JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education

Journal of Educational Measurement
Summer 2010, Vol. 47, No. 2, pp. 150–174

How Often Do Subscores Have Added Value? Results from Operational and Simulated Data

Sandip Sinharay
Educational Testing Service

Recently, there has been an increasing level of interest in subscores for their potential diagnostic value. Haberman suggested a method based on classical test theory to determine whether subscores have added value over total scores. In this article I first provide a rich collection of results regarding when subscores were found to have added value for several operational data sets. Following that I provide results from a detailed simulation study that examines what properties subscores should possess in order to have added value. The results indicate that subscores have to satisfy strict standards of reliability and correlation to have added value. A weighted average of the subscore and the total score was found to have added value more often.

There is an increasing interest in subscores, which are scores on subtests, because of their potential diagnostic value. Individual examinees want to know their strengths and weaknesses in different content areas to plan for future remedial work. States and academic institutions such as colleges and universities often want a summary of performance for their students to better evaluate their training and focus on areas that need instructional improvement (Haladyna & Kramer, 2004). The U.S. Government's No Child Left Behind (NCLB) Act of 2001 demands, among other things, that students should receive diagnostic reports that allow teachers to address their specific academic needs; subscores could be used in such a diagnostic report. The total score, consisting of a single number, is often argued to be too deterministic; showing how the examinee's abilities vary over the different subtests may be of more use. Finally, it may be possible to improve predictive validity by using the subscores.

Despite this apparent usefulness of subscores, certain important factors must be considered before making a decision on whether to report subscores at either the individual or institutional level. According to Standard 5.12 of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 1999), scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established and the standard applies to subscores as well. Further, Standard 1.12 of the Standards for Educational and Psychological Testing demands that if a test provides more than one score, the distinctiveness of the separate scores should be demonstrated. Several researchers such as Wainer et al. (2001) and Tate (2004) also emphasized the importance of ensuring reasonable subscore performance.

Inspired by the above need to assess the quality of subscores, Haberman (2008) and Haberman, Sinharay, and Puhan (2009) recently suggested statistical methods

based on classical test theory (CTT) to examine whether subscores have added value (the next section describes when a subscore is defined to have added value) over total scores. These papers, as well as Puhan, Sinharay, Haberman, and Larkin (2008), Sinharay and Haberman (2008), and Lyren (2009), analyzed data sets from a variety of testing programs. They found that there are only a handful of tests for which subscores had added value.

A question that testing program staff often face, especially while designing new tests that intend to report subscores, is: "What properties should the subscores possess in order to have added value?" Especially, the testing program staff would like to know more about how reliable and how distinct their subscores should be in order to have added value. These questions became even more pertinent after the research work of Haberman (2008), as previously there was no obvious method to determine whether a subscore has added value. Only partial answers to the above-mentioned question are provided by Haberman (2008) and Sinharay, Haberman, and Puhan (2007), who explained that a subscore is more likely to have added value when (a) it has high reliability, (b) the total score has low reliability, and (c) it is distinct from the other subscores. More research attempting to answer these questions will be beneficial.

In this article I first summarize results obtained from the analysis of operational data that are relevant to the above-mentioned question; some of these results are taken from existing literature. Then, I provide results from a detailed simulation study that was designed to obtain more information on when subscores can be expected to have added value. Data are simulated from a multivariate item response theory (MIRT) model. In the simulation study I used estimated item parameters from operational tests as the generating item parameters in the MIRT model. This makes the simulation study somewhat realistic. Several factors that are likely to affect whether subscores have added value are manipulated in the simulation study.

Methods from Classical Test Theory

In this section I describe the approach of Haberman (2008) to determine whether and how to report subscores. Let us denote the subscore and the total score of an examinee as s and x, respectively. Haberman (2008) and Sinharay et al. (2007), taking a CTT viewpoint, assumed that a reported subscore is intended to be an estimate of the true subscore s_t and considered the following estimates of the true subscore:

- An estimate $s_s = \bar{s} + \alpha(s - \bar{s})$ based on the observed subscore, where $\bar{s}$ is the average subscore for the sample of examinees and α is the reliability of the subscore.
- An estimate $s_x = \bar{s} + c(x - \bar{x})$ based on the observed total score, where $\bar{x}$ is the average total score and c is a constant that depends on the reliabilities and standard deviations of the subscore and the total score and the correlations between the subscores.
- An estimate $s_{sx} = \bar{s} + a(s - \bar{s}) + b(x - \bar{x})$ that is a weighted average of the observed subscore and the observed total score, where a and b are constants that depend on the reliabilities and standard deviations of the subscore and the

total score and the correlations between the subscores. This will be referred to as the *weighted average* hereafter.

It is also possible to consider an augmented subscore s_{aug} that is an appropriately weighted average of all the subscores of an examinee (Wainer et al., 2001) as an estimate of s_t. The estimate s_{sx} is a special case of the augmented subscore s_{aug}; s_{sx} places the same weight on the subscores other than the one of interest while s_{aug} places different weights on all the subscores. However, for the simulated and operational data examples discussed later in this article, s_{aug} yielded results that are very similar to those for s_{sx}. Hence, in this article I do not provide any results for s_{aug}.

To compare the performances of s_s, s_x, and s_{sx} as estimates of s_t, Haberman (2008) suggested the use of the proportional reduction in mean squared error (PRMSE). The larger the PRMSE, the more accurate is the estimate.[1] In this article I will denote the PRMSE for s_s, s_x, and s_{sx} as $PRMSE_s$, $PRMSE_x$, and $PRMSE_{sx}$, respectively. The PRMSE is conceptually similar to reliability, and the quantity $PRMSE_s$ can be shown to be exactly equal to the reliability of the subscore. Haberman (2008) recommended the following strategy regarding determination of whether a subscore or a weighted average has added value:

- A subscore has added value over the total score only if $PRMSE_s$ is larger than $PRMSE_x$, because the subscore will provide more accurate diagnostic information (in the form of a lower mean squared error in estimating the true subscore) than the observed total score in that case. Sinharay et al. (2007) discussed why this strategy is reasonable and how this ensures that a subscore satisfies professional standards.
- The quantity $PRMSE_{sx}$ will always be at least as large as $PRMSE_s$ and $PRMSE_x$. However, s_{sx} requires a bit more computation than either s_s or s_x. Hence, a weighted average has added value only if $PRMSE_{sx}$ is substantially larger compared to both $PRMSE_s$ and $PRMSE_x$.

If neither the subscore nor the weighted average has added value, diagnostic information should not be reported for the test, and alternatives such as scale anchoring (Beaton & Allen, 1992) should be considered. The computations for application of the method of Haberman (2008) are simple and involve only the sample variances, correlations, and reliabilities of the total score and the subscores. Note that the methodology does not involve any assumptions except those of classical test theory.

The computations in the methodology of Haberman (2008) use the disattenuated correlations among the subscores. For some data sets, due to extremely high correlations between subscores, disattenuated correlations between subscores may occasionally be larger than 1. If this happens, while it is possible to apply the method of correction for attenuation (Bock & Petersen, 1975), it is obvious that neither subscores nor weighted averages should be reported.

Review of Results from Operational Data Analysis

Table 1 summarizes results from several operational data sets. For all the tests considered in the table, the subscores refer to the operationally reported subscores or section scores unless mentioned otherwise.

The P-ACT®+ English and Mathematics tests were discussed by Harris & Hanson (1991), who used three forms each of these tests. Table 1 shows results for one form each of these tests. The PRMSEs were computed using the information in Harris and Hanson. The subscore reliabilities were not provided in Harris and Hanson. However, the correlation and disattenuated correlation between the subscores were provided; these were used to compute the product of the reliabilities of the two subscores, and then the Spearman–Brown prophecy formula was used to estimate the reliabilities of the subscores and the total score (as the length of the subtests is known).[2]

The results for SAT® I Verbal, SAT I Math, and SAT I were discussed by Haberman (2008), who analyzed one form of the October 2002 administration of the SAT I examination.[3] For SAT I Verbal, the subscores refer to the critical reading, analogies, and sentence completion scores. For SAT I Math, the subscores refer to the scores on four-choice multiple choice questions, five-choice multiple choice questions, and student-produced responses; these scores were not operationally reported. For SAT I, the subscores actually refer to the SAT I Verbal and SAT I Math scores. The results for Praxis were discussed by Haberman (2008).

Table 1 includes the results for the Spring 2006 assessment of the Delaware Student Testing Program (DSTP) 8th grade mathematics assessment. The data from the test were analyzed in Stone, Ye, Zhu, & Lane (2010), who reported, using an exploratory factor analysis method, the presence of only one factor in the data set. The four subscores that were proposed, but are not reported, correspond to four content domains: numeric reasoning, algebraic reasoning, geometric reasoning, and quantitative reasoning. The summary statistics reported in Stone et al. were used to compute the PRMSEs required in the method of Haberman (2008).[4]

"State Reading (5th grade)" refers to a fifth grade end-of-grade assessment of reading ability in a Midwestern state. Results for one form of the test are shown in Table 1. The purpose of the test is to assess the proficiency level of examinees for meeting NCLB requirements. The data were analyzed by Ackerman and Shu (2009) and Henson, Templin, and Irwin (2009). The test has 73 multiple choice items. According to the test specification manual, 55 items were intended to measure ability to understand the meaning of words and phrases and 18 items were intended to measure comprehension; these are the two subscores considered in this article, as in Ackerman and Shu and Henson et al.[5] though they are not operationally reported. The summary statistics reported in Ackerman and Shu were used to compute the PRMSEs.

Table 1 includes the results for the 2006 spring administration of the Swedish Scholastic Assessment Test (SweSAT), which is used for selection to higher education in Sweden (Lyren, 2009). SweSAT reports five subscores: Vocabulary; Swedish Reading Comprehension; English Reading Comprehension; Data Sufficiency; and Diagrams, Tables, and Maps. Some of the results shown in Table 1 are from Lyren, who applied the method of Haberman (2008) to five forms of SweSAT. The average

Table 1
Results from Analysis of Operational Data Sets

Name/Nature of the Test	No. of Subscores	Av. Length	Av. α	Av. Corr.	Average Corr. (disatt.)	How many Subscores Have Added Value?	How Many wtd. avs Have Added Value?
CC (for beginning teachers)	4	19	.38	.44	1.00	0	0
TA (measures cognitive and technical skills)	7	11	.42	.51	1.00	0	0
CB (for teachers of special ed. programs)	3	19	.46	.42	.96	0	0
CG (for teachers of mathematics)	3	16	.62	.59	.95	0	0
TD1 (measures school and individual student progress)	4	15	.70	.73	.98	0	0
TD2 (measures school and individual student progress)	6	13	.70	.75	1.00	0	0
DSTP Math (8th grade)	4	19	.77	.77	1.00	0	0
SAT I Math	3	20	.78	.75	.97	0	0
P-ACT+ Eng	2	25	.79	.76	.95	0	0
P-ACT+ Math	2	20	.71	.67	.94	0	1
State Reading (5th grade)	2	37	.72	.65	.92	0	1
SAT I Verbal	3	26	.79	.74	.95	0	1
CH (for paraprofessionals)	3	24	.85	.76	.89	0	3
CF (for principals and school leaders)	4	15	.48	.41	.85	0	4
MFT Business	7	17	.56	.47	.85	0	6

(Continued)

Table 1
Continued

Name/Nature of the Test	No. of Subscores	Av. Length	Av. α	Av. Corr.	Average Corr. (disatt.)	How many Subscores Have Added Value?	How Many wtd. avs Have Added Value?
CD (for teachers of social studies)	6	22	.63	.54	.87	0	6
CE (for teachers of Spanish)	4	29	.80	.65	.80	1	2
TB1 (tests mastery of a language)	2	44	.85	.77	.90	1	2
TC1 (measures achievement in a discipline)	3	68	.85	.76	.90	1	3
CA (for teachers in elementary schools)	4	30	.74	.59	.79	1	4
TB2 (tests mastery of a language)	2	43	.90	.68	.75	2	2
SAT I	2	69	.92	.70	.76	2	2
TC2 (measures achievement in a discipline)	3	67	.87	.72	.82	2	3
Praxis	4	25	.72	.56	.78	2	4
SweSAT	5	24	.78	.55	.71	4	5

Note. The reliability is denoted as α. Weighted averages are denoted as "wtd. avs".

disattenuated correlation of .69 for the test is the lowest among all the tests shown in Table 1.

Table 1 also includes the results for the major field test of business (MFT Business) that is a comprehensive outcomes assessment of knowledge obtained by students in a business major (for an associate, bachelor, or MBA degree). The test form considered here, administered between 2002 and 2006, consists of 118 multiple choice items. Seven subscores (accounting, economics, management, quantitative business analysis and information systems, finance, marketing, and legal and social environment) are reported at aggregate levels (e.g., average subscore of a class or a program) and not for individual examinees. The data considered here were analyzed in Ling (2009).

The eight tests denoted as CA-CH are certification tests discussed in Puhan et al. (2008). The seven tests denoted as TA, TB1, TB2, TC1, TC2, TD1, and TD2 were discussed in Sinharay and Haberman (2008). For test TA, the seven subscores, each corresponding to a skill area the test is supposed to measure, were originally intended to be reported, but actually are not reported.

Each row in Table 1 shows, for a particular test, the number of subscores, average length of the subscores, average reliability of the subscores, average correlation among the subscores, average disattenuated correlation[6,7], the number of subscores that have added value, and the number of weighted averages that have added value (where the assumption was made that a weighted average has added value if the corresponding $PRMSE_{sx}$ is larger than the maximum of $PRMSE_s$ and $PRMSE_x$ by .01 or larger[8]). The tests are sorted first by the number of subscores that had added value (low to high), then by the number of weighed averages that had added value, and finally by average reliability of the subscores.

Most of the tests had only multiple choice items. For these tests, "length" refers to the number of items. Some tests such as CF had constructed response (CR) items with score categories 0, 1, 2, ... For a subscore involving CR items, "length" refers to the maximum score (for example, for a subscore with four items, each with three score categories, 0, 1, and 2, the length is 8).

The reliability of the different scores and subscores that are reported in Table 1 were estimated using Cronbach's α. Some researchers such as Kamata, Turhan, and Darandari (2003) have argued that when a test consists of several subtests (which is usually the case when subscores are considered), the stratified α may be a more accurate estimate of the true reliability than α. This article uses α because it is most often used in operational assessments. For a few of the data sets for which stratified α was computed, the values of stratified α were very close to those of α. For example, the stratified α was the same as α up to 2 decimal places for seven out of the eight tests CA-CH. This is primarily due to the usually high correlation between the subscores. Kamata et al. found in a simulation study that stratified α, although better than α overall, was not much different from α for highly correlated subtests. Because stratified α is slightly higher than α for the total score and will be the same as α for the subscores, the use of stratified α rather than α will make a subscore slightly less likely to have added value.

Figures 1 to 3 show, for the operational data sets, the percentages of subscores (Figures 1 and 2) or weighted averages (Figure 3) that had added value. In each of

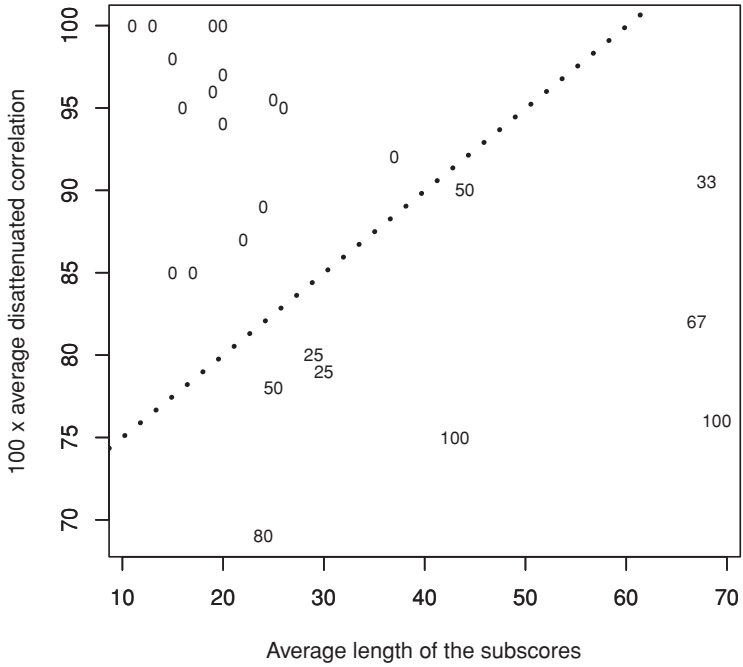

Figure 1. The percent of subscores that have added value for different average subscore length and average disattenuated correlation for the operational data.

these figures, the *Y*-axis corresponds to the average disattenuated correlation among the subscores. In Figure 1, the *X*-axis denotes the average length of the subscores, while, in Figures 2 and 3, the *X*-axis denotes average subscore reliability. In the three figures there are plots for each row listed in Table 1, a number that is the same as the percentage of subscores (or weighted averages) that have added value at the point (x, y), where x is the corresponding average subscore reliability multiplied by 100 (or length in Figure 1) and y is 100 times the average disattenuated correlation. For example, in Table 1, SAT I has average length 69, average disattenuated correlation .76, and two subscores (that is 100% of all subscores) that had added value. Hence Figure 1 has the number 100 plotted at the point (69,76).

Table 1 and Figures 1 to 3 show that, in general, subscores consisting of a large number of items (which have high reliability) tend to have added value. For example, for the test TC2, subscores consisting of about 67 items had added value. However, not all subscores consisting of a large number of items have added value. For example, for the test TC1, which has an average subtest length of 68, only one of three subscores has added value. Tests with low average disattenuated correlations tended to have subscores with added value. However, for the test CF, the average disattenuated correlation is .85, and none of the subscores have added value, while, for the test TB1, the average disattenuated correlation is .90, but one of the two subscores has added value.

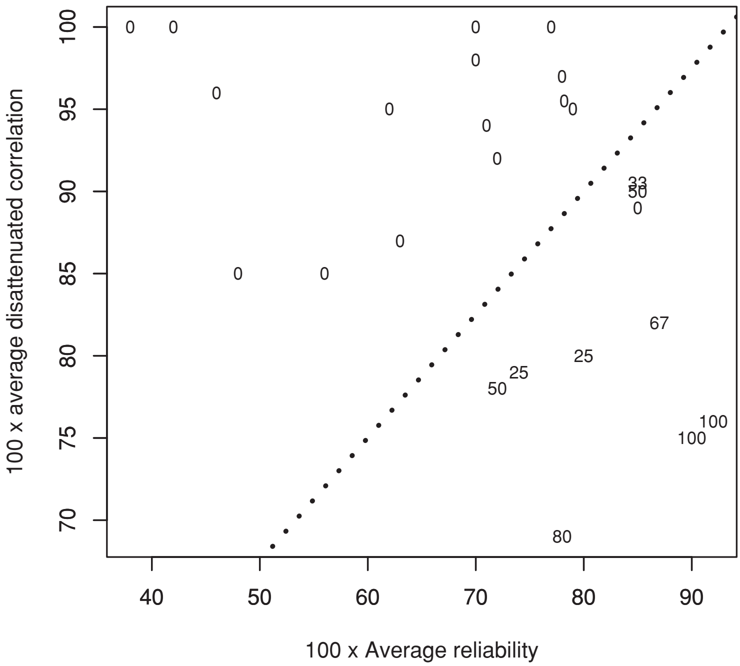

Figure 2. The percent of subscores that have added value for different average subscore reliability and average disattenuated correlation for the operational data.

Often, percentage of subscores with a specific average length (or average reliability) that have added value depends on the average disattenuated correlations.[9] Hence, each of Figures 1 to 3 includes a bold dotted line roughly dividing the plot into two regions in which the percentage is low (zero) or high (positive). Note that this line is not unique and was drawn after a visual examination of the points in the plot and not using any mathematical formula. In each of these figures, as one goes from the top left corner to the bottom right corner (that is, as the average length/reliability increases and the average disattenuated correlation decreases), the subscores show more tendency to have added value. Figure 3 shows that the weighted averages have added value for many of the operational data sets and that weighted averages are much more likely to have added value compared to the subscores themselves. While there are 16 zeroes in Figures 1 or 2, there are only nine in Figure 3.

While Table 1 and Figures 1 to 3 have the advantage of being based on real data, one could argue that they were based on only a few data sets so that if one obtains another collection of data sets, the table and the figure might change. In addition, few of these tests had subscores that have added value (for example, there are only four points in Figure 1 with the percentage larger than 50). Besides, there are extraneous factors such as the nature of the test that affect the above results and it is difficult to remove their effects from these results. Finally, there are some gaps in Figures 1 to 3. For example, there are only two tests with average length around 50, with none of

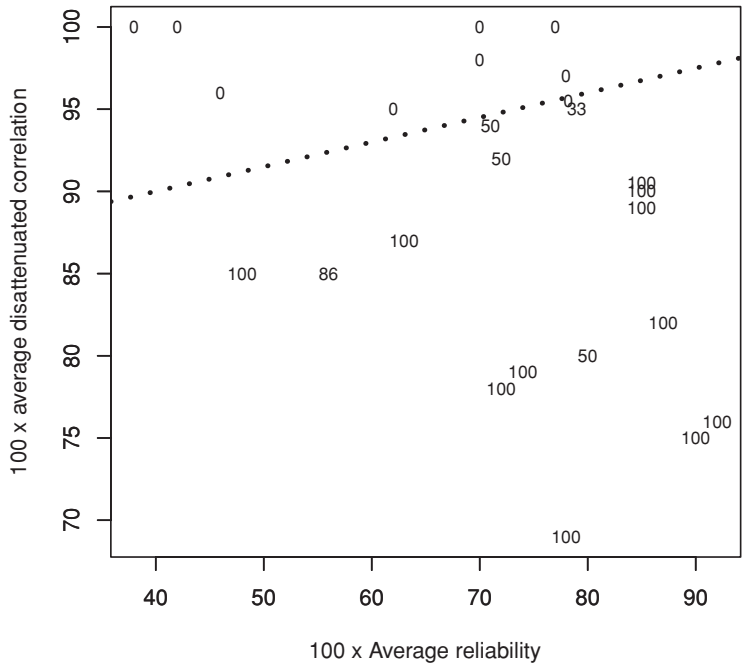

Figure 3. The percent of weighted averages that have added value for different average subscore reliability and average disattenuated correlation for the operational data.

them having average disattenuated correlation between .75 and .90. A decision based on Figures 1 to 3 for a data set whose corresponding point falls in one of these gaps will require extrapolation and may end up being inaccurate. Hence, while the results from Table 1 can provide some guidance to testing programs, they leave some gaps.

Hence a simulation study was performed where it was easy to control different factors and study the effects of the factors of interest. The results from the simulated data are expected to augment the findings from the real data in providing better guidance to the testing programs interested in subscores. The simulation study is discussed in the next section.

Simulation Study

The MIRT Model

In this section I discuss results for data simulated from the 2-parameter logistic MIRT model (Reckase, 2007; Haberman, von Davier, & Lee, 2008) for which the item response function for item i is given by

$$\left[1 + e^{-(a_{1i}\theta_1 + a_{2i}\theta_2 + \cdots + a_{Ki}\theta_K - b_i)}\right]^{-1}, \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ is the K-dimensional ability parameter for an examinee, b_i is a scalar item parameter that is related to the difficulty of item i, and $(a_{1i}, a_{2i}, \ldots a_{Ki})$ is a vector of item discrimination parameters. Each component of

$\boldsymbol{\theta}$ corresponds to a subscore. For any item i, only one among the parameters a_{1i}, $a_{2i}, \ldots, a_{Ki}$ is assumed to be nonzero, depending on the subscore the item contributes to (e.g., for an item belonging to the first subscore, a_{1i} is nonzero while $a_{2i} = a_{3i} = \cdots = a_{Ki} = 0$), so that the simulations are performed from a simple-structure MIRT model (which means that the subscores do not share common items).

Simulation Design

Generating item parameters. One data set was obtained from each of three tests on which subscores or section scores are operationally reported. On the first test, which is a test in English, two section scores are reported (they are treated as subscores here), each of which is based on 100 multiple choice items. On the second test, which is the test TC2 in Table 1, three subscores are reported, which consist of 66–67 multiple choice items (Sinharay & Haberman, 2008). On the third test, which is the test CA in Table 1, four subscores are reported (language arts/reading, mathematics, social studies, and science), each based on 30 items (Puhan et al., 2008).

The marginal maximum likelihood estimates of the item parameters of the model given by Expression 1 were computed using the stabilized Newton-Raphson algorithm (Haberman et al., 2008) from each of the three data sets. The prior distribution assumed, as in Haberman et al., was

$$\boldsymbol{\theta} \sim \mathcal{N}_K(\mathbf{0}, \Sigma), \tag{2}$$

where $\mathcal{N}_K$ denotes the density of the normal distribution with K dimensions and $\mathbf{0} = (0, 0, \ldots, 0)'$. The diagonals of Σ are set to 1 to ensure identifiability of the model parameters, and the off-diagonals are estimated from the data. For each data set, each operationally reported subscore or section score is considered to measure a skill area and is assumed to contribute to one dimension of $\boldsymbol{\theta}$. The estimated item parameter values were used later as generating item parameters in the simulation study and the estimated correlations between the components of $\boldsymbol{\theta}$ were used later to generate the ability parameters. There was a need to generate test data with different numbers of items per subscore using the estimated item parameters from each test. For example, using the item parameters estimated from a data set from the test CA, that has 30 items per subscore, data were generated with 10, 20, 30, and 50 items per subscore. To do that in a systematic manner, the distribution of the estimated item parameters was used to generate the item parameters for the simulated data sets. For each test data set, a bivariate normal distribution $\mathcal{B}_k$ was fitted to the log-a and b-parameter estimates of the items belonging to kth subscore, $k = 1, 2, \ldots K$, and the generating item parameters for the kth subscore were randomly drawn from $\mathcal{B}_k$. Thus, for example, a bivariate normal distribution was fitted to the log-a and b-parameter estimates of the items contributing to the first subscore of test CA and then the generating parameters of 10, 20, 30, or 50 items, depending on the simulation condition, for the first subscore of test CA were randomly drawn from this distribution.

Factors controlled in the simulation study. The following factors were controlled in the simulation studies:

- Number of subscores. For each of the three above-mentioned tests (that have two, three, and four subscores, respectively), the estimated item parameters were used to simulate data for which the number of subscores (or the dimension of $\boldsymbol{\theta}$) is the same as that reported for the test. For example, the estimated item parameters from the data set from the test TC2 (that reports three subscores) were used to simulate data that have three subscores. Hence, in the simulations, the "number of subscores" can take one of three values: 2, 3, and 4. However, the "number of subscores" refers to more than simply the number of subscores. Each level of this factor also has its own set of item parameters obtained from an operational test data set as described above.
- Length of the subscores. In this study I used four values for the length: 10, 20, 30, and 50. Note that the reliability of a test increases as the test length increases. For simplicity, in this article I assumed that the different subscores for a given test have the same length. The appendix shows some results for the case when the different subscores for a given test may have different lengths.
- Level of correlation (ρ) among the components of $\boldsymbol{\theta}$. This article used six levels: .70, .75, .80, .85, .90, and .95. If the correlation level for a simulation case is ρ, the mean of all the off-diagonal elements of Σ (which denote the correlations between the components of $\boldsymbol{\theta}$) in Equation 1 was set equal to ρ to simulate the data sets. The starting point was the estimated correlation matrix $\mathcal{C}$ between the components of $\boldsymbol{\theta}$ from the fit of the model given by Equation 1 to an operational data set. The matrix $\mathcal{C}$ was

$$\begin{pmatrix} 1.00 & .82 & .70 \\ .82 & 1.00 & .86 \\ .70 & .86 & 1.00 \end{pmatrix} \tag{3}$$

  for the three-subscore case, and

$$\begin{pmatrix} 1.00 & .78 & .80 & .84 \\ .78 & 1.00 & .72 & .78 \\ .80 & .72 & 1.00 & .85 \\ .84 & .78 & .85 & 1.00 \end{pmatrix} \tag{4}$$

  for the four-subscore case.[10] To obtain a Σ with a mean correlation ρ, m, the mean of the correlations, was computed from $\mathcal{C}$ and then the (i, j)th element of Σ was set as the (i, j)th element of $\mathcal{C} - m + \rho$, where $i \neq j$.[11] This strategy ensured that the average of the correlations in Σ is ρ, but allowed the correlations between the subscores to be realistically different. Note that the correlations among the components of $\boldsymbol{\theta}$ are similar to the disattenuated correlations between the subscores. Hence, from Table 1, the choice of the above six levels of the correlation (especially, the lowest and highest of them) is reasonable.
- Sample size N. In this study I used three levels of the sample size: 100, 1,000, and 4,000.

Steps in the simulation study. For each simulation condition (determined by a value each of the number of subscores, length of the subscores, level of correlation, and sample size), the generating item parameters were drawn once as described above (from the bivariate normal distributions $\mathcal{B}_k$s), and then $R = 100$ replications[12] were performed. Each replication involved the following steps:

1. Generate the ability parameter $\boldsymbol{\theta}$ for each of the N examinees from the multivariate normal distribution $\mathcal{N}_K(\mathbf{0}, \Sigma)$, where the diagonals of Σ are 1 and the off-diagonals are computed as described above.
2. Simulate a data set, that is, simulate scores on each item of the test for each examinee, using Equation 1, the draws of $\boldsymbol{\theta}$ in the above step and the above-mentioned generating item parameters for the test.
3. Calculate, for the simulated data set, several quantities, such as correlations among the subscores and the PRMSEs.

Results

Table 2 shows results for sample size of 1,000. The table shows results for four (out of six) values of the level of correlation. Each of the 18 cells (where a cell corresponds to a simulation case) of the table shows the following eight quantities:

1. $100 \times$ the average reliability of the total score (denoted as α_{tot} in the table), where the average is taken over the R replications.
2. $100 \times$ the average reliability (remember that reliability $= PRMSE_s$) of the subscores (denoted as PR_s), where the average is taken over the appropriate number of subscores (for example, two subscores when the number of subscores $= 2$) in each replication and then over the R replications.
3. $100 \times$ the average correlation between the subscores (denoted as r), where the average is taken over the appropriate number of correlations (for example, six correlations when the number of subscores $= 4$) in each replication and then over the R replications.
4. $100 \times$ the average disattenuated correlation between the subscores. This is denoted as r_d in the tables.
5. $100 \times$ average $PRMSE_x$ (denoted PR_x), where the average is taken over the appropriate number of subscores in each replication and then over the R replications.
6. $100 \times$ average $PRMSE_{sx}$ (denoted as PR_{sx}).
7. Overall percent of subscores that have added value (denoted as % sub). This is the overall percent of cases (out of a total of $R \times K$, where K is the number of subscores) when $PRMSE_s$ is larger than $PRMSE_x$.
8. Overall percent of weighted averages that have added value (denoted as % wtd). This is the overall percent of cases (out of a total of $R \times K$) when $PRMSE_{sx}$ is larger than the maximum of $PRMSE_s$ and $PRMSE_x$ by .01 or more.

The first eight lines of numbers of Table 2 show the results for two subscores, the next eight lines for three subscores, and the last eight lines for four subscores.

Table 2
Summary of the Simulated Data for Sample Size 1,000

No. of Subscores	Length of the Subscores															
	10 Items Correlation				20 Items Correlation				30 Items Correlation				50 Items Correlation			
	.70	.80	.90	.95	.70	.80	.90	.95	.70	.80	.90	.95	.70	.80	.90	.95
2																
α_{tot}	73	75	76	77	85	86	86	87	89	90	90	91	93	94	94	94
PR_s	62	62	63	63	77	77	77	77	83	83	83	83	89	89	89	89
r	44	51	58	61	54	62	69	74	57	66	75	79	62	71	80	85
r_d	71	82	92	97	70	80	90	96	69	79	90	95	69	79	90	95
PR_x	63	68	73	76	72	77	82	85	75	80	86	88	79	84	89	92
PR_{sx}	68	70	74	76	80	81	84	85	85	86	87	89	90	91	92	93
% sub	36	00	00	00	100	46	00	00	100	100	01	00	100	100	57	00
% wtd	100	94	08	01	100	100	65	01	100	100	98	03	05	96	100	16
3																
α_{tot}	75	77	78	79	86	87	88	88	90	91	92	92	94	94	95	95
PR_s	56	56	56	56	72	72	72	72	80	80	80	80	87	87	87	87
r	39	45	51	54	50	58	65	69	56	64	71	76	61	69	78	82
r_d	70	80	91	96	70	80	90	95	70	80	90	95	70	80	90	95
PR_x	61	67	74	77	69	76	82	85	72	79	86	89	75	82	89	92
PR_{sx}	67	70	75	77	78	81	84	86	83	85	88	90	89	90	92	93
% sub	41	03	00	00	67	46	01	00	76	67	26	00	100	67	65	04
% wtd	95	71	59	21	100	86	67	33	100	97	67	39	75	100	67	53
4																
α_{tot}	80	82	83	84	89	90	91	91	92	93	94	94	95	96	96	96
PR_s	57	57	57	57	72	72	72	72	80	80	80	80	87	87	87	87
r	40	46	52	55	50	57	65	69	55	63	72	76	60	69	78	82
r_d	70	81	91	97	69	80	90	95	69	80	90	95	69	79	90	95
PR_x	62	70	78	82	69	76	84	88	71	79	87	90	73	81	89	93
PR_{sx}	69	73	79	82	79	82	86	89	84	86	89	91	89	90	92	94
% sub	25	07	00	00	71	25	07	00	99	39	25	00	100	100	25	23
% wtd	100	89	32	23	100	100	54	25	100	100	72	26	75	98	90	31

Note. α_{tot} denotes average reliability of the total score; PR_s denotes 100 × the average reliability of the subscores; r denotes 100 × the average correlation between the subscores; r_d denotes 100 × the average disattenuated correlation between the subscores; PR_x denotes 100 × average $PRMSE_x$; PR_{sx} denotes 100 × average $PRMSE_{sx}$; PR_{aug} denotes 100 × average $PRMSE_{aug}$; % sub denotes the overall percent of subscores that have added value; % wtd denotes the overall percent of weighted averages that have added value.

Note that Table 2 shows the average values of different quantities for any simulation case. Because the correlations between the subscores are different for the three- and four-subscore cases, the values of, for example, reliability or the percent of times when a subscore has added value will vary from one subscore to another. So, for example, a subscore that has lower correlations with the other subscores on an average

will be more likely to have added value than the others. For example, for the case with three subscores, length = 10, and correlation = .70, the percent of times when a subscore has added value was 94, 0, and 28 for the three subscores; their average, 41, is reported in Table 2.

For each simulation case in Table 2, the number of items contributing to the different subscores in a test was the same. The appendix shows results of a similar simulation where the number of items contributing to the subscores was different. For a fixed average number of items per subscore, the percent of subscores that have added value was virtually the same whether the number of items contributing to the subscores in a test are all the same or different.

Figures 4 to 7 show, for simulated data with sample size of 1,000, the overall percentage of subscores (Figures 4 to 6) or weighted averages (Figure 7) that have added value. These plots (unlike Table 2) show results for all the six levels of correlation from .70 to .95. Figure 4 is a three-dimensional scatter plot showing the overall percentage of subscores that have added value (shown along the Z-axis using a vertical line) for each subscore length and level of correlation. Figures 5 to 7 are like Figure 1

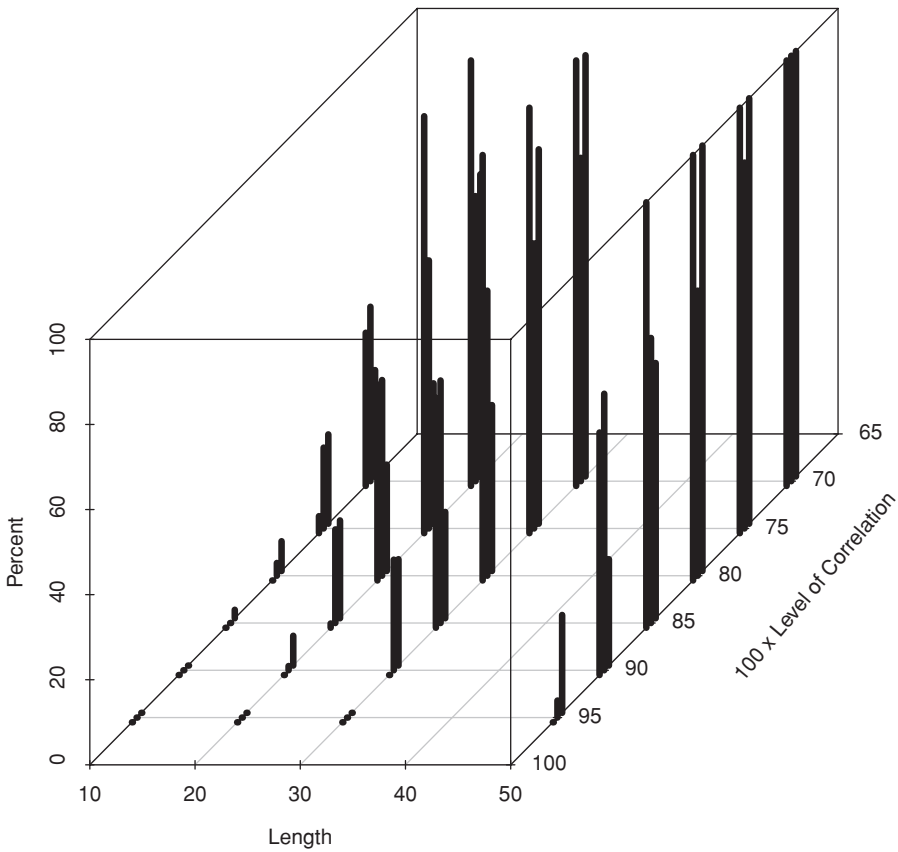

Figure 4. A Three-dimensional scatter plot showing the overall percent of subscores that have added value versus length and level of correlation for simulated data with sample size 1,000.

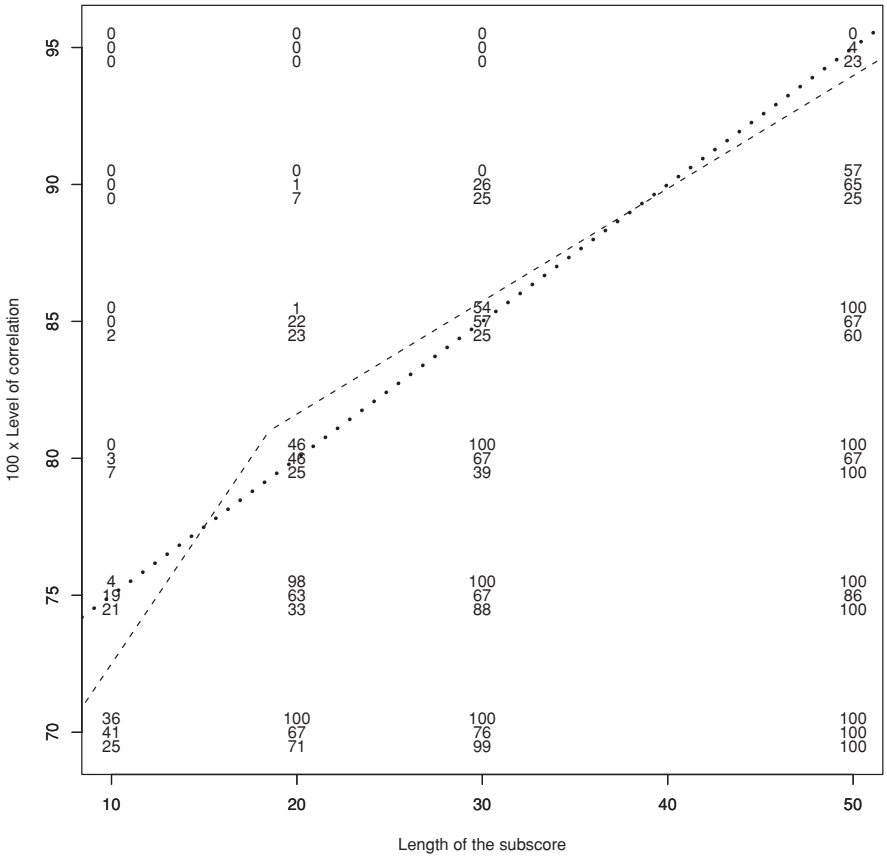

Figure 5. The overall percent of subscores that have added value for different subscore length and level of correlation for simulated data with sample size 1,000.

and show, for each combination of subscore length (or average reliability) and level of correlation, a number showing the percentage. Figures 5 to 7 show dashed lines roughly dividing the plot into two regions in which the percentage is low (less than 25%, roughly) and high (more than 25%). These three figures also reproduce the corresponding bold dotted lines from Figures 1 to 3 to assist a comparison of results from the operational and simulated data.

In Figures 4 to 7, there may be up to three points (corresponding to the three values of number of subscores) at each (X, Y)-coordinate denoting the percentages for the two-, three-, and four-subscore cases, respectively. For convenience of viewing, the percentages for the two-subscore cases are shown slightly above those for the three-subscore cases and the percentages for the four-subscore case are shown slightly below those for the three-subscore cases.

Table 2 and Figures 4 to 7 lead to the following conclusions:

- Overall, the percent of times when the subscores have added value increases with an increase in their lengths (or reliability) and with a decrease in the

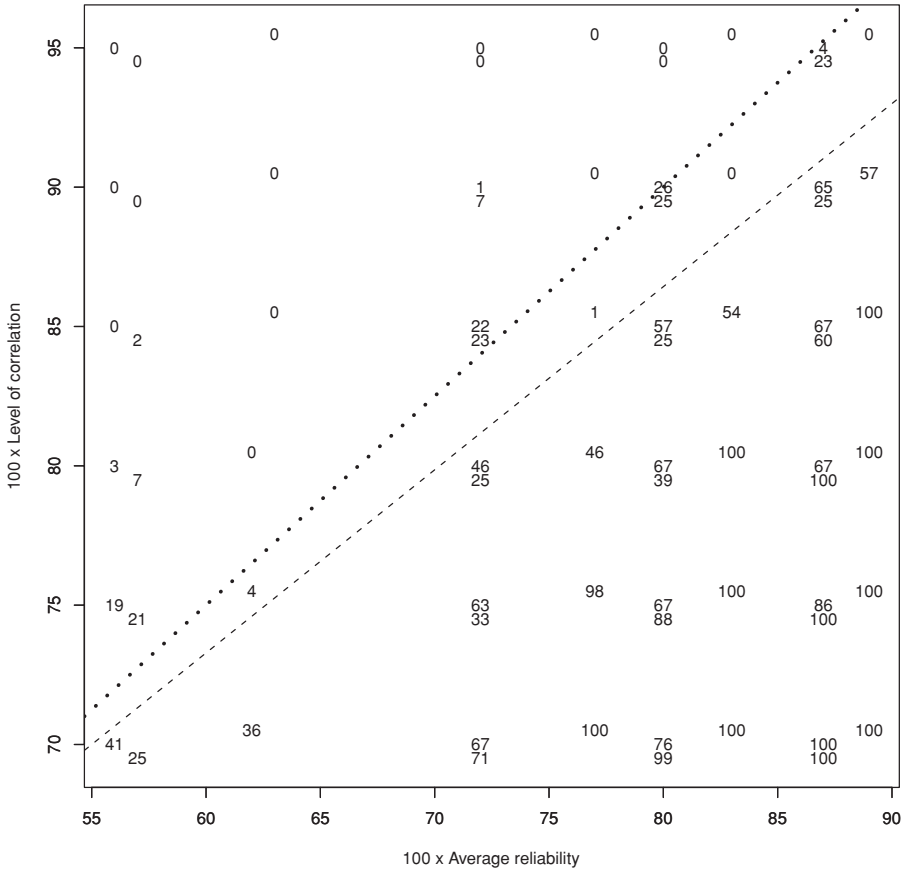

Figure 6. The overall percent of subscores that have added value for different average subscore reliability and level of correlation for simulated data with sample size 1,000.

correlations among them (that is, as they become more distinct). This is expected from the discussions in Haberman (2008).

- If the average length of the subscores is 10, subscores are rarely of added value. Of 16 such cases in Table 2, the percent of times when the subscores have added value is less than 1 in nine cases and has a significant nonzero value only when the level of correlation is .70, which, according to Table 1, is rare in practice. This conclusion supports the findings of Table 1 in which none of the subscores that consist of a few items had any added value, but is stronger because the tests considered in Table 1 had very few subscores with length 10 or less. If the length of the subscores is 10, the weighted averages have added value

  - always for level of correlation .7,
  - often for level of correlation between .75 and .85,
  - sometimes for level of correlation .9, and
  - rarely for level of correlation .95.

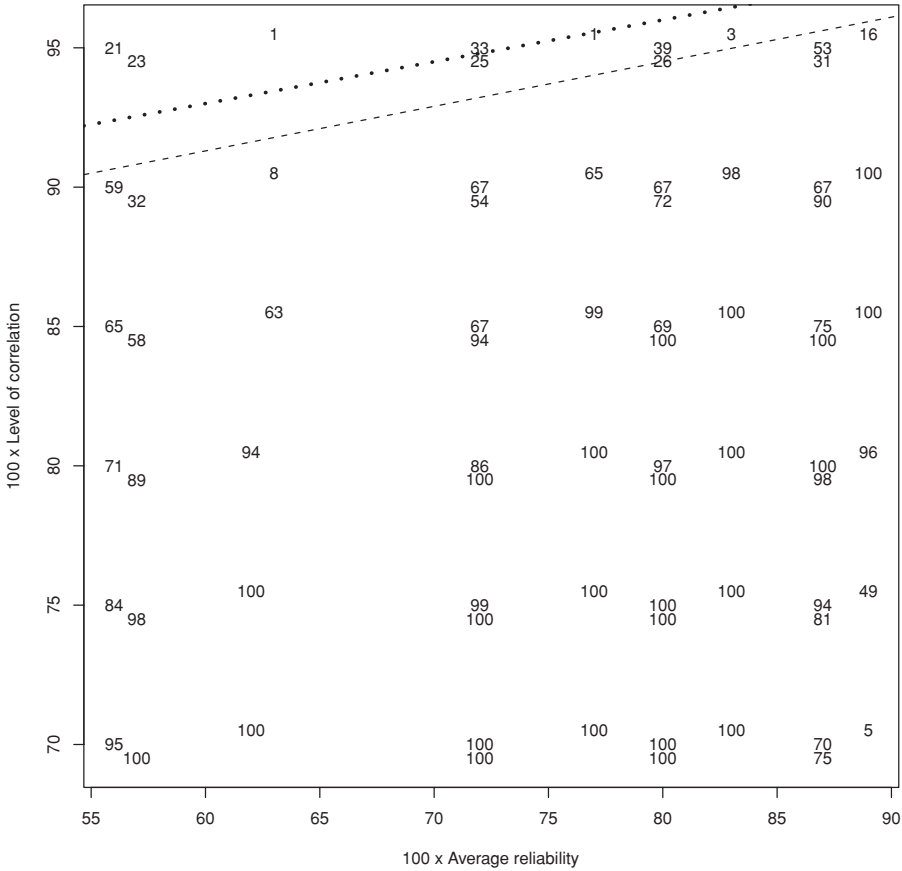

Figure 7. The overall percent of weighted averages that have added value for different average subscore reliability and level of correlation for simulated data with sample size 1,000.

- If the level of correlation is .9 or higher, subscores rarely have added value. Weighted averages often have added value if the level of correlation is .9, but even they do not have any added value if the level is .95. This finding essentially agrees with the findings from Table 1, but appears to be more general (for example, because of a gap at the top right corner of Figure 1).
- If the average length of the subscores is 20 or larger, whether subscores have added value depends on the level of correlation. For example, for length 20, subscores have added value more than 50% of the time if the level of correlation is less than or equal to .75, while, for length 50, they have added value more than 50% of the time if the level of correlation is less than or equal to .85. Thus, there is an interaction between the length of the subscores and the level of correlation.
- The dotted and dashed lines in Figures 5 to 7 agree quite closely, which indicates that the conclusions are roughly similar from operational data and simulated data regarding when a subscore has added value. While the results from operational data have the advantage that they correspond to real data, the results

from the simulated data have the advantage that they are based on several data sets and are more stable than those for the real data.

- The table and the figures show that it is not straightforward to have subscores that have added value. The subscores have to consist of several items (at least 20) and be sufficiently distinct from each other (with disattenuated correlations less than .85) to have any hope of having added value. On the other hand, it is much easier to have weighted averages that have added value. In Figure 7, which shows results for weighted averages, most of the percentages are higher than 50.
- The average disattenuated correlation (r_d in Table 2) among the subscores is always very close to the level of correlation (among the components of $\boldsymbol{\theta}$) for any simulation case.
- The number of subscores does not affect the percentage of cases when the subscores have added value, but the values of reliability, etc. in Table 2 often change as this number changes.
- The PRMSEs of the augmented subscores (Wainer et al., 2001) are almost always very close to those of the weighted averages (Haberman, 2008). The percent of times when the PRMSE of an augmented subscore exceeds that of the corresponding weighted average by at least .01 is 0 for 33 out of the 48 simulation cases shown in Table 2 and is at most 12 (for the case with four subscores, 50 items per subscore, and correlation = .90). Hence the results for the augmented subscore are not shown in Table 2.
- As the level of correlation increases, $PRMSE_{sx}$ becomes closer to the total test reliability (because the weighted average becomes closer to the total test score).

The results discussed above correspond to sample size of 1,000. The results were very similar for other sample sizes (and hence are not shown) with the only exception that for sample size of 100, the difference between the PRMSE of the augmented subscores (Wainer et al., 2001) and the PRMSE of the weighted averages (Haberman, 2008) increases slowly as the number of subscores increases and the length of subscores decrease; for example, the percent of times when the PRMSE of an augmented subscore exceeds that of the corresponding weighted average by at least .01 is more than 50 for 12 out of the 24 simulation cases for the four-subscore case (with six of these cases belonging to length = 10). However the average difference of these two PRMSEs is never more than .02.

Conclusions

Testing program staff interested in reporting subscores often would like to know the properties their subscores should possess in order for them to have added value. Especially, they would like to know more about how reliable and how distinct the subscores should be in order for them to have added value. This article is an attempt to some provide guidance to developing these testing programs.

In this article I first summarized relevant findings from analyses of operational data sets using a table and easily understandable graphical plots. These findings provide some guidance about when subscores can be expected to have added value, but were not conclusive because of too many confounding factors and the small number

of data sets analyzed. Hence, to augment the findings from the operational data, a detailed and realistic simulation study was performed to examine when subscores can be expected to have added value.

There were several interesting findings from the combination of results from the operational and simulated data that promise to be useful to testing program staff interested in reporting subscores. The most important finding is that it is not easy to have subscores that have added value. Based on the results here, the subscores have to consist of at least about 20 items and have to be sufficiently distinct from each other to have any hope of having added value. Several practitioners believe that subscores consisting of a few items may have added value if they are sufficiently distinct from each other. However, the results in this study provide evidence that is contrary to that belief. Subscores with 10 items[13] were not of any added value even for a realistically extreme (low) disattenuated correlation of .7. The practical implication of this finding is that the test developers have to work hard (to make the subtests long and distinct) if they want subscores that have added value.

Weighted averages (Haberman, 2008), on the other hand, were found to have added value more often. For the most part, they had added value as long as the disattenuated correlation between the subscores was less than .95. Even for subtest length of 10, the weighted averages were primarily found to have added value when the disattenuated correlation was .85 or less. This finding should come as good news to testing companies. Weighted averages may be difficult to explain to the general public, who may not like the idea that, for example, a reported reading subscore is based not only on the observed reading subscore, but also on the observed writing subscore. However, this difficulty is more than compensated by the higher PRMSE (that is, more precision) of the weighted average. Note that if a test has only a few very short and distinct subtests, a weighted average may have added value, but should not be reported because its PRMSE, although substantially larger than $PRMSE_s$ and $PRMSE_x$, may still not be adequately high.

The several figures in this article summarize the results in an easily understandable manner and may be used to provide guidance to testing companies. For example, if subscores are to be reported on a testing program, and only 20 questions per subscore can be afforded, Figures 1 and 5 indicate that the average disattenuated correlations between the subscores should be less than .80 (approximately). The figures should be used with caution, however. It is possible to find a unique test for which these figures do not provide accurate guidance. It will be a wise strategy to compute PRMSE's for each test data set before reporting subscores (even after the use of the above-mentioned figures to construct the test).

The usual limitations of simulation studies apply to the results reported in Table 2 and Figures 4 to 7. However, the results of the simulation study essentially agree with those in Table 1, which is based on analysis of operational data; this fact makes the results of the simulation study trustworthy. In addition, the simulations used item parameters estimated from operational data to generate the simulated data sets to make them more realistic. Haberman et al. (2008) found that MIRT models fit operational data better than a univariate IRT model and provide a reasonably good fit to operational data sets. So the data simulated from a MIRT model in this study can

be expected to retain the important features of the operational data reasonably well. In reality, model misfit often occurs. In addition to the simulations reported in this article, limited simulations were performed under different extents of model misfit. For example, some data sets were simulated under the assumption that some items do not follow the form given by Equation 1, but, instead have item response functions of the so called "bad items" in Sinharay (2006). The results for such data did not differ much from those reported in Table 2.

There are several related issues that can be examined in further research. The simulations considered only dichotomous items. It is possible to simulate data with polytomous items. It may be worthwhile to simulate data that mimic those from tests other than the three considered in this article. This article considers subscores that do not share common items (that is the most common phenomenon in practice; all except one of the tests shown in Table 1 deal with such subscores). It is possible to analyze data from tests with subscores that share common items and perform simulations to emulate data from such tests. It is anticipated, however, that the requirements for subscores that share common items to have added value will be even stricter; for example, the information contained in 20 items contributing to such a subscore is probably less than that in 20 items contributing to a subscore that does not share common items. Augmented subscores performed very similar to weighted averages in the simulations here; it may be worthwhile to try to find cases where there will be a difference in performance of these two. The most likely candidate for such a case will be a test of a type not considered in Table 1 or, based on the above-mentioned results for simulated data with sample size of 100, a test that has several subscores and small sample size. One could consider other methods, for example the method of fitting beta-binomial models to the observed subscore distributions (Harris and Hanson, 1991) or factor analysis, to determine when a subscore has added value. However, the method of Harris and Hanson involves significance testing with a χ^2 statistic whose null distribution is not well established (p. 5), and factor analysis involves several issues such as determining whether to analyze at the item level or at the item parcel level, determining whether to use exploratory or confirmatory factor analysis, and determining which test statistics to use to find the number of factors in the data, which complicate the process of determining whether a subscore has added value. The method to determine if subscores based on MIRT models have added value (Haberman, & Sinharay, in press) could be another possible candidate (though the method provided results similar to the CTT-based method considered in this article). The CTT-based method (Haberman, 2008) was chosen in this article because the method is conceptually and computationally simple, provides a simple and unambiguous rule as to when a subscore has added value, and has a strong theoretical basis.

Appendix: Results from a Simulation Where the Number of Items Contributing to the Subscores Is Different

Table A1 shows the results from a simulation whose design is similar to that in the Simulation Study section above, except that only the case with three subscores

Table A1
Summary of the Simulated Data for Sample Size 1,000

Level of Diffe-scores	Average Length of the Subscores															
	10 Items Correlation				20 Items Correlation				30 Items Correlation				50 Items Correlation			
	.70	.80	.90	.95	.70	.80	.90	.95	.70	.80	.90	.95	.70	.80	.90	.95
0																
α_{tot}	75	77	78	79	86	87	88	88	90	91	92	92	94	94	95	95
PR_s	56	56	56	56	72	72	72	72	80	80	80	80	87	87	87	87
r	39	45	51	54	50	58	65	69	56	64	71	76	61	69	78	82
r_d	70	80	91	96	70	80	90	95	70	80	90	95	70	80	90	95
PR_x	61	67	74	77	69	76	82	85	72	79	86	89	75	82	89	92
PR_{sx}	67	70	75	77	78	81	84	86	83	85	88	90	89	90	92	93
% sub	41	03	00	00	67	46	01	00	76	67	26	00	100	67	65	04
% wtd	95	71	59	21	100	86	67	33	100	97	67	39	75	100	67	53
20																
α_{tot}	76	77	79	79	86	87	88	89	90	91	92	92	94	95	95	95
PR_s	56	56	56	56	72	72	72	72	79	79	79	79	86	86	86	86
r	39	44	50	53	50	57	64	68	55	63	71	75	60	69	77	82
r_d	70	80	91	95	70	80	90	95	70	80	90	95	70	80	90	95
PR_x	61	67	74	77	69	75	82	86	72	79	86	89	75	82	89	92
PR_{sx}	67	71	75	78	78	81	85	86	83	85	89	90	89	90	92	93
% sub	48	07	00	00	67	56	07	00	71	67	37	00	100	67	67	12
% wtd	95	74	62	20	100	89	67	35	100	97	67	46	66	74	67	62
40																
α_{tot}	76	78	79	80	87	88	88	89	91	91	92	92	94	95	95	95
PR_s	55	55	55	55	70	70	70	70	78	78	78	78	85	85	85	85
r	37	42	48	51	48	55	62	66	54	61	59	73	59	68	76	81
r_d	70	80	91	96	70	80	90	95	70	80	90	95	70	80	90	95
PR_x	61	67	74	78	68	75	82	86	71	78	86	89	74	81	88	92
PR_{sx}	67	71	76	78	78	81	85	87	83	85	89	90	88	90	92	93
% sub	55	16	00	00	67	64	13	00	68	67	53	00	94	67	67	21
% wtd	93	76	70	24	100	90	67	35	76	97	67	56	58	67	57	65

Note. α_{tot} denotes average reliability of the total score; PR_s denotes 100 × the average reliability of the subscores; r denotes 100 × the average correlation between the subscores; r_d denotes 100 × the average disattenuated correlation between the subscores; PR_x denotes 100 × average $PRMSE_x$; PR_{sx} denotes 100 × average $PRMSE_{sx}$; PR_{aug} denotes 100 × average $PRMSE_{aug}$; % sub denotes the overall percent of subscores that have added value; % wtd denotes the overall percent of weighted averages that have added value.

was considered here and the number of items contributing to the two subscores was different. Two levels of the difference were considered: 20% and 40%. A 20% level of difference means that if the length of the subscores (actually, it is now an average length of the subscores) is, for example, 30, the number of items contributing to the three subscores are 24 (=30 – 20% of 30), 36 (=30 + 20% of 30), and 30. For comparison purposes, the results for 0% (the case when the three subscores have

equal numbers of items) is also given; these are the results for three subscores from Table 2.

It is clear from Table A1 that the level of difference does not affect the results much. The results (especially the percent of times when the subscores or weighted averages have added value) are almost the same when the three subscores have 30 items each, or have 24, 30, and 36 items, respectively, or 18, 30, and 42 items, respectively. The percent of times when the PRMSE of an augmented subscore (Wainer et al., 2001) exceeds that of the corresponding weighted average by at least .01 ranges between only 0 and 8 for the simulation cases shown in Table A1, although, intuitively, many would expect the augmented subscore to perform better than the weighted average in these cases of unequally long subtests.

Acknowledgments

The author thanks Shelby Haberman, Gautam Puhan, Mark Reckase, Helena Jia, Per-Erik Lyren, Jonathan Templin, and Terry Ackerman. The author gratefully acknowledges the help of Denise Schmutte with proofreading. Any opinions expressed in this publication are those of the author and not necessarily of Educational Testing Service.

Notes

[1]A larger PRMSE is equivalent to a smaller mean squared error in estimating the true subscore and hence is desirable.

[2]Changing these estimated reliabilities slightly did not affect the conclusions.

[3]Note that the examination has been changed significantly since 2002.

[4]The fact that summary statistics published in another paper can be used to perform all the required calculations demonstrates the simplicity of the method of Haberman (2008).

[5]Note that Ackerman and Shu (2009) and Henson et al. (2009) actually considered diagnostic scores based on IRT models and not subscores per se.

[6]Where the disattenuated correlation between two subscores is equal to the simple correlations between them divided by the square root of the product of the reliabilities of the two subscores.

[7]Note that although the table reports the averages to summarize a lot of information in a compendious manner, for some of these tests, the lengths, reliabilities, and correlations of the subscores are substantially unequal.

[8]Changing .01 to other small values such as .02 or .03 did not affect the conclusions much.

[9]In other words, there is an interaction between average length (or average reliability) and average disattenuated correlation.

[10]There was no need to use correlations from operational data sets for the 2-subscore case, for which the only correlation was set equal to the level of correlation for a simulation case.

[11]When the level of correlation is .95, some of the correlations in $\mathcal{C}$ were changed before this calculation to ensure that Σ is positive definite.

[12]The standard errors of relevant quantities were examined to make sure that the choice of $R = 100$ produced sufficiently precise results.

[13]In practice, it is not difficult to find operationally reported subscores that are based on 10 or fewer items.

References

Ackerman, T., & Shu, Z. (2009, April). *Using confirmatory MIRT modeling to provide diagnostic information in large scale assessment*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*, 191–204.

Bock, R. D., & Petersen, A. C. (1975). A multivariate correction for attenuation. *Biometrika*, *62*, 673–678.

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

Haberman, S. J., & Sinharay, S. (in press). Reporting of subscores using multidimensional item response theory. *Psychometrika*.

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, *62*, 79–95.

Haberman, S. J., von Davier, M., & Lee, Y. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous distributions* (ETS Research Report No. RR-08-45). Princeton, NJ: Educational Testing Service.

Haladyna, S. J., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation and the Health Professions*, *24*, 349–368.

Harris, D. J., & Hanson, B. A. (1991, April). *Methods of examining the usefulness of subscores*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Henson, R., Templin, J., & Irwin, P. (2009, April). *Ancillary random effects: A way to obtain diagnostic information from existing large scale tests*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Kamata, A., Turhan, A., & Darandari, E. (2003, April). *Estimating reliability for multidimensional composite score scale scores*. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Ling, G. (2009, April). *Why the major field (business) test does not report subscores of individual test-takers—reliability and construct validity evidence*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

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

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2008). *Comparison of subscores based on classical test theory methods* (ETS Research Report No. RR-08-54). Princeton, NJ: Educational Testing Service.

Reckase, M. D. (2007). Multidimensional item response theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 607–642). Amsterdam, The Netherlands: North-Holland.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*, *59*, 429–449.

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

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, *26*(4), 21–28.

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education, 23*, 63–86.

Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education*, *17*, 89–112.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., & Nelson, L. (2001). Augmented scores—"Borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Hillsdale, NJ: Lawrence Erlbaum.

Author

SANDIP SINHARAY is a Senior Research Scientist, Educational Testing Service, MS 12T, Rosedale Road, Princeton NJ 08541; ssinharay@ets.org. His primary research interests include item response theory, equating, diagnostic score reporting, Bayesian methods, and application of statistics to education.