\widetilde{U} > 1$ (Z appears to provide more X-type information than X alone). The discussion is merely a suggestion for consideration. As noted previously, many issues not treated here would need to be studied to ensure that the reported scores are satisfactory from both reliability and validity perspectives.

Suppose $\widetilde{U} > 1$ for more than one subscore, say, for example, X_1 and X_2. If the procedures outlined in Sections 3.4.1 and 3.4.2 are employed, then the rank ordering of examinees based on X_1' and X_2' will be the same, as will the rank orderings based on $SS(X_1')$ and $SS(X_2')$. This may be quite problematic depending on the uses to be made of the scores.

In any case, if the procedures in Sections 3.4.1 and 3.4.2 (or anything like them) are employed, then it would seem prudent (and probably imperative) that users be notified accordingly to guard against potential misues.

3.5 Relationships with Haberman's Statistics

Using RSEs, Haberman (2008) provides statistics for deciding when to report X. For consistency with Haberman (2008), in this section attention is focused on a specific true score, τ_X. Also, to simplify notation, Haberman's S_X is simply designated X, and S_Z is designated Z. As shown in the Appendix, (Section 6.3), Haberman's (2008) $\psi(\tau_X|\mathrm{L}(\tau_X|S_X))$ statistic is algebraically equivalent to ρ_X^2, and his $\psi(\tau_X|\mathrm{L}(\tau_X|S_Z))$ statistic is algebraically equivalent to $U = \rho^2(T_X, Z))$, although Haberman (2008) makes substantially different assumptions from those made in this paper.

Haberman's (2008) decision rule supports reporting X if

$$\psi(\tau_X|\mathrm{L}(\tau_X|X)) \geq \psi(\tau_X|\mathrm{L}(\tau_X|Z)),$$

which is equivalent to $\rho_X^2 \geq U$, which in turn is equivalent to $\widetilde{U} \leq 1$ (see Section 6.3). In other words, if the only consideration is whether or not to report X, Haberman's (2008) approach and the approach considered in this

Table 1: Summary Statistics for SAT Verbal (Sub)scores (from Haberman, 2008, p. 218)

X	n	$\overline{X}$	$\widehat{\sigma}(X)$	$\widehat{\sigma}(E_X)$	Z	$\widehat{\rho}(X,Z)$
CR	40	19.4	8.6	3.4	V	.96
A	19	9.3	4.1	2.1	V	.87
SC	19	10.6	4.4	2.1	V	.90
V	78	32.7	16.3	4.4	V+M	.93
M	60	26.9	14.1	3.6	V+M	.91

Note. CR = critical reading; A = analogies; SC = sentence completion; V = Verbal; M = Math. CR, A, and SC are subscores. Math section statistics are included only because $Z = V + M$ for the Verbal section.

Table 2: Utility and Relative Utility of using Z Instead of X for SAT Verbal (Sub)scores

X	Z	$\widehat{\rho}^2_X = \widehat{\rho}^2(T_X, X)$	$\text{Est}(U) = \widehat{\rho}^2(T_X, Z)$	$\text{Est}(\widetilde{U})$	$100(\widetilde{U}-1)\%$
CR	V	.84	.89	1.54	54%
A	V	.74	.87	2.35	135%
SC	V	.78	.88	2.07	107%
V	V+M	.91	.85	.56	

Note. CR = critical reading; A = analogies; SC = sentence completion; V = Verbal; M = Math. CR, A, and SC are subscores. The results in columns three and four are the same as those in columns three and four of Table 5 in Haberman (2008, p. 219).

paper lead to the same decision. However, Haberman's justification for his decision rule rests on logic surrounding RSEs, whereas the logic proposed here makes no use of RSEs.

4 Haberman's SAT Example

Haberman (2008) provides an example of his approach to examining the question of when subscores have value. His example involves subscores for both the Verbal (V) and Math (M) sections of the "old" SAT — i.e., the SAT used prior to March, 2005. Here we focus on the Verbal score and its subscores [critical reading (CR), analogies (A), and sentence completion (SC)]. Table 1 provides summary statistics from Haberman (2008, Table 2, p. 218).

Table 2 provides estimates of reliability (ρ^2_X), utility (U), relative utility ($\widetilde{U}$), and $100(\widetilde{U}-1)\%$. In Table 2 the values of $\widehat{\rho}^2_X$ and $\text{Est}(U)$ are the values in columns three and four, respectively, of Table 5 in Haberman (2008, p. 219).

Presumably due to rounding error, the Haberman values (columns three and

four of his Table 5 and the present paper's Table 2) cannot always be obtained exactly from the Haberman values in Table 1. For example, for SC,

$$\widehat{\rho}_X^2 = 1 - \left[\frac{\widehat{\sigma}(E_X)}{\widehat{\sigma}(X)}\right]^2 = 1 - \left[\frac{2.1}{4.4}\right]^2 = .77,$$

rather than .78, as reported by Haberman (2008, Table 5, third column). Also, for CR, using Equation 10,

$$\text{Est}(U) = \frac{[(8.6)(16.3)(.96) - (3.4)^2]^2}{(.84)(8.6)^2(16.3)^2} = .91,$$

rather than .89, as reported by Haberman (2008, Table 5, fourth column).

The results in the last two columns of Table 2 clearly indicate that, with respect to reliability issues, the total observed Verbal score is a better choice than X for any of the observed subtest scores. By contrast, for the Verbal score, the total $V + M$ score is not nearly as good as the observed Verbal score, with respect to reliability issues. These conclusions are consistent with those suggested by Haberman (2008), but the psychometric rationale here differs substantially from that in Haberman (2008). Furthermore, when $\widetilde{U} > 1$ it seems advantageous to be able to quantify results in terms of $100(\widetilde{U}-1)\%$, namely, the percentage increase in test length for X needed to obtain a reliability consistent with the X-type information in Z.

5 Summary and Concluding Comments

If test scores fit a unidimensional model, a psychometically compelling argument cannot be mounted for reporting any subscores since, by definition, there is only one proficiency or latent trait. The results in this paper do not circumvent this problem in any way. So, the results reported in this paper are meaningful and useful only when unidimensionality does *not* hold.

This paper considers whether or not to report X based on reliability considerations, primarily. The principal argument is that reporting X is justified, based on reliability considerations, if $\widetilde{U} \leq 1$. By contrast, if $\widetilde{U} > 1$, then reporting Z or some transformation of it *may* be justified.

5.1 Caveats on using Z as a Proxy for X

There are caveats on using Z as a proxy for X, however.

1. If X is not judged to be reliable enough for reporting purposes, almost always the best solution is to increase the length of X, although this may not be a practical alternative.

2. If X has unacceptably low reliability, it is sometimes argued that reporting scores for X can be justified provided interpretations are restricted to only those items that are actually in X with all other conditions of measurement held constant. This is a severe constraint on interpretations, however.

3. It is possible for X to have acceptably high reliability even when there is evidence that Z is a better proxy for T_X.

4. Typically there is information in Z that is only partially related to T_X. This fact may be judged sufficient to rule out using Z (or any transformation of it) as a proxy for T_X.

5. Using Z (or a transformation of it) as a proxy for T_X should be done with caution, in conjunction with appropriate documentation and warnings.

One always important caution involves paying attention to validation issues (see especially Kane, in press, and Brennan, in press), namely, evidence arguments based on claims made about interpretations and uses of reported subscores.

5.2 Metrics

Some of the discussions in this paper have been presented, explicitly or implicitly, in terms of the number-correct metric or number-of-points metric, which means that X and Z are sums of integer scores. This was done for purposes of simplicity and consistency with most of the current literature on the topic of subscores. Also, much of the literature on classical test theory uses the number-correct metric (or, occasionally, the number-of-points metric). Many of the equations, however, are not metric specific, and certainly the logic applies to any metric.

5.3 The Variable Y

The results in this paper have been presented using Y as an undifferentiated variable. More generally, however, these results apply when Y is a composite variable,

$$Y = Y_1 + Y_2 + Y_3 \cdots + Y_q, \tag{29}$$

provided $\sigma(T_X, E_{Y_i}) = 0$ and $\sigma(E_X, E_{Y_i}) = 0$ for all i.

In almost all the literature on the value of subscores, Z is the total score on a test. For the results in this paper, however, there is no restriction on Z other than that it must include X. So, for example, if Y is given by Equation 29, then $Z = X + Y_1 + Y_3 + \cdots + Y_q$ is permissible. Further, the Y_i need not correspond to reported subscores; any of the Y_i could be associated with a subset of items from a reported subscore. Although this is mathematically true, it may not have much practical value if Z is not a reported score.

5.4 Reliability Coefficients

It can be argued that reliability coefficients are overused in much of the literature on, and technical documentation for, testing. The usual argument is that statistics such as standard errors of measurement are more useful when making decisions about examinees. The author would generally agree, but not necessarily for the matters discussed in this paper. When comparing the measurement

characteristics of two observed scores such as X and Z, it is important to recognize that, for many metrics, any X-type information in Z affects true score variance and error variance differentially for a lengthened version of X (see for example, Section 2.4).

In the literature on subscores, most of the time it is assumed that coefficient alpha is used for reliability. For the results in this paper, however, any coefficient could be used. Ideally, the coefficient chosen should be one that includes all the potential sources of error of interest to an investigator.

Also, the coefficients for X and Z need not be of the same type. For example, suppose the X-type items are dichotomously-scored multiple-choice items, while the Y-type items are constructed response items with a scoring rubric of, say, 0–4. Then coefficient alpha might be used for X, while a multivariate generalizability coefficient might be used for Z (see Brennan, 2001).

Although the reliability coefficients for X and Z may be of different types, investigators need to be thoughtful in their choice of coefficients and the data to estimate them. For example, since X is part of Z it would be problematic to use coefficient alpha for X and a test-retest coefficient for Z, since the latter includes occasion as a source of error while the former generally does not. Similarly, almost certainly, the data used to compute reliability coefficients and $\sigma(X,Z)$ should be obtained at the same time (or nearly so), and these data should represent the same population.

6 Appendix

6.1 Linear Transformations and Equation 10

Although U is invariant with respect linear transformations of Z, Equation 10 is expressed assuming that $Z = X + Y$, and computations should be done using $Z = X + Y$. By contrast, if $l(Z) = a + b(Z)$ is used, a computational formula for U is:

$$U = \frac{[\sigma(X, l(Z)) - b\,\sigma^2(E_X)]^2}{\rho_X^2\,\sigma^2(X)\,\sigma^2(l(Z))}.$$

It may not be entirely evident why b multiplies the second term in the numerator. The reason is that the numerator is the square of $\sigma(T_X, l(Z)) = \sigma(X, l(Z)) - \sigma(E_X, l(Z))$, and

$$\begin{aligned}
\sigma(E_X, l(Z)) &= \sigma(E_X, a + b\,Z) \\
&= \sigma[E_X, a + b\,(X + Y)] \\
&= \sigma\{E_X, a + b\,[(T_X + E_X) + (T_Y + E_Y)]\} \\
&= \sigma(E_X, b\,E_X) \\
&= b\,\sigma^2(E_X).
\end{aligned}$$

6.2 Proofs of Equations in Section 3.3

Assume that $\rho_X^2 = \rho_Y^2$ and $\sigma(X) = \sigma(Y)$; i.e., X and Y are equally reliability and transformed to have equal standard deviations. Also, to simplify the notation, let ρ stand for $\rho(T_X, T_Y)$. Using Equation 9, the utility index can be expressed as

$$\begin{aligned} U &= \left[\left(\frac{\rho_X\,\sigma(X)}{\sigma(Z)}\right) + \rho\,\left(\frac{\rho_X\,\sigma(X)}{\sigma(Z)}\right)\right]^2 \\ &= \left(\frac{\sigma^2(X)\,\rho_X^2}{\sigma^2(Z)}\right)\,(1+\rho)^2, \end{aligned} \tag{30}$$

where

$$\begin{aligned} \sigma^2(Z) &= 2\,\sigma^2(X) + 2\,\sigma(X,Y) \\ &= 2\,\sigma^2(X) + 2\,\sigma^2(X)\,\rho(X,Y) \\ &= 2\,\sigma^2(X) + 2\,\sigma^2(X)\,\rho_X^2\,\rho \\ &= 2\,\sigma^2(X)(1+\rho_X^2\,\rho). \end{aligned} \tag{31}$$

It follows from Equations 30 and 31 that

$$U = \frac{\rho_X^2\,(1+\rho)^2}{2\,(1+\rho_X^2\,\rho)}, \tag{32}$$

and

$$\begin{aligned} 1-U &= \frac{2\,(1+\rho_X^2\,\rho) - \rho_X^2\,(1+\rho)^2}{2\,(1+\rho_X^2\,\rho)} \\ &= \frac{2\,+2\,\rho_X^2\,\rho - \rho_X^2\,-2\,\rho_X^2\,\rho - \rho_X^2\,\rho^2}{2\,(1+\rho_X^2\,\rho)} \\ &= \frac{2\,-\rho_X^2\,-\rho_X^2\,\rho^2}{2\,(1+\rho_X^2\,\rho)} \\ &= \frac{2\,-\rho_X^2(1+\rho^2)}{2\,(1+\rho_X^2\,\rho)}. \end{aligned} \tag{33}$$

Therefore, from Equations 32 and 33

$$\frac{U}{1-U} = \frac{\rho_X^2\,(1+\rho)^2}{2\,-\rho_X^2(1+\rho^2)},$$

and using Equation 20,

$$\begin{aligned} \widetilde{U} &= \left(\frac{U}{1-U}\right) \div \left(\frac{\rho_X^2}{1-\rho_X^2}\right) \\ &= \left(\frac{U}{1-U}\right)\left(\frac{1-\rho_X^2}{\rho_X^2}\right) \\ &= \frac{(1-\rho_X^2)(1+\rho)^2}{2\,-\rho_X^2(1+\rho^2)}. \end{aligned} \tag{34}$$

Using X is defensible, given the perspective adopted in this paper, when $\widetilde{U} \leq 1$; i.e., when

$$\begin{aligned}
(1-\rho_X^2)(1+\rho)^2 &\leq 2-\rho_X^2(1+\rho^2) \\
(1-\rho_X^2)(1+2\,\rho+\rho^2) &\leq 2-\rho_X^2(1+\rho^2) \\
1+2\,\rho+\rho^2-\rho_X^2-2\,\rho_X^2\rho-\rho_X^2\rho^2 &\leq 2-\rho_X^2-\rho_X^2\rho^2 \\
1+2\,\rho+\rho^2-2\,\rho_X^2\rho &\leq 2 \\
1+2\,\rho(1-\rho_X^2)+\rho^2 &\leq 2 \\
2\,\rho(1-\rho_X^2) &\leq 1-\rho^2 \\
1-\rho_X^2 &\leq \left(\frac{1-\rho^2}{2\,\rho}\right) \\
\rho_X^2 &\geq 1-\left(\frac{1-\rho^2}{2\,\rho}\right). \qquad (35)
\end{aligned}$$

Alternatively, using Equation 35, letting $v = 1-\rho_X^2$, and solving for ρ in terms of ρ_X^2 gives

$$\begin{aligned}
\left(\frac{1-\rho^2}{2\,\rho}\right) &\geq v \\
1-\rho^2 &\geq 2\,v\,\rho \\
\rho^2+2\,v\,\rho-1 &\leq 0.
\end{aligned}$$

Using the quadratic formula gives

$$\begin{aligned}
\rho &\leq -v \pm \sqrt{v^2+1} \\
&\leq -(1-\rho_X^2) \pm \sqrt{(1-\rho_X^2)^2+1}
\end{aligned}$$

Since ρ_X^2 must be between 0 and 1, assuming ρ is nonnegative,[9] it follows that when $\widetilde{U} \leq 1$,

$$\rho \leq -(1-\rho_X^2) + \sqrt{(1-\rho_X^2)^2+1}.$$

which is equivalent to

$$\rho \leq (\rho_X^2-1) + \sqrt{(1-\rho_X^2)^2+1}. \qquad (36)$$

Note that as $\rho_X^2 \to 0$, the right side of Equation 36 approaches $(-1+\sqrt{2}) \doteq .414$. Roughly speaking, this means that, even for very small values of ρ_X^2, using X is preferable to using Z when ρ is quite small. Of course, in such extreme cases, neither X nor Z may be adequate estimators of T_X.

[9] Mathematically, ρ could be negative, but we disregard this possibility because it generally leads to an ambiguous, if not meaningless, measurement procedure.

6.3 Haberman's Statistics

The utility index, U, and reliability of X, ρ_X^2, are related to quantities discussed by Haberman (2008). His principal statistics are special cases of the proportional reduction in mean-squared error (PRMSE) based on using A to estimate true score rather than using $\boldsymbol{E}(X)$ to estimate true score. For consistency with Haberman (2008), here we focus on a specific true score, τ_X. Also, Haberman's S_X is simply designated X, here; similarly, S_Z is designated Z.

Under classical test theory assumptions,

$$\text{MSE}(\tau_X|\,\boldsymbol{E}(X)) = \text{MSE}(\tau_X|\,\boldsymbol{E}(\tau_X)) = \sigma^2(\tau_X).$$

It follows that PRMSE can be expressed as

$$\psi(\tau_X|A) \quad = \quad 1 - \frac{\text{MSE}(\tau_X|A))}{\sigma^2(\tau_X)}. \tag{37}$$

As Kelley (1947) showed decades ago, for the linear regression (L) of τ_X on X

$$\text{MSE}(\tau_X|\text{L}(\tau_X|X)) = \sigma^2(\tau_X)[1 - \rho^2(\tau_X, X)]. \tag{38}$$

It follows from Equation 37 that

$$\psi(\tau_X|\text{L}(\tau_X|X)) = \rho^2(\tau_X, X), \tag{39}$$

which is reliability, previously abbreviated ρ_X^2. Replacing X with Z in Equations 38 and 39 gives

$$\psi(\tau_X|\text{L}(\tau_X|Z)) = \rho^2(\tau_X, X), \tag{40}$$

which is the utility index U in this paper.

Given the above results, according to Sinharay (2010, p. 152), a subscore has added (or equivalent) value over the total score when $\psi(\tau_X|\text{L}(\tau_X|X)) \geq \psi(\tau_X|\text{L}(\tau_X|Z))$, which is equivalent to $\rho_X^2 \geq U$, which in turn is equivalent to $\widetilde{U} \leq 1$, as proven next:

$$\begin{aligned}
\widetilde{U} \leq 1 \quad &\Rightarrow \quad \frac{U/(1-U)}{\rho_X^2/(1-\rho_X^2)} \leq 1 \\
&\Rightarrow \quad U/(1-U) \leq \rho_X^2/(1-\rho_X^2) \\
&\Rightarrow \quad U - U\rho_X^2 \leq \rho_X^2 - U\rho_X^2 \\
&\Rightarrow \quad U \leq \rho_X^2 \\
&\Rightarrow \quad \rho_X^2 \geq U.
\end{aligned}$$

The derivation presented above is the author's, but it closely mirrors comments made by Sinharay (personal communication, September 27, 2012).

6.4 Proof of Equations in Section 3.4.1

Recall from Section 3.4.1 that $X' = a + b\,Z$, where $b = \sigma(X)/\sigma(Z)$ and $a = \overline{X} - b\,\overline{Z}$. It follows that:

$$\begin{aligned}
\boldsymbol{E}(X' - T_X) &= \boldsymbol{E}(a + b\,Z - T_X) \\
&= \boldsymbol{E}\left\{\overline{X} - \left[\frac{\sigma(X)}{\sigma(Z)}\right]\overline{Z} + \left[\frac{\sigma(X)}{\sigma(Z)}\right] Z - T_X\right\} \\
&= \boldsymbol{E}\left\{[\mu(T_X) - T_X] + \left[\frac{\sigma(X)}{\sigma(Z)}\right](Z - \overline{Z})\right\} \\
&= 0.
\end{aligned}$$

Since a is a constant for all persons, and the expectation of $(X' - T_X)$ is zero,

$$\begin{aligned}
\mathrm{MSE}(X' - T_X) &= \mathrm{var}\left[\frac{\sigma(X)}{\sigma(Z)}\,Z - T_X\right] \\
&= \mathrm{var}\left[\frac{\sigma(X)}{\sigma(Z)}\,Z\right] + \mathrm{var}(T_X) - 2\,\mathrm{cov}\left[\frac{\sigma(X)}{\sigma(Z)}\,Z, T_X\right] \\
&= \sigma^2(X) + \sigma^2(T_X) - 2\,\frac{\sigma(X)}{\sigma(Z)}\,\mathrm{cov}(Z, T_X) \\
&= 2\,\sigma^2(X) - \sigma^2(E_X) - 2\,\frac{\sigma(X)}{\sigma(Z)}\rho(T_X, Z)\,\sigma(T_X)\,\sigma(Z) \\
&= 2\,\sigma^2(X) - \sigma^2(E_X) - 2\,\sigma(X)\,\sigma(T_X)\,\rho(T_X, Z) \\
&= 2\,\sigma^2(X) - \sigma^2(E_X) - 2\,\sigma^2(X)\,\rho_X\,\rho(T_X, Z) \\
&= 2\,\sigma^2(X)[1 - \rho_X\,\rho(T_X, Z)] - \sigma^2(E_X),
\end{aligned}$$

and the square root is Equation 6.4.

7 References

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

Brennan, R. L. (2012, December). *Kelley's regressed score estimates and inconsistencies with classical test theory.* (CASMA Technical Note Report No. 5). 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. (in press). Commentary on "Validating the interpretations and uses of test scores." *Journal of Educational Measurement.*

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. (in press), Validating the interpretations and uses of test scores. *Journal of Educational Measurement.*

Kelley, T. L. (1947). *Fundamentals of statistics.* Cambridge, MA: Harvard University Press.

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

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47*, pp. 150–174.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B.B., Swygert, K. A., & Thissen, D. (2001). Augumented scores—"Borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 342–387). Mahwah, NJ: Lawrence Erlbaum.