Article

Journal of Educational and Behavioral Statistics
Vol. XX, No. X, pp. 1–20
DOI: 10.3102/1076998618788862

Added Value of Subscores and Hypothesis Testing

Sandip Sinharay
Educational Testing Service

The value-added method of Haberman is arguably one of the most popular methods to evaluate the quality of subscores. According to the method, a subscore has added value if the reliability of the subscore is larger than a quantity referred to as the proportional reduction in mean squared error of the total score. This article shows how well-known statistical tests can be used to determine the added value of subscores and augmented subscores. The usefulness of the suggested tests is demonstrated using two operational data sets.

Keywords: *multiple correlation; PRMSE; value-added method*

The value-added method of Haberman (2008) has attracted wide attention. According to the method, a subscore has added value if the estimate of the *proportional reduction in mean squared error* (PRMSE) of the subscore is larger than the estimate of the PRMSE of the total score. Because the PRMSEs estimated from a sample can be affected by sampling variability (e.g., Wedman & Lyren, 2015), Sinharay and Haberman (2014) and Feinberg and Jurich (2017) suggested the use of statistical hypothesis testing to determine the added value of subscores and the use of resampling methods (e.g., Efron & Tibshirani, 1993) to compute the variance of the difference (in the former paper) or the ratio (latter paper) of two estimated PRMSEs.

Feinberg and Jurich (2017) noted that obtaining a theoretical expression of the estimated variance of the difference or the ratio of two estimated PRMSEs was problematic because these two estimates are correlated and a theoretical expression of their covariance was unavailable. This article suggests an approach to overcome this problem by utilizing the approach of Sinharay (2013) of expressing the PRMSEs as squared correlations and then using well-known statistical tests for testing the significance of the difference between two correlated correlations (e.g., Olkin, 1967; Olkin & Finn, 1995; Williams, 1959). The outcome is a new hypothesis-testing approach to determine the added value of subscores. In addition, a hypothesis-testing approach is suggested for determining the added value of augmented subscores (Haberman, 2008).

A review of the method of Haberman (2008) is included in the next section. The Method and the Real Data sections provide the expressions of the relevant variance and the resulting hypothesis-testing approaches and an illustration using

TABLE 1.
Some Notations

Notation	Quantity Denoted by the Notation
S	Observed subscore of an examinee
X	Observed total score of the examinee
T_S	True subscore of the examinee
μ_S	Mean of S
μ_X	Mean of X
σ_S^2	Variance of S
σ_X^2	Variance of X
$\rho_{ST_S}^2$	Reliability coefficient of a subscore

Note. $\rho_{Y_1Y_2}$ denotes the correlation coefficient between the variables Y_1 and Y_2.

two real data sets, respectively. The last section includes the conclusions. The Appendix includes software codes for implementing the suggested approaches.

Literature Review

The Haberman Method of Evaluating the Added Value of Subscores

Table 1 includes some of the notation that is used throughout this article. Such notation has been used by, for example, Lord and Novick (1968) and Sinharay (2018).

The method of Haberman (2008) primarily focuses on the following two linear regressions:

- The linear regression of the true subscore (T_S) on the observed subscore (S):

$$\mathcal{L}(T_S|S) = \mu_S + \rho_{ST_S}^2(S - \mu_S). \quad (1)$$

- The linear regression of the true subscore on the observed total score (X):

$$\mathcal{L}(T_S|X) = \mu_S + \frac{\text{Cov}(X, T_S)}{\sigma_X^2}(X - \mu_X). \quad (2)$$

The quality of $\mathcal{L}(T_S|S)$ and $\mathcal{L}(T_S|X)$ is evaluated using their mean squared error (MSE) and their PRMSE in comparison to those of a constant predictor μ_S. The predictors μ_S, $\mathcal{L}(T_S|S)$, and $\mathcal{L}(T_S|X)$ have corresponding MSEs of

$$\text{MSE}_{\text{const}} = E(T_S - \mu_S)^2 = \text{Var}(T_S),$$

$$\text{MSE}_{\text{sub}} = E(T_S - \mathcal{L}(T_S|S))^2 = \text{Var}(T_S)\left(1 - \rho_{ST_S}^2\right),$$

$$\text{and MSE}_{\text{total}} = E(T_S - \mathcal{L}(T_S|X))^2 = \text{Var}(T_S)\left(1 - \rho_{XT_S}^2\right),$$

respectively, where Var(X) denotes the variance of X.

The PRMSE in predicting T_S by $\mathcal{L}(T_S|S)$ instead of by μ_S, denoted henceforth as PRMSE$(T_S|S)$, is computed as

$$\text{PRMSE}(T_S|S) = 1 - \frac{\text{MSE}_{\text{sub}}}{\text{MSE}_{\text{const}}} = 1 - \frac{\text{Var}(T_S)\left(1 - \rho^2_{ST_S}\right)}{\text{Var}(T_S)} = \rho^2_{ST_S}. \tag{3}$$

The PRMSE in predicting T_S by $\mathcal{L}(T_S|X)$ instead of by μ_S is computed as

$$\text{PRMSE}(T_S|X) = 1 - \frac{\text{MSE}_{\text{total}}}{\text{MSE}_{\text{const}}} = 1 - \frac{\text{Var}(T_S)\left(1 - \rho^2_{XT_S}\right)}{\text{Var}(T_S)} = \rho^2_{XT_S}. \tag{4}$$

A larger value of a PRMSE indicates an increased accuracy of the corresponding linear regression. Equation 3 shows that PRMSE$(T_S|S)$ is equal to the subscore reliability.

According to the method of Haberman (2008), a subscore has added value if

$$\text{PRMSE}(T_S|S) > \text{PRMSE}(T_S|X),$$

a requirement which, from Equations 3 and 4, is identical to the requirement that

$$\rho^2_{ST_S} > \rho^2_{XT_S}. \tag{5}$$

For real data sets, the PRMSEs are unknown and can be estimated using approaches outlined by Haberman (2008). The method of Haberman (2008) was applied to data from a wide variety of tests by Brennan (2012); Liu, Robin, Yoo, and Manna (2018); Lyren (2009); Meijer, Boev, Tendeiro, Bosker, and Albers (2017); Puhan, Sinharay, Haberman, and Larkin (2010); Reise, Bonifay, and Haviland (2013); Sinharay (2010); Sinharay and Haberman (2008); Sawaki and Sinharay (2017); and Wedman and Lyren (2015).

Interpretation of Haberman's Method in Terms of Parallel Forms

Let S' denote the subscore on a parallel test form (Sinharay, 2013). Then,

$$E(S) = E(S') = \mu_S, \text{Var}(S) = \text{Var}(S') = \sigma^2_S, \text{ and correlation}(S, S') = \rho_{SS'} = \rho^2_{ST_S}. \tag{6}$$

One can predict S' by the constant μ_S; the corresponding MSE is given by

$$E(S' - \mu_S)^2 = \text{Var}(S') = \sigma^2_S. \tag{7}$$

The linear regression of S' on S is given by

$$\mathcal{L}(S'|S) = E(S') + \frac{\rho_{SS'}\sqrt{\text{Var}(S')}}{\sqrt{\text{Var}(S)}}(S - E(S)) = \mu_S + \rho^2_{ST_S}(S - \mu_S), \tag{8}$$

and has a corresponding MSE of

$$E(S' - \mathcal{L}(S'|S))^2 = E(S' - \mu_S - \rho^2_{ST_S}(S - \mu_S))^2 = \sigma^2_S\left(1 - \rho^4_{ST_S}\right). \tag{9}$$

Equations 6, 7, and 9 indicate that the PRMSE in predicting S' by $\mathcal{L}(S'|S)$ instead of μ_S is given by

$$\text{PRMSE}(S'|S) = \frac{\sigma_S^2 - \sigma_S^2\left(1 - \rho_{ST_S}^4\right)}{\sigma_S^2} = \rho_{ST_S}^4 = [\text{PRMSE}(T_S|S)]^2 = \rho_{SS'}^2. \quad (10)$$

The linear regression of S' on X is given by

$$\mathcal{L}(S'|X) = \mu_S + \rho_{XS'}\frac{\sigma_S}{\sigma_X}(X - \mu_X), \quad (11)$$

and has a corresponding MSE of

$$E(S' - \mathcal{L}(S'|X))^2 = \sigma_S^2\left(1 - \rho_{XS'}^2\right). \quad (12)$$

From Equations 7 and 12, the PRMSE in predicting S' by $\mathcal{L}(S'|X)$ instead of μ_S is

$$\text{PRMSE}(S'|X) = \frac{\sigma_S^2 - \sigma_S^2\left(1 - \rho_{XS'}^2\right)}{\sigma_S^2} = \rho_{XS'}^2. \quad (13)$$

Sinharay (2013) also showed that

$$\rho_{XS'}^2 = \rho_{XT_S}^2\rho_{ST_S}^2 = \text{PRMSE}(T_S|X)\text{PRMSE}(T_S|S). \quad (14)$$

A subscore has added value if $\text{PRMSE}(S'|S)$ is larger than $\text{PRMSE}(S'|X)$, which, because of Equations 10, 13, and 14, is identical to the requirement that

$$\rho_{ST_S}^4 > \rho_{XT_S}^2\rho_{ST_S}^2, \quad (15)$$

or to the requirement that

$$\rho_{ST_S}^2 > \rho_{XT_S}^2,$$

which, because of Equation 5, is identical to the requirement by Haberman (2008) for a subscore to have added value.

Existing Hypothesis-Testing Approaches to Determine the Added Value of Subscores

The original method of Haberman (2008) does not involve any hypothesis testing. Let us denote, for example, the estimated value of $\text{PRMSE}(T_S|S)$ as $\widehat{\text{PRMSE}(T_S|S)}$. According to the Haberman method, a larger value of $\widehat{\text{PRMSE}(T_S|S)}$ compared to $\widehat{\text{PRMSE}(T_S|X)}$, irrespective of the magnitude of the difference between these two quantities, indicates an added value of the subscore.

The PRMSEs estimated from a sample can be affected by sampling variability (e.g., Wedman & Lyren, 2015), which may lead to incorrect conclusions on added value of subscores, especially for small samples. Therefore, Sinharay and

Haberman (2014) suggested considering that a subscore has added value only if $\widehat{\text{PRMSE}}(T_S|S) - \widehat{\text{PRMSE}}(T_S|X)$ divided by its estimated standard deviation (*SD*) is larger than a critical value (of, e.g., 1.64 at 5% significance level). Feinberg and Jurich (2017) suggested considering a subscore to have added value only if

$$\widehat{\text{VAR}} = \widehat{\text{PRMSE}}(T_S|S)/\widehat{\text{PRMSE}}(T_S|X)$$

divided by its estimated *SD* is larger than a critical value. One common element of these two hypothesis-testing approaches is the use of a resampling method (e.g., Efron & Tibshirani, 1993) to compute the relevant variances. While Sinharay and Haberman (2014) used the jackknife procedure (Efron, 1979) to estimate the variance of $\widehat{\text{PRMSE}}(T_S|S) - \widehat{\text{PRMSE}}(T_S|X)$, Feinberg and Jurich (2017) used the bootstrap procedure (e.g., Efron & Tibshirani, 1993) to estimate the variance of $\widehat{\text{VAR}}$.

Feinberg and Jurich (2017) stated that it is difficult to obtain a theoretical expression for the estimated variance of the difference or the ratio of $\widehat{\text{PRMSE}}(T_S|S)$ and $\widehat{\text{PRMSE}}(T_S|X)$ because these two estimates are correlated and there is no known theoretical expression of their covariance or correlation; therefore, for example, the variance of $\widehat{\text{PRMSE}}(T_S|S) - \widehat{\text{PRMSE}}(T_S|X)$ cannot be derived as the sum of the variances of $\widehat{\text{PRMSE}}(T_S|S)$ and $\widehat{\text{PRMSE}}(T_S|X)$.[1]

Hypothesis testing has not been used to determine the added value of augmented subscores even though the advantages of augmented subscores have been acknowledged by several researchers such as Sinharay (2010) and Thissen (2013).

Method: New Hypothesis-Testing Approaches to Determine the Added Value of Subscores and Augmented Subscores

Added Value of Subscores

Given the earlier discussion that the requirement

$$\text{PRMSE}(T_S|S) \geq \text{PRMSE}(T_S|X)$$

for a subscore to have added value is identical to the requirement that

$$\text{PRMSE}(S'|S) \geq \text{PRMSE}(S'|X),$$

one can test the hypothesis

$$H_0 : \text{PRMSE}(T_S|S) - \text{PRMSE}(T_S|X) \text{ versus } H_1 : \text{PRMSE}(T_S|S) > \text{PRMSE}(T_S|X)$$

or the hypothesis

$$H_0 : \text{PRMSE}(S'|S) - \text{PRMSE}(S'|X) \text{ versus } H_1 : \text{PRMSE}(S'|S) > \text{PRMSE}(S'|X)$$

to determine whether a subscore has added value. The latter approach is undertaken in this article because of the parallel-forms interpretation of $\text{PRMSE}(S'|S)$ and $\text{PRMSE}(S'|X)$. Also, the question of interest in applications of the Haberman method almost always is "Do the subscores have added value?" This question corresponds to a one-sided alternative hypothesis (H_1).

Let r_{01}, r_{02}, and r_{12}, respectively, denote the sample correlations between variables V_0 and V_1, between V_0 and V_2, and between V_1 and V_2, respectively, where these variables are all measured on one sample of size N of examinees. Let ρ_{01}, ρ_{02}, and ρ_{12} denote the corresponding population correlations. Note that r_{01} and r_{02} have a variable (V_0) in common—so r_{01} and r_{02} are "correlated correlations" according to the terminology of, for example, Olkin and Finn (1990).

Olkin's Z statistic. Pearson and Filon (1898) provided the result that

$$\text{Var}(r_{01}) = \frac{1}{N}(1-\rho_{01}^2)^2, \tag{16}$$

and that an estimate of the covariance of r_{01} and r_{02} is given by

$$\text{Cov}(r_{01}, r_{02}) = \frac{1}{N}\left[\frac{1}{2}(2\rho_{12} - \rho_{01}\rho_{02})(1-\rho_{01}^2-\rho_{02}^2-\rho_{12}^2)+\rho_{12}^3\right]. \tag{17}$$

Equations 16 and 17 lead to the result that

$$\widehat{\text{Var}}(r_{01}-r_{02}) = \frac{1}{N}\left[\left(1-r_{01}^2\right)^2+\left(1-r_{02}^2\right)^2-2r_{12}^3-(2r_{12}-r_{01}r_{02})\left(1-r_{01}^2-r_{02}^2-r_{12}^2\right)\right] \tag{18}$$

after replacing the population correlations by the corresponding sample correlations. Then, one can test the hypothesis $H_0 : \rho_{01} = \rho_{02}$ using the Olkin's Z statistic (e.g., Olkin, 1967) that is defined as

$$Z = \frac{r_{01}-r_{02}}{\sqrt{\widehat{\text{Var}}(r_{01}-r_{02})}}$$

and follows the standard normal distribution under the null hypothesis (H_0) for large samples when $\mathbf{V} = (V_0, V_1, V_2)'$ follows the trivariate normal distribution (e.g., Olkin, 1967; Olkin & Siotani, 1976; Steiger & Hakstian, 1982).

The application of the Olkin's Z statistic to determine the added value of subscores is facilitated by the interpretation of the square roots of the estimated $\text{PRMSE}(S'|S)$ and $\text{PRMSE}(S'|X)$ as sample correlations both of which involve S'. Equation 10 shows that $\sqrt{\widehat{\text{PRMSE}}(S'|S)}$ can be interpreted as the sample correlation between S' and S and Equation 13 shows that $\sqrt{\widehat{\text{PRMSE}}(S'|X)}$ can be interpreted as the sample correlation between S' and X. Therefore, one can compute the estimated variance of $\sqrt{\widehat{\text{PRMSE}}(S'|S)}-\sqrt{\widehat{\text{PRMSE}}(S'|X)}$ using

Equation 18 by setting $r_{01} = \sqrt{\widehat{\text{PRMSE}(S'|S)}}$, $r_{02} = \sqrt{\widehat{\text{PRMSE}(S'|X)}}$, and r_{12} = the sample correlation between S and X.

Then, to determine whether a subscore has added value, one can perform a hypothesis test (without using any resampling method) by examining whether the Olkin's Z statistic given by

$$Z = \frac{\sqrt{\widehat{\text{PRMSE}(S'|S)}} - \sqrt{\widehat{\text{PRMSE}(S'|X)}}}{\sqrt{\widehat{\text{Var}}\left(\sqrt{\widehat{\text{PRMSE}(S'|S)}} - \sqrt{\widehat{\text{PRMSE}(S'|X)}}\right)}}, \tag{19}$$

is larger than 1.64 or not, or, equivalently, by examining whether the (one sided) 95% confidence interval $\{l, u\}$ includes 0 or not, where

$$l = \sqrt{\widehat{\text{PRMSE}(S'|S)}} - \sqrt{\widehat{\text{PRMSE}(S'|X)}} - 1.64\sqrt{\widehat{\text{Var}}\left(\sqrt{\widehat{\text{PRMSE}(S'|S)}} - \sqrt{\widehat{\text{PRMSE}(S'|X)}}\right)}$$

and $u = 1$; the estimated variance in the denominator of Equation 19 is computed using Equation 18.

The value of the Olkin's Z statistic can be computed from given values of N, ρ_{01}, ρ_{02}, and ρ_{12} using the `R` package `cocor` (Diedenhofen & Musch, 2015).

Williams's t *statistic.* Another popular statistic to test for the significance of the difference between correlated correlations is the Williams's t statistic (Williams, 1959), which, in the context of testing the H_0, $H_0 : \rho_{01} = \rho_{02}$, is defined as

$$t = \frac{\sqrt{2}(r_{01} - r_{02})}{\left[\frac{(N-1)(1+r_{12})}{\frac{N-1}{N-3}|R| - \left(\frac{r_{01}+r_{02}}{2}\right)^2 (1-r_{12})^3}\right]^{1/2}}, \tag{20}$$

where $|R|$ is the determinant of the sample correlation matrix between V_0, V_1, and V_2. The statistic follows the t distribution with $N - 3$ degrees of freedom for large samples under the trivariate normality of $\mathbf{V}$ under the H_0. The statistic performed well compared to other statistics for testing correlation correlations in the comparison study of Hittner, May, and Silver (2003). The Williams's t statistic for testing the added value of subscores can be obtained, as with Olkin's Z, by setting $r_{01} = \sqrt{\widehat{\text{PRMSE}(S'|S)}}$, $r_{02} = \sqrt{\widehat{\text{PRMSE}(S'|X)}}$, and r_{12} = the sample covariance between S and X. The value of the Williams's t statistic can also be obtained from the `R` package `cocor`.

Added Value of Augmented Subscores

Haberman (2008) suggested the augmented subscore, which is the linear regression of T_S on S and X and is given by

$$\mathcal{L}(T_S|S,X) = \mu_S + (\mathrm{Cov}(S,T_S), \mathrm{Cov}(X,T_S))\begin{pmatrix} \mathrm{Var}(S) & \mathrm{Cov}(S,X) \\ \mathrm{Cov}(S,X) & \mathrm{Var}(X) \end{pmatrix}^{-1} \begin{pmatrix} S-\mu_S \\ X-\mu_X \end{pmatrix}. \tag{21}$$

Haberman (2008) provided expressions of $\mathrm{PRMSE}(T_S|S,X)$, which is the PRMSE associated with $\mathcal{L}(T_S|S,X)$, and its estimate. According to Haberman (2008), an augmented subscore has added value when $\mathrm{PRMSE}(T_S|S,X)$ is considerably larger than both $\mathrm{PRMSE}(T_S|S)$ and $\mathrm{PRMSE}(T_S|X)$.

Sinharay (2018) suggested the augmented subscore for predicting S' from S and X or the linear regression of S' on S and X that would henceforth be denoted as $\mathcal{L}(S'|S,X)$. The regression is given by

$$\mathcal{L}(S'|S,X) = \mu_S + (\mathrm{Cov}(S,S'), \mathrm{Cov}(X,S'))\begin{pmatrix} \mathrm{Var}(S) & \mathrm{Cov}(S,X) \\ \mathrm{Cov}(S,X) & \mathrm{Var}(X) \end{pmatrix}^{-1} \begin{pmatrix} S-\mu_S \\ X-\mu_X \end{pmatrix}. \tag{22}$$

Sinharay (2018) also proved that $\mathcal{L}(S'|S,X)$ is identical to $\mathcal{L}(T_S|S,X)$ and that $\mathrm{PRMSE}(S'|S,X)$, the PRMSE associated with $\mathcal{L}(S'|S,X)$, is given by

$$\mathrm{PRMSE}(S'|S,X) = \mathrm{PRMSE}(T_S|S,X)\mathrm{PRMSE}(T_S|S). \tag{23}$$

The result in Equation 23 is similar to those provided by Equations 10 and 14. An augmented subscore has added value when $\mathrm{PRMSE}(S'|S,X)$ is considerably larger than both $\mathrm{PRMSE}(S'|S)$ and $\mathrm{PRMSE}(S'|X)$, which, from a hypothesis-testing point of view, is equivalent to the condition that both $\mathrm{PRMSE}(S'|S,X) - \mathrm{PRMSE}(S'|S)$ and $\mathrm{PRMSE}(S'|S,X) - \mathrm{PRMSE}(S'|X)$ are significantly larger than 0. Thus, to determine the added value of augmented subscores, one can test if $\mathrm{PRMSE}(S'|S,X) - \mathrm{PRMSE}(S'|S)$ is significantly larger than 0 for a subscore for which $\mathrm{PRMSE}(S'|S) > \mathrm{PRMSE}(S'|X)$ and test if $\mathrm{PRMSE}(S'|S,X) - \mathrm{PRMSE}(S'|X)$ is significantly larger than 0 for a subscore for which $\mathrm{PRMSE}(S'|S) < \mathrm{PRMSE}(S'|X)$.

Sinharay (2018) showed that $\mathrm{PRMSE}(S'|S,X)$ is the squared multiple correlation coefficient corresponding to the linear regression of S' on S and X. Equations 10 and 13 show that $\mathrm{PRMSE}(S'|S)$ and $\mathrm{PRMSE}(S'|X)$ are squared correlations, the former between S' and S and the latter between S' and X.

Let $\rho^2_{0(12)}$ denote the population value of the squared multiple correlation coefficient corresponding to the linear regression of V_0 from V_1 and V_2, and let $r^2_{0(12)}$ denote the corresponding sample squared multiple correlation coefficient. To test the H_0, $H_0 : \rho^2_{0(12)} = \rho^2_{01}$, one can use the test statistic

$$\frac{r^2_{0(12)} - r^2_{01}}{\sqrt{\widehat{\mathrm{Var}}(r^2_{0(12)} - r^2_{01})}}, \tag{24}$$

which follows a standard normal distribution for large samples under the H_0 under the trivariate normality of $\mathbf{V}$ (e.g., Hedges & Olkin, 1983; Olkin & Finn, 1995; Steiger & Hakstian, 1982). This statistic would be referred to as Hedges–Olkin's Z statistic. The variance estimate $\widehat{\text{Var}}(r^2_{0(12)} - r^2_{01})$ is given by (e.g., Olkin & Finn, 1995)

$$\widehat{\text{Var}}(r^2_{0(12)} - r^2_{01}) = \mathbf{v}'\boldsymbol{\Omega}\mathbf{v},$$

where $\mathbf{v} = (v_1, v_2, v_3)'$,

$$v_1 = \frac{2r_{12}}{1 - r^2_{12}}(r_{01}r_{12} - r_{02}); v_2 = \frac{2}{1 - r^2_{12}}(r_{02} - r_{01}r_{12});$$
$$v_3 = \frac{2}{(1 - r^2_{12})^2}(r_{12}r^2_{01} + r_{12}r^2_{02} - r_{01}r_{02} - r_{01}r_{02}r^2_{12}),$$

and

$$\boldsymbol{\Omega} = \begin{pmatrix} \widehat{\text{Var}}(r_{01}) & \widehat{\text{Cov}}(r_{01}, r_{02}) & \widehat{\text{Cov}}(r_{01}, r_{12}) \\ & \widehat{\text{Var}}(r_{02}) & \widehat{\text{Cov}}(r_{02}, r_{12}) \\ & & \widehat{\text{Var}}(r_{12}) \end{pmatrix}.$$

The individual terms in $\boldsymbol{\Omega}$ can be computed using Equations 16 and 17.

Thus, to test whether $\text{PRMSE}(S'|S,X) - \text{PRMSE}(S'|S)$ is significantly larger than 0, given the interpretation of $\text{PRMSE}(S'|S,X)$ and $\text{PRMSE}(S'|S)$ as squared multiple/simple correlation coefficients, one can use the test statistic provided in Equation 24 by letting S', S, and X play the roles of V_0, V_1, and V_2, respectively, or, equivalently, by letting $\widehat{\text{PRMSE}}(S'|S,X)$, $\sqrt{\widehat{\text{PRMSE}}(S'|S)}$, $\sqrt{\widehat{\text{PRMSE}}(S'|X)}$, and the sample covariance between S and X play the roles of $r^2_{0(12)}$, r_{01}, r_{02}, and r_{12}, respectively. To test whether $\text{PRMSE}(S'|S,X) - \text{PRMSE}(S'|X)$ is significantly larger than 0, one can use the test statistic provided in Equation 24 by letting S', X, and S play the roles of V_0, V_1, and V_2, respectively.

The Appendix includes R codes for computing the values of the Olkin's Z statistic, the Williams's t statistic, and the Hedges–Olkin's Z statistic for a data set. The codes make use of the R package "subscore" (Dai, Wang, & Svetina, 2016). The data set used in the codes can be obtained upon request from the authors of the package.

Real Data Example

Let us consider a data set from the TerraNova test that is a series of standardized achievement tests designed for K–12 students. Yao (2010) analyzed this data set that includes the item-level scores of 3,953 examinees on five subject areas—language (LG), mathematics (MT), reading (RD), dcience (SC), and

docial dtudies (SS). The number of items (all multiple choice) in the five subject areas are 34, 57, 46, 40, and 40, respectively. Thus, the test includes 217 items. Each subject area is assumed to contribute to one subscore in the rest of this analysis.

Hypothesis Testing

The data from the test were used to compute the estimated PRMSE$(T_S|S)$, PRMSE$(T_S|X)$, PRMSE$(S'|S)$, PRMSE$(S'|X)$, and PRMSE$(S'|S,X)$. The Cronbach's α was used as the estimate of the reliabilities of the subscores and the stratified α was used as the estimate of the reliability of the total score. Finally, the estimated *SD* of $\sqrt{\widehat{\text{PRMSE}}(S'|S)} - \sqrt{\widehat{\text{PRMSE}}(S'|X)}$ was computed using Equation 18, the Olkin's Z statistic was computed using Equation 19, the Williams's t statistic was computed using Equation 20, and the Hedges–Olkin's Z statistic was computed using Equation 24. The estimated *SD* of $\sqrt{\widehat{\text{PRMSE}}(S'|S)} - \sqrt{\widehat{\text{PRMSE}}(S'|X)}$ was also computed using the jackknife procedures (e.g., Efron, 1979). [2]

For each subscore, Table 2 includes the number of items on the subtests, the sample mean and sample correlations of the subscores, the estimates of the PRMSEs,[3] the estimated *SD* of $\sqrt{\widehat{\text{PRMSE}}(S'|S)} - \sqrt{\widehat{\text{PRMSE}}(S'|X)}$ using the theoretical expression provided in Equation 18, the estimated *SD* of $\sqrt{\widehat{\text{PRMSE}}(S'|S)} - \sqrt{\widehat{\text{PRMSE}}(S'|X)}$ using the jackknife procedure, and the Olkin's Z, the Williams's t, and the Hedges–Olkin's Z statistics. The next row of the table indicates whether each subscore has added value; a "yes" for a subscore indicates that the Olkin's Z for the subscore is larger than 1.64 and a "no" indicates an Olkin's Z smaller than 1.64. The following row provides information on whether each augmented subscore has added value, that is, whether the Hedges–Olkin's Z is larger than 1.64 or not. The last three rows of the table will be discussed later.

Table 2 shows that if one intends to perform hypothesis testing to determine the added value of the subscores, then the values of Olkin's Z and Williams's t indicate that the mathematics and reading subscores have added value while the other three subscores do not. Note that $\widehat{\text{PRMSE}}(S'|S)$ is slightly larger than $\widehat{\text{PRMSE}}(S'|X)$ for the language and social studies subscores so that one not performing hypothesis testing and just comparing $\widehat{\text{PRMSE}}(S'|S)$ to $\widehat{\text{PRMSE}}(S'|X)$ would consider these two subscores to be of added value, but the Z and t values for these two subscores are not large enough for these subscores to be considered to have added value from a hypothesis-testing point of view. For each subscore, the values of the Olkin's Z and the Williams's t statistics are quite

TABLE 2.
Results for the TerraNova Data

Subscore	LG	MT	RD	SC	SS
Number of items	34	57	46	40	40
Sample mean	22.8	40.2	31.2	26.7	27.2
Correlation coefficients	1.00	0.76	0.88	0.72	0.77
		1.00	0.75	0.74	0.77
			1.00	0.75	0.80
				1.00	0.81
Estimated PRMSE($T_S\|S$)	0.890	0.917	0.922	0.830	0.894
Estimated PRMSE($T_S\|X$)	0.889	0.840	0.892	0.857	0.890
Estimated PRMSE($S'\|S$)	0.792	0.840	0.850	0.689	0.799
Estimated PRMSE($S'\|X$)	0.791	0.770	0.822	0.711	0.796
Estimated PRMSE($S'\|S, X$)	0.829	0.855	0.869	0.749	0.833
Estimated *SD*: Theoretical	0.0028	0.0029	0.0023	0.0041	0.0027
Estimated *SD*: Jackknife	0.0020	0.0024	0.0018	0.0031	0.0026
Olkin's Z	0.18	13.50	6.82	−3.33	0.75
Williams's t	0.18	14.36	6.93	−3.34	0.75
Hedges–Olkin's Z	13.66	9.53	11.46	11.73	13.36
Does the subscore have added value?	No	Yes	Yes	No	No
Does the augmented subscores have added value?	Yes	Yes	Yes	Yes	Yes
% Added value for subsamples: Z	3	100	85	0	6
% Added value for subsamples: t	3	100	85	0	6
% Added value for subsamples: Jackknife	3	100	91	0	7

Note. LG = language; MT = mathematics; RD = reading; SC = science; SS = social studies; *SD* = standard deviation.

close. Further, the estimates of the *SD* of $\sqrt{\widehat{\text{PRMSE}}(S'|S)} - \sqrt{\widehat{\text{PRMSE}}(S'|X)}$ from the jackknife procedure were reasonably close to those obtained using Equation 18, which indicates that the corresponding theoretical expression was accurate for this data set. Finally, the values of the Hedges–Olkin's Z statistic indicate that all the augmented subscores have added value.

Examining the Accuracy of the Suggested Hypothesis-Testing Approach

The Olkin's Z, the Williams's t, and the Hedges–Olkin's Z statistics follow their theorized distributions for large samples and trivariate normal variables (e.g., Hedges & Olkin, 1983; Olkin & Finn, 1995; Steiger, 1980). However, in practice, the theorized distributions may not hold for a given sample, especially if the sample size is small or moderately large or if the score distributions are far

from the normal distribution. For example, Hittner et al. (2003) found these statistics to have inflated Type I error rates when values of three variables V_0, V_1, and V_2 for up to 300 individuals were simulated from the exponential distribution.[4] Therefore, a simulation based on real data was performed to evaluate whether the Olkin's Z and the Williams's t statistics perform satisfactorily in determining the added value of subscores.

For the TerraNova data set described above, the following steps were repeated 1,000 times:

1. Draw a subsample of size 500 examinees from the original sample.[5]
2. Compute the values of the Olkin's Z and the Williams's t statistic from the subsample; use the subsample to compute a Z statistic like that given in Equation 19 with the difference that the variance involved in the Z statistic is obtained not using Equation 18 but is obtained using the jackknife procedure (Efron, 1979).[6]
3. Examine whether the Olkin's Z and the Williams's t statistic and the jackknife procedure indicate that the subscores have added value for the subsample.

The last three rows of Table 2 provide the percentage of times when the Olkin's Z, the Williams's t statistic, and the jackknife procedure, when applied on subsamples of size 500, indicate that the corresponding subscore has added value. For example, the Value 3 for the jackknife for the Language subscore in the last row implies that the jackknife procedure indicates that the Language subscore has added value for 3% of the 1,000 subsamples (i.e., for 30 subsamples). These percentages indicate that all the three procedures lead to accurate inferences regarding added value of subscores. Only the MT and RD subscores in the full sample have added value and the three procedures indicate added value for these two subscores in at least 85% of the subsamples. For the three subscores that do not have added value for the full sample, the three procedures indicate added value in at most 7% of the subsamples. Thus, in some sense, the power of Olkin's Z and Williams's t is at least 85% and their Type I error rate does not exceed 7%. In addition, both the Olkin's Z and the Williams's t statistic agreed with the jackknife procedure on the added value of the subscores in 97% of the subsamples (combined over all the subscores)—this result points to the satisfactory performance of the Olkin's Z and the Williams's t statistics for these data.

Similar calculations were performed for a data set from a licensure test that involved four subscores and 6,641 examinees. Table 3 provides the results for the data set for which only the second subscore had added value in the full sample. Both the Olkin's Z and the Williams's t statistics agreed with the jackknife procedure on the added value of the subscores in 100% of the subsamples for each subscore. The table leads to conclusions similar to those from Table 2 and points to the satisfactory performance of the Olkin's Z and the Williams's t statistics.

It should be noted that in simulations like those described above, the results on added value for subsamples would agree with those for the original sample if

TABLE 3.
Results for the Licensure Data

Subscore	1	2	3	4
Number of items	30	30	30	30
Estimated PRMSE($T_S\|S$)	0.628	0.857	0.752	0.780
Estimated PRMSE($T_S\|X$)	0.725	0.754	0.757	0.816
Estimated PRMSE($S'\|S$)	0.394	0.735	0.565	0.609
Estimated PRMSE($S'\|X$)	0.455	0.646	0.569	0.637
Estimated PRMSE($S'\|S,X$)	0.487	0.752	0.621	0.669
Estimated standard deviation (*SD*): Theoretical	0.0062	0.0035	0.0045	0.0037
Estimated *SD*: Jackknife	0.0059	0.0026	0.0037	0.0030
Olkin's Z	−7.50	15.49	−0.63	−4.81
Williams's t	−7.56	16.36	−0.63	−4.84
Hedges–Olkin's Z	10.26	10.66	14.82	12.51
Does the subscore have added value?	No	Yes	No	No
Does the augmented subscores has added value?	Yes	Yes	Yes	Yes
% Added value for subsamples: Z	0	100	0	0
% Added value for subsamples: t	0	100	0	0
% Added value for subsamples: Jackknife	0	100	1	0

PRMSE($S'|S$) and PRMSE($S'|X$) are vastly different in the original sample; for example, if PRMSE($S'|S$) is much larger than PRMSE($S'|X$) for a subscore in the original sample, each hypothesis-testing approach would indicate that the subscore has added value both for the original sample and for all subsamples. However, in the two data sets considered here, PRMSE($S'|S$) and PRMSE($S'|X$) are quite close in the original sample for three subscores (subscores LG and SS in Table 2 and the third subscore in Table 3). The high level of agreement of results from the subsamples and samples for these three subscores point to the satisfactory performance of the Olkin's Z and the Williams's t statistics and the jackknife procedure.

In some limited simulations based on these real data, the Hedges–Olkin's Z statistic also was found to lead to satisfactory results regarding the test of the hypothesis that an augmented subscore has added value.

Conclusions

Haberman (2008) suggested a method to determine whether subscores have added value for a test. With the exception of Sinharay and Haberman (2014), the application of the method has not involved any hypothesis testing. This article shows how one can perform statistical hypothesis tests to determine whether a subscore or an augmented subscore has added value. Two real data illustrations demonstrate the utility of the suggested hypothesis-testing approaches and show

that the suggested tests perform quite satisfactorily. Overall, this article promises to increase the appeal of the method of Haberman (2008) that has been of increasing interest among the measurement practitioners.

The suggested tests promise to be most useful for small sample sizes (of, e.g., a few hundred examinees) for which even a difference of 0.15 between $\widehat{\text{PRMSE}}(S'|S)$ and $\widehat{\text{PRMSE}}(S'|X)$ may not be statistically significant.[7] For large sample sizes (of, e.g., a few thousand examinees), even a difference of 0.02 would most likely be statistically significant[8] and application of hypothesis-testing would not be required.

Among the tests suggested in this article, the test for determining the added value of augmented subscores promises to be the most useful to practitioners. The current approach for determining the added value of augmented subscores (e.g., Haberman, 2008; Sinharay, 2018) comprises a subjective judgment of whether $\widehat{\text{PRMSE}}(S'|S,X)$ is substantially larger than both of $\widehat{\text{PRMSE}}(S'|S)$ and $\widehat{\text{PRMSE}}(S'|X)$. The appplication of the Hedges–Olkin test suggested in this article provides a rigorous basis of judging whether $\widehat{\text{PRMSE}}(S'|S,X)$ is substantially larger than both $\widehat{\text{PRMSE}}(S'|S)$ and $\widehat{\text{PRMSE}}(S'|S,X)$.

In this article, expressions were suggested for hypothesis testing using $\text{PRMSE}(S'|S)$, $\text{PRMSE}(S'|X)$, and $\text{PRMSE}(S'|S,X)$ suggested by Sinharay (2013) and Sinharay (2018), but similar expressions can be obtained for $\text{PRMSE}(T_S|S)$, $\text{PRMSE}(T_S|X)$, and $\text{PRMSE}(T_S|S,X)$ suggested by Haberman (2008). In some limited simulations, hypothesis testing was performed using the PRMSEs suggested by Haberman (2008), but the results were very similar to those reported in this article—so results using the PRMSEs suggested by Haberman (2008) are not reported here. In addition, although a test typically involves several subscores, no adjustment for multiple comparisons was applied in this article; in practice, the investigator may adjust for multiple comparisons by controlling the family-wise error rate (using, e.g., a Bonferroni correction) or controlling the false discovery rate (using the procedure of Benjamini and Hochberg, 1995).

There exist several tests of correlated correlations in addition to Olkin's Z and Williams's t. See, for example, reviews of them in the study by Hittner et al. (2003) and Steiger (1980). Dunn and Clark (1969) suggested two such statistics, denoted as Z_1^* and Z_2^*, that have been found to have performed well in simulation studies. Hittner et al. (2003) found Williams's t and one of the statistics of Dunn and Clark (1969) to perform well under a variety of situations.[9] However, the existing simulations including those by, for example, Hittner et al. (2003) and Boyer, Palachek, and Schucany (1983) were not performed in the context of determination of added value of subscores. In addition, the existing simulations involved analysis of data generated from distributions such as the uniform distribution that are quite distinct from typical distributions of test scores/subscores.

Therefore, future research could examine several such statistics in simulation studies in the context of determining the added value of subscores.

Even though an approach of hypothesis testing was suggested in this article, hypothesis testing has its own limitations (e.g., Wasserstein & Lazar, 2016). Therefore, some practitioners may prefer effect size over hypothesis testing and consider a subscore to be of added value only if the difference of PRMSE$(S'|S)$ and PRMSE$(S'|X)$ is larger than a value that they consider to be practically significant.[10] Similarly, given that subscores with added value are quite rare (see, e.g., a survey by Sinharay, 2010), some practitioners may consider a subscore to be of added value if the difference of PRMSE$(S'|S)$ and PRMSE$(S'|X)$ is positive, irrespective of the magnitude of the difference (i.e., without testing any hypothesis), following the recommendation of Haberman (2008). However, the methods suggested in this article may be useful to investigators who choose to perform hypothesis testing to determine if the added value of subscores or augmented subscores exceeds the extent that can be attributed to sampling variability.

Appendix

The R Code for Computing the Olkin's *Z*, Williams's *t*, and Hedges–Olkin's *Z* Statistics

```
library(subscore)
# Set up the data
test.inf=c(4,9,9,6,8)#4 subscores to be computed from 9, 9, 6, and 8 items
ns=test.inf[1]
Dat=read.csv("TIMSSdata.csv",header=T)# The data used in R package
''subscore''
n=nrow(Dat)
test.d=data.prep(Dat,test.inf)
# Use R package ''subscore'' to compute Haberman's PRMSEs etc.
Subscores = CTTsub(test.d,method="Haberman")
SubTot = Subscores$subscore.original
SP=Subscores$PRMSE # Haberman's PRMSEs
# Compute PRMSEs suggested by Sinharay (2013) from Haberman's PRMSEs
PRs=SP[1:ns,2]*SP[1:ns,2]
PRx=SP[1:ns,3]*SP[1:ns,2]
PRsx=SP[1:ns,4]*SP[1:ns,2]
rsx=cor(SubTot)[1:ns,(ns+1)]#Correlations between subscore & total
# Use R package ''cocor'' to compute Olkin's Z and Williams' t
library(cocor)
olk=rep(0,ns)
wil=rep(0,ns)
```

```
for (j in 1:ns)
{olk[j]=cocor.dep.groups.overlap(sqrt(PRs[j]),sqrt(PRx[j]),
 rsx[j],n)@olkin1967$statistic
 wil[j]=cocor.dep.groups.overlap(sqrt(PRs[j]),
 sqrt(PRx[j]),rsx[j],n)@williams1959$statistic}
 # A function to compute the Hedges-Olkin's Z statistic
 compsd=function(r01,r02,r12,r012,ns,n)
{a2=2*(r02-r12*r01)/(1-r12*r12)
 a1=-r12*a2
 a3=2*(r12*r01*r01+r12*r02*r02-r01*r02*(1+r12^2))/((1-r12^2)**2)
 V=matrix(0,3,3)
 v11s=(1-r01^2)^2/n
 v22s=(1-r02^2)^2/n
 v33s=(1-r12^2)^2/n
 z=rep(0,4)
 for (j in 1:ns)
 {V[1,2]=(0.5*(2*r12[j]-r01[j]*r02[j])*(1-r12[j]^2-r01[j]^2-
  r02[j]^2)+r12[j]^3)/n
  V[2,3]=(0.5*(2*r01[j]-r12[j]*r02[j])*(1-r12[j]^2-r01[j]^2-
  r02[j]^2)+r01[j]^3)/n
  V[1,3]=(0.5*(2*r02[j]-r12[j]*r01[j])*(1-r12[j]^2-r01[j]^2-
  r02[j]^2)+r02[j]^3)/n
  V[1,1]=v11s[j]
  V[2,2]=v22s[j]
  V[3,3]=v33s[j]
  for (i in 1:2){for (k in (i+1):3){V[k,i]=V[i,k]}}
  vec=c(a1[j],a2[j],a3[j])
  SD=sqrt(t(vec)%*%V%V*%vec)
  z[j]=(r012[j]*r012[j]-r01[j]^2)/SD}
  return(z)}
 zs=compsd(sqrt(PRs),sqrt(PRx),rsx,sqrt(PRsx),ns,n)#Compute z
 comparing PRMSEsx & PRMSEs
 zx=compsd(sqrt(PRx),sqrt(PRs),rsx,sqrt(PRsx),ns,n)#Compute z
 comparing PRMSEsx & PRMSEx
 hedgesolkin=ifelse(PRs>PRx,zs,zx)
# Write the results
 cat("\n  Haberman's PRMSEs: ",round(SP[1:ns,2],2),"\n")
 cat("\n  Haberman's PRMSEx: ",round(SP[1:ns,3],2),"\n")
 cat("\n Haberman's PRMSEsx: ",round(SP[1:ns,4],2),"\n")
 cat("\n Olkin's Z         : ",round(olk,2),"\n")
 cat("\n Williams' t       : ",round(wil,2),"\n")
 cat("\n Hedges-Olkin's Z  : ",round(hedgesolkin,2),"\n")
```

Acknowledgments

The author wishes to express sincere appreciation and gratitude to Li Cai, the editor, and the two anonymous reviewers for several helpful comments. The author thanks Shelby Haberman, Gautam Puhan, and Richard Feinberg for their helpful comments on an earlier version of this article.

Declaration of Conflicting Interests

The author prepared the work as employee of Educational Testing Service. Any opinions expressed in this publication are those of the author and not necessarily of ETS.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The variance of $\widehat{\text{PRMSE}}(T_S|S) - \widehat{\text{PRMSE}}(T_S|X)$ actually is the sum of the variances of $\widehat{\text{PRMSE}}(T_S|S)$ and $\widehat{\text{PRMSE}}(T_S|X)$ minus 2 times their covariance, and a theoretical expression for the covariance term is unknown.
2. The bootstrap procedure (e.g., Efron & Tibshirani, 1993) produced almost identical values as the jackknife procedure and hence is not considered henceforth.
3. Note that the $\sqrt{\widehat{\text{PRMSE}}(S'|S)}$ s are identical to the Cronbach's αs of the subscores.
4. Note that the distributions of test scores and subscores are far from exponential—so this result may not have any implication for testing for added value of subscores.
5. Some other subsample sizes were also used, but they led to very similar results.
6. The bootstrap procedure (Efron & Tibshirani, 1993) performed very similarly to the jackknife procedure—so the results for the former are not included here.
7. For example, Sinharay and Haberman (2014) found the standard deviation (*SD*; using the jackknife procedure of Efron, 1979) using the jackknife procedure of $\widehat{\text{PRMSE}}(S'|S) - \widehat{\text{PRMSE}}(S'|X)$ for a sample with 156 examinees to be 0.0891.
8. For example, Sinharay and Haberman (2014) found the *SD* of $\widehat{\text{PRMSE}}(S'|S) - \widehat{\text{PRMSE}}(S'|X)$ for a sample with 5,251 examinees to be 0.0071.
9. Steiger (1980) and Boyer, Palachek, and Schucany (1983) also reported satisfactory performance of Williams's *t*.
10. It may be easy to arrive at such a value given that these quantities are like reliability and practitioners typically have an idea on the gain in reliability that can be considered practically significant for their own tests.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, *57*, 289–300.

Boyer, J. E., Palachek, A. D., & Schucany, W. R. (1983). An empirical study of related correlation coefficients. *Journal of Educational Statistics*, *8*, 75.

Brennan, R. L. (2012). *Utility indexes for decisions about subscores* (CASMA Research Report No. 33). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment.

Dai, S., Wang, X., & Svetina, D. (2016). *subscore: Computing subscores in classical test theory and item response theory* (R package Version 2.0).

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLOS One*, *10*, e0121945.

Dunn, O. J., & Clark, V. (1969). Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association*, *64*, 366.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*, 1–26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman and Hall.

Feinberg, R., & Jurich, D. P. (2017). Guidelines for interpreting and reporting subscores. *Educational Measurement: Issues and Practice*, *36*, 5–13.

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

Hedges, L. V., & Olkin, I. (1983). Joint distributions of some indices based on correlation coefficient. In S. Carlin, T. Amemiya, & L. A. Goodman (Eds.), *Studies in econometrics, time series, and multivariate statistics* (pp. 437–454). New York, NY: Academic Press.

Hittner, J. B., May, K., & Silver, N. C. (2003). A Monte Carlo evaluation of tests for comparing dependent correlations. *The Journal of General Psychology*, *130*, 149–168.

Liu, Y., Robin, F., Yoo, H., & Manna, V. (2018). *Statistical properties of the GRE Psychology test subscores* (Educational Testing Service Research Report No. 18–19). Princeton, NJ: Educational Testing Service.

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

Lyren, P. (2009). Reporting subscores from college admission tests. *Practical Assessment, Research, and Evaluation*, *14*, 1–10.

Meijer, R. R., Boev, A. J., Tendeiro, J. N., Bosker, R. J., & Albers, C. J. (2017). The use of subscores in higher education: When is this useful? *Frontiers in Psychology*, *8*. doi:10.3389/fpsyg.2017.00305

Olkin, I. (1967). Correlations revisited. In J. C. Stanley (Ed.), *Improving experimental design and statistical analysis* (pp. 102–128). Chicago, IL: Rand McNally.

Olkin, I., & Finn, J. D. (1990). Testing correlated correlations. *Psychological Bulletin*, *108*, 330–333.

Olkin, I., & Finn, J. D. (1995). Correlations redux. *Psychological Bulletin*, *118*, 155–164.

Olkin, I., & Siotani, M. (1976). Asymptotic distribution of functions of a correlation matrix. In S. Ikeda (Ed.), *Essays in probability and statistics* (pp. 235–251). Tokyo, Japan: Shinko Tsusho.

Pearson, K., & Filon, L. N. G. (1898). Mathematical contributions to the theory of evolution. IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *191*, 229–311.

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: An evaluation of methods using empirical data. *Applied Measurement in Education*, *23*, 266–285.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*, 129–140.

Sawaki, Y., & Sinharay, S. (2017). Do the TOEFL iBT section scores provide value-added information to stakeholders? *Language Testing*. Advance online publication. doi:10.1177/0265532217716731

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

Sinharay, S. (2013). A note on assessing the added value of subscores. *Educational Measurement: Issues and Practice*, *32*, 38–42.

Sinharay, S. (2018). An interpretation of augmented subscores and their added value in terms of parallel forms. *Journal of Educational Measurement*, *55*, 177–193.

Sinharay, S., & Haberman, S. J. (2008). *Reporting subscores: A survey* (ETS Research Memorandum No. 08–18). Princeton, NJ: ETS.

Sinharay, S., & Haberman, S. J. (2014). An empirical investigation of population invariance in the value of subscores. *International Journal of Testing*, *14*, 22–48.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251.

Steiger, J. H., & Hakstian, A. R. (1982). The asymptotic distribution of elements of a correlation matrix: Theory and application. *British Journal of Mathematical and Statistical Psychology*, *35*, 208–215.

Thissen, D. (2013). Using the testlet response model as a shortcut to multidimensional item response theory subscore computation. In R. Millsap, L. van der Ark, D. Bolt, & C. Woods (Eds.), *New developments in quantitative psychology—Presentations from the 77th Annual Psychometric Society Meeting* (pp. 29–40). New York, NY: Springer.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*, 129–133.

Wedman, J., & Lyren, P. (2015). Methods for examining the psychometric quality of subscores: A review and application. *Practical Assessment, Research, and Evaluation*, *20*, 1–14.

Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society, Series B*, *21*, 396–399.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement*, *47*, 339–360.

Author

SANDIP SINHARAY is a principal research scientist at Educational Testing Service, Princeton, NJ 08541, USA; email: ssinharay@ets.org. His research interests include item response theory, assessment of model fit, reporting of subscores, statistical methods for detecting test fraud, and Bayesian methods.

Manuscript received February 20, 2018
Revision received May 15, 2018
Accepted June 18, 2018