Research Report

When Can Subscores Be Expected To Have Added Value? Results From Operational and Simulated Data

Sandip Sinharay

August 2010

ETS RR-10-16

When Can Subscores Be Expected To Have Added Value? Results From Operational and Simulated Data

Sandip Sinharay
ETS, Princeton, New Jersey

August 2010

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Matthias von Davier

Technical Reviewers: Andreas Oranje and Isaac Bejar

Abstract

Recently, there has been an increasing level of interest in using subscores for their potential diagnostic value. Haberman (2008) suggested a method based on classical test theory to determine whether subscores have added value over total scores. This paper provides a literature review and reports when subscores were found to have added value for several operational data sets. Then this paper provides 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. Augmented subscores (Haberman, 2008; Wainer et al., 2001) were found to have added value more often.

Key words: augmented subscore, mean squared error

Acknowledgments

The author thanks Shelby Haberman, Gautam Puhan, Mark Reckase, Helena Jia, Per-Erik Lyren, Jonathan Templin, and Terry Ackerman for helpful suggestions. The author gratefully acknowledges the help of Denise Schmutte and Kim Fryer with proofreading.

There is an increasing interest in subscores because of their potential diagnostic value. Failing candidates 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 profile of performance for their graduates to better evaluate their training and focus on areas that need instructional improvement (Haladyna & Kramer, 2004).

Despite this apparent usefulness of subscores, certain important factors must be considered before making a decision whether to report subscores at either the individual or institutional level. Standard 5.12 of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) states, "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* (American Educational Research Association et al., 1999) 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 papers by Puhan, Sinharay, Haberman, and Larkin (2008) and Sinharay and Haberman (2008), analyzed data sets from a variety of testing programs. They found that there are only a handful of tests for which subscores have added value.

A question that testing programs 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?" In particular, the testing programs would like to know more about how many items should comprise their subscores 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.

This paper first provides a literature review and summarized results that are relevant to the above mentioned question and were obtained from the analysis of operational data. Then, this paper provided 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 (which is a natural choice when subscores are under consideration). The simulation study uses 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.

Section 1 provides a brief overview of the CTT-based methods of Haberman (2008). Section 2 provides a literature review—it discusses the results from several operational data sets regarding the question of when subscores can be expected to have added value. Section 3 describes the simulation study. Section 4 provides conclusions based on the results from the operational and simulated data sets.

1 Methods From Classical Test Theory

This section describes 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.

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 the true subscore. However, for simulated and operational data, s_{aug} yielded results that are very similar to those for s_{sx}. Hence this paper does not provide any results for s_{aug}. Note that the estimate s_{sx} is a special case of the augmented subscore s_{aug}; s_{sx} places the same weight on all the subscores other than the one of interest instead of weighing them differently. Unless otherwise stated, s_{sx} will be referred to as the augmented subscore in the rest of the paper.

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] This paper will denote the PRMSE for s_s, s_x, and s_{sx} as $PRMSE_s$, $PRMSE_x$, and $PRMSE_{sx}$ respectively. The quantity $PRMSE_s$ can be shown to be exactly equal to the reliability of the subscore. Haberman (2008) recommended the following strategy to decide whether a subscore or an augmented subscore has added value:

- If $PRMSE_s$ is less than $PRMSE_x$, declare that the subscore "does not provide added value over the total score," because the observed total score will provide more accurate diagnostic information (in the form of a lower mean squared error in estimating the true subscore) than the observed subscore 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, declare that an augmented subscore has added value only if $PRMSE_{sx}$ is substantially larger compared to both $PRMSE_s$ and $PRMSE_x$.

If neither the subscore nor the augmented subscore 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. Haberman (2008) and Sinharay et al. (2007) 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 other subscores. The appendix provides more details about the method of Haberman (2008).

2 Review of Results From Operational Data Analysis

Table 2 summarizes the findings from Haberman (2008), Harris and Hanson (1991), Puhan et al. (2008), and Sinharay and Haberman (2008) from operational data sets.

Each row in the table shows, for a test, the number of subscores, average number of items in the subscores, average reliability of the subscores, average correlation among the subscores, average disattenuated correlation,[2,3] the number of subscores that have added value, and the number of augmented subscores that have added value (where the assumption was made that an augmented subscore has added value if the corresponding $PRMSE_{sx}$ is larger than the maximum of $PRMSE_s$ and $PRMSE_x$ by 0.01 or larger[4]).

For SAT® Verbal (the first row of numbers in Table 2), the subscores refer to the critical reading, analogies, and sentence completion scores, percentile scores for which used to be reported to the examinees. For SAT Math, the subscores refer to the scores on four-choice multiple choice questions, five-choice multiple choice questions, and student-produced responses—these were not operationally reported. For SAT (the third row of numbers in Table 2), the subscores actually refer to the SAT Verbal and SAT Math scores. 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 now. For all the other tests considered in Table 2, the subscores refer to operationally reported subscores.

For the P-ACT+ English and Mathematics tests, the numbers shown in Table 2 are from Harris and Hanson (1991), who used three forms each of these tests. The subscore reliabilities were not provided in Harris and Hanson (1991). However, for each form, 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 (as the number of items comprising the subscores is known). For these data, the methods of Haberman (2008) were not applied because of the lack of information. However, Harris and Hanson (1991) concluded that the P-ACT+ subscores do not provide information distinct from the total scores, using an approach that involves fitting of

Table 1
Results From Analysis of Operational Data Sets

Name/nature of the test	No. of sub-scores	Av. length	Av. α	Av. corr.	Average corr. (disatt.)	How many subscores have added value?	How many aug. subs have added value?
SAT Verbal	3	26	0.79	0.74	0.95	None	One
SAT Math	3	20	0.78	0.75	0.97	None	None
SAT	2	69	0.92	0.70	0.76	Both	Two
Praxis	4	25	0.72	0.56	0.78	Two	Four
P-ACT+ English	2	25	0.80	0.76	0.96	None	NA
P-ACT+ Mathematics	2	20	0.80	0.71	0.95	None	NA
DSTP Math (8th grade)	4	19	0.77	0.77	1.00	None	None
CA (for teachers in elementary schools)	4	30	0.74	0.59	0.79	One	Four
CB (for teachers of special ed. programs)	3	19	0.46	0.42	0.96	None	None
CC (for beginning teachers)	4	19	0.38	0.44	1.00	None	None
CD (for teachers of social studies)	6	22	0.63	0.54	0.87	None	Six
CE (for teachers of Spanish)	4	29	0.80	0.65	0.80	One	Two
CF (for principals and school leaders)	4	25	0.48	0.41	0.85	None	Four
CG (for teachers of mathematics)	3	16	0.62	0.59	0.95	None	None
CH (for paraprofessionals)	3	24	0.85	0.76	0.89	None	Three
TA (measures cognitive and technical skills)	7	11	0.42	0.51	1.00	None	None
TB1 (tests mastery of a language)	2	44	0.85	0.77	0.90	One	Two
TB2 (tests mastery of a language)	2	43	0.90	0.68	0.75	Two	Two
TC1 (measures achievement in a discipline)	3	68	0.85	0.76	0.90	One	Three
TC2 (measures achievement in a discipline)	3	67	0.87	0.72	0.82	Two	Three
TD1 (measures school and individual student progress)	4	15	0.70	0.73	0.98	None	No
TD2 (measures school and individual student progress)	6	13	0.70	0.75	1.00	None	No

Note. The reliability is denoted as α. Augmented subscores are denoted as *aug. subs.* The first four tests were discussed by Haberman (2008). The next two tests were discussed in Harris and Hanson (1991). The next, DSTP Math, is discussed in Stone, Ye, Zhu, and Lane (2009). The next eight, denoted CA-CH, are certification tests discussed in Puhan et al. (2008). The next seven, denoted TA, TB1, ... TD2, were discussed in Sinharay and Haberman (2008).

beta-binomial models to the observed subscore distributions. Hence, it was assumed that none of the P-ACT+ subscores have added value for any of these forms.

The seventh row of Table 2 shows 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 et al. (2009), who reported, using the exploratory factor analysis method, the presence of only one factor in the data set. The four subscores, which 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. (2009) were used to compute the PRMSEs required in the method of Haberman (2008).[5]

For some tests such as CC, average disattenuated correlation was larger than 1—they were set to 1.00. Most of the tests had only multiple choice items. Some tests such as CF had constructed response (CR) items. For a subscore with CR items, the *length* refers to the total number of score categories minus the number of items (for example, for a subscore with 4 items, each with score categories 0, 1, and 2, the length is $4 \times 3 - 4 = 8$).

Figures 1 to 3 show, for the operational data sets, the percentage of subscores (Figures 1 and 2) or augmented subscores (Figure 3) that had added value. In each of 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. The three figures plot, for each row listed in Table 2, a number that is the same as the percentage of subscores (or augmented subscores) 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 2, the third row shows that the SAT had average length 69, average disattenuated correlation 0.76, 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) at the bottom right corner.

Table 2 and Figures 1 to 3 show that there are few subscores that have added value. It is also worth noting that the disattenuated correlation is 0.95 or larger for many of the tests. In general, subscores with 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 with a large number of items have added value. For example, for the test TC1, which has an average subscore length of 68, only one of three subscores has added value. Tests

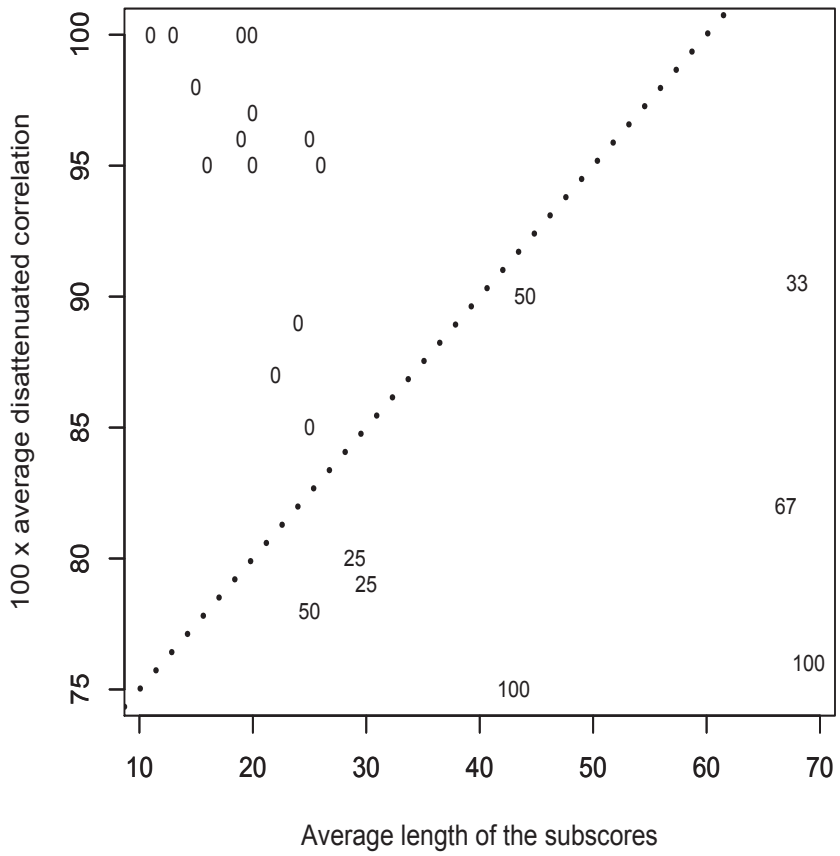

***Figure* 1. The percentage of subscores that had added value for different average subscore number of items and average disattenuated correlation for the operational data.**

with low average disattenuated correlation tended to have subscores with added value. However, for the test CF, the average disattenuated correlation is 0.85, and none of the subscores have added value, while, for the test TB1, the average disattenuated correlation is 0.90, but one of the two subscores has added value.

Often, the percentage of subscores with a specific average length (or average reliability) that have added value depends on the average disattenuated correlations.[6] Hence, each of the figures shows a bold dotted line roughly dividing the plot into two regions in which the percentage is low (zero) and high (positive). Note that this line is arbitrary and was drawn after a visual

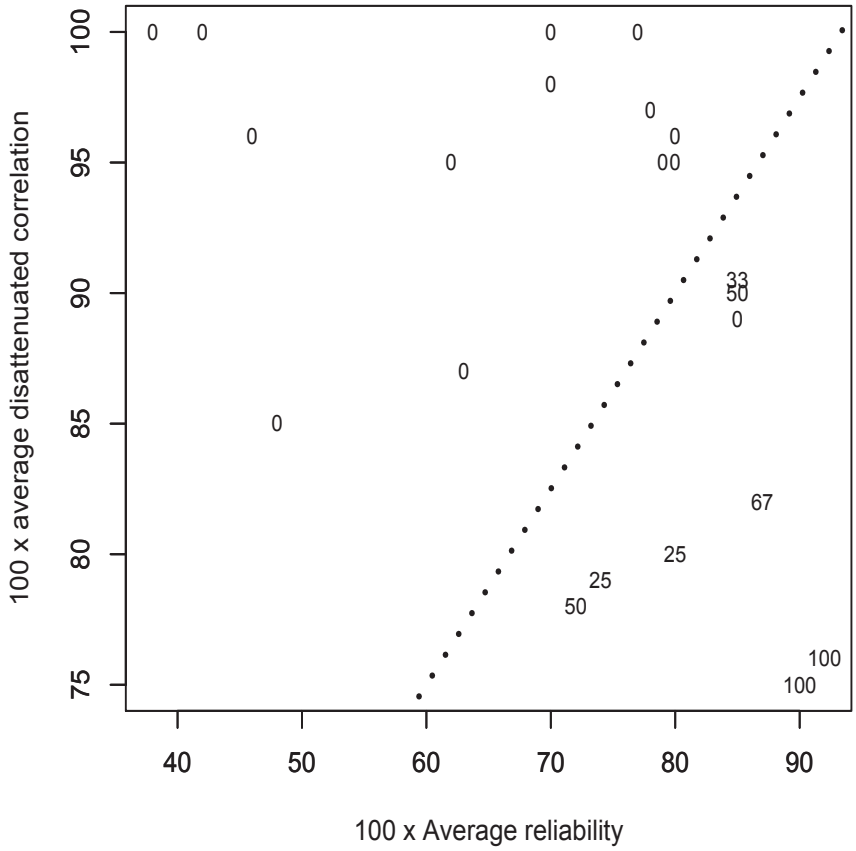

***Figure* 2. The percentage of subscores that had added value for different average subscore reliability and average disattenuated correlation for the operational data.**

examination of the points in the plot and not created 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 augmented subscores have added value for many of the operational data sets and that augmented subscores are much more likely to have added value compared to the subscores themselves.

However, Table 2 and the Figures 1 to 3 were based on only a few data sets, so that they are not expected to be very stable (for example, if one obtains another collection of data sets, a figure like Figure 1 for those data sets may look substantially different). In addition, few of these tests

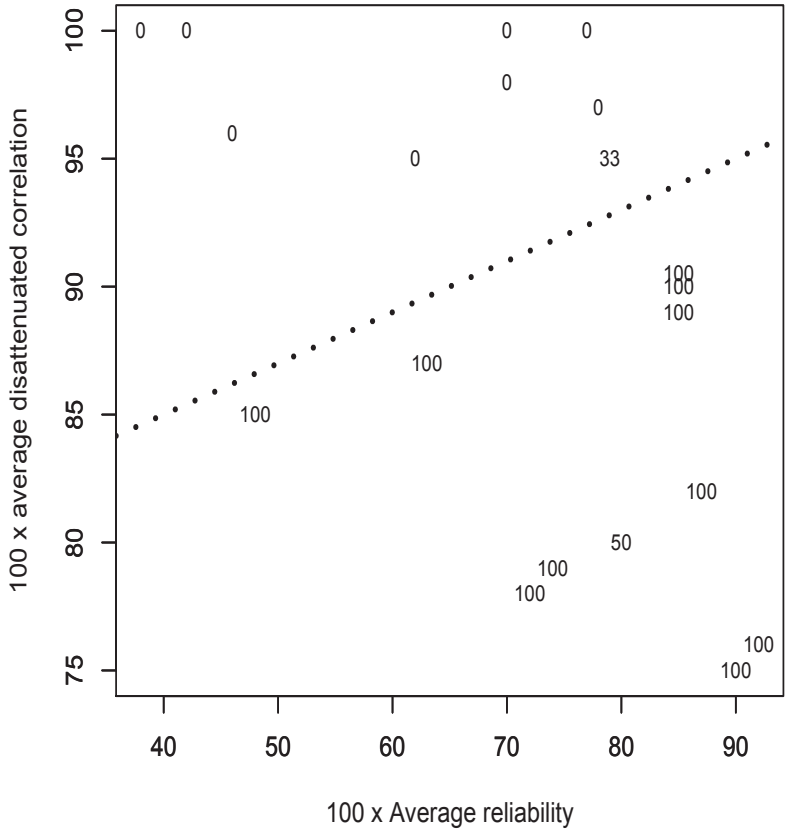

***Figure* 3. The percentage of augmented subscores that had added value for different average subscore reliability and average disattenuated correlation for the operational data.**

had subscores that have added value (for example, there are only three points in Figure 1 with the percentage of subscores that have added value larger than 50). Besides, there are other extraneous factors, such as the nature of the test that affect the 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 points with average length around 50, with none of them having average disattenuated correlation between 0.75 and 0.90. A decision based on Figures 1 to 3 for a data set whose correponding point falls in one of these gaps will require extrapolation and may not be correct. Hence, the results from Table 2 can provide some guidance to testing programs, but cannot be

used to give precise advise as to how long or how distinct their subscores have to be in order to have added value.

Hence there was the need to perform a simulation study, where it is easier to control different factors and study the effect of the factors of interest. The simulation study is discussed in the next section.

3 Simulation Study

The MIRT Model

This section discusses 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

$$(1 + e^{-(a_{1i}\theta_1 + a_{2i}\theta_2 + \cdots + a_{Ki}\theta_K - b_i)})^{-1}, \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)' \sim \mathcal{N}_K\left((0, 0, \ldots, 0)', \Sigma\right), \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ is the K-dimensional ability parameter for an examinee, b_i is the difficulty parameter for item i, $a_{1i}, a_{2i}, \ldots a_{Ki}$ are the K slope parameters for item i (a_{ki} denotes the loading of item i on the k-th dimension), and $\mathcal{N}_K$ denotes the density of the normal distribution with K dimensions. Each component of $\boldsymbol{\theta}$ corresponds to a subscore. The diagonals of Σ are set to 1 to ensure identifiability of the model parameters. For any item i, only one among the slope parameters $a_{1i}, a_{2i}, \ldots, a_{Ki}$ is assumed to be non-zero, 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 (that is equivalent to assuming that the subscores do not share common items).

Simulation Design

Generating item parameters. One data set was obtained from each of three tests that operationally report subscores or section scores. The first test, which is a test in English, reports two section scores, each of which is based on 100 multiple choice items. The second test, which is the test TC2 in Table 2, reports three subscores, which consist of 66-67 multiple choice items (Sinharay & Haberman, 2008). The third test, which is the test CA in Table 2, reports four subscores—language arts/reading, mathematics, social studies, and science—each based on 30 items (Puhan et al., 2008).

The model given by Equation 1 was fitted using the stabilized Newton-Raphson algorithm (Haberman et al., 2008) for each of the three data sets to obtain estimated item parameters. 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 later used as generating item parameters in the simulation study. For each test, a bivariate normal distribution $\mathcal{B}_k$ was fitted to the log-slope and difficulty parameter estimates of the items belonging to k-th subscore, $k = 1, 2, \ldots K$. The generating item parameters for the k-th subscore in the test were randomly drawn from $\mathcal{B}_k$.

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 (which 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 3 subscores) was used to simulate data that have 3 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. For this reason, quotes are put around "number of subscores."

- Length of the subscores. This paper 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, this paper assumed that the different subscores for a given test have the same length.

- Level of correlation (ρ) among the components of $\boldsymbol{\theta}$. This paper used six levels: 0.70, 0.75, 0.80, 0.85, 0.90, and 0.95. If the correlation level for a simulation case is ρ, it was assumed, to simulate the data sets, that all the off-diagonal elements of Σ (which denote the correlations between the components of $\boldsymbol{\theta}$) in Equation 1 are equal to ρ. Note that the correlations among the components of $\boldsymbol{\theta}$ are similar to the disattenuated correlations between the subscores. Hence, from Table 2, the choice of these levels of this correlation (especially, the lowest of them) is reasonable.

- Sample size N. This paper 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 of 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 distributions $\mathcal{B}_k$s), and then R = 100 replications[7] 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 $\mathcal{N}_K\left((0,0,\ldots,0)',\Sigma\right)$, where the diagonals of Σ are 1 and the off-diagonals are the same as the correlation level for the simulation case.

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.

Simulation Results

Table 3 shows results for sample size of 1,000. The table shows results for four (out of six) values of the level of correlation. The results were very similar for other sample sizes and hence are not shown.

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.

Table 2

Summary of the Simulated Data for Sample Size 1,000

No. of sub-scores	Length of the subscores: 10, Correlation 70	80	90	95	20, Correlation 70	80	90	95	30, Correlation 70	80	90	95	50, Correlation 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
% aug	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	72	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	75	82	85	72	79	86	89	75	82	88	92
PR_{sx}	66	70	74	77	78	80	84	86	83	85	87	90	89	89	91	93
% sub	19	00	00	00	85	19	00	00	100	66	00	00	100	100	20	00
% aug	100	92	38	10	100	100	74	13	100	100	91	16	97	100	100	39
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	86	87	87
r	40	46	52	55	50	58	65	69	55	63	72	76	60	69	78	82
r_d	70	81	91	96	69	80	90	95	69	80	90	95	69	79	90	95
PR_x	62	70	78	82	69	76	84	88	72	79	87	90	73	81	89	93
PR_{sx}	69	73	79	82	79	81	85	88	83	85	89	91	89	90	92	94
% sub	22	01	00	00	81	23	00	00	99	67	01	00	100	100	26	00
% aug	100	92	41	08	100	100	74	13	100	100	91	24	83	100	100	58

3. 100 × 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 × the average disattenuated correlation between the subscores. This is denoted as r_d in the tables.

5. 100 × 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 × average $PRMSE_{sx}$ (denoted as PR_{sx}).

7. Overall percentage of subscores that have added value (denoted as % sub). This is the overall percentage of cases (out of a total of R × K, where K is the number of subscores) when $PRMSE_s$ is larger than $PRMSE_x$.

8. Overall percentage of augmented subscores that have added value (denoted as % aug). This is the overall percentage of cases (out of a total of R × K) when $PRMSE_{sx}$ is larger than the maximum of $PRMSE_s$ and $PRMSE_x$ by 0.01 or more.

Figures 4 to 8 show, for simulated data with sample size of 1,000, the overall percentage of subscores or augmented subscores that have added value. These plots (unlike Table 3) show results for all the six levels of correlation from 0.70 to 0.95. Figures 4 and 7 are three-dimensional scatterplots showing the overall percentage of subscores (Figure 4) or augmented subscores (Figure 7) that have added value (shown along the Z-axis using a vertical line) versus subscore length and level of correlation. Figures 5, 6, and 8 are like Figure 1 and show, for each combination of subscore length (or average reliability) and level of correlation, a number showing the percentage. Figures 5, 6, and 8 show dashed lines roughly dividing the plot into two regions in which the percentage is low (less than 10%, roughly) and high (more than 10%). 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 8, there are up to three points (corresponding to the three values of "number of subscores") for each (x, y) coordinate. To avoid overlapping points in the figures, one of the three points was moved slightly up and another slightly down.

Examination of Table 3 and Figures 4 to 8 leads to the following conclusions:

- Overall, the percentage of times when the subscores have added value increases with an increase in their lengths (or reliability) and with a decrease in the 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 3, the percentage of times when the subscores have added value is less than 1

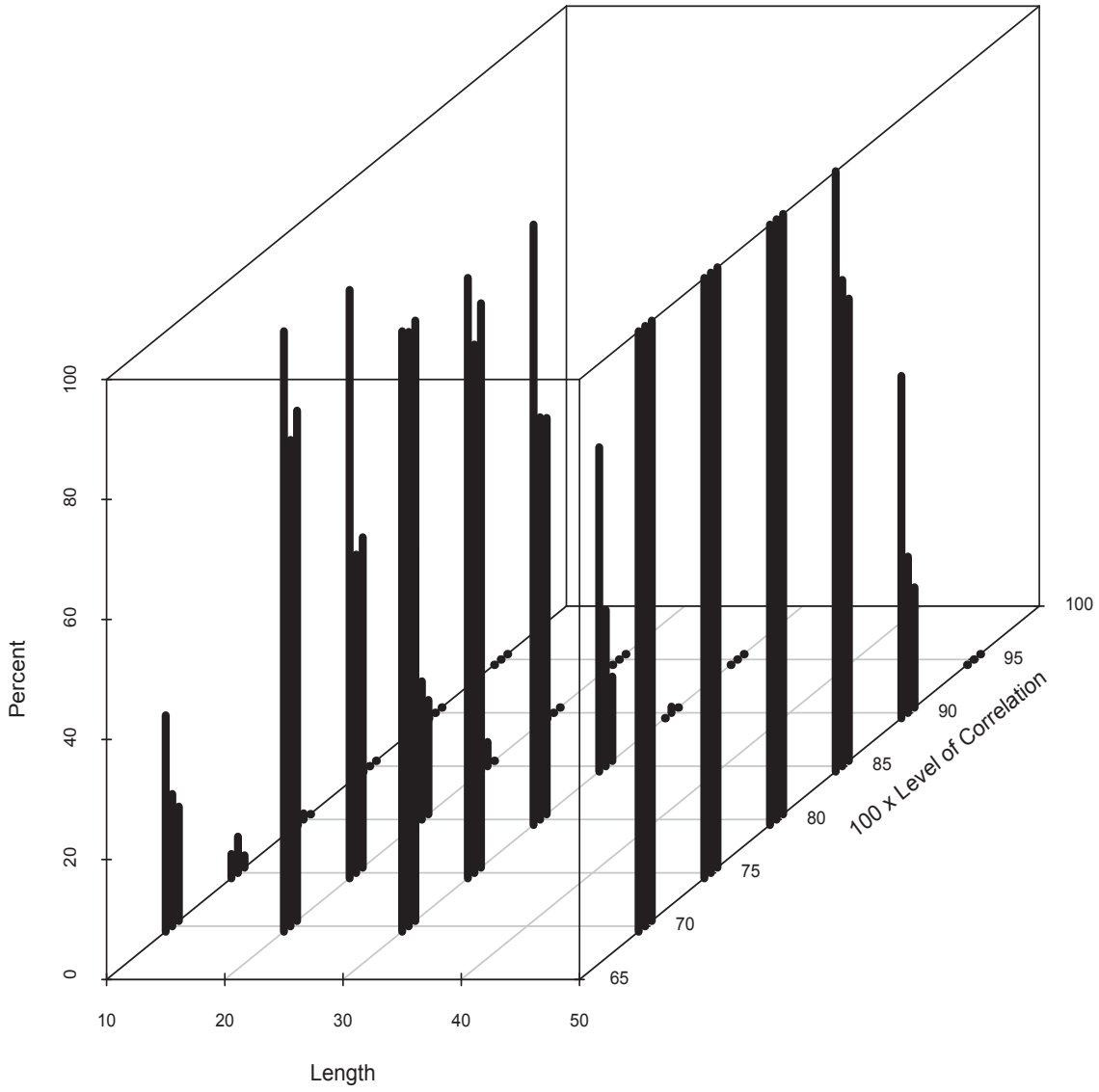

***Figure* 4. A three-dimensional scatterplot showing the overall percentage of subscores that had added value for sample size 1,000.**

in nine cases and has a significant non-zero value only when the level of correlation is only 0.70, which, according to Table 2, is rare in practice. This conclusion supports the findings of Table 2 in which none of the subscores with few items had any added value, but is stronger because the tests considered in Table 2 had very few subscores with length 10 or less. If the

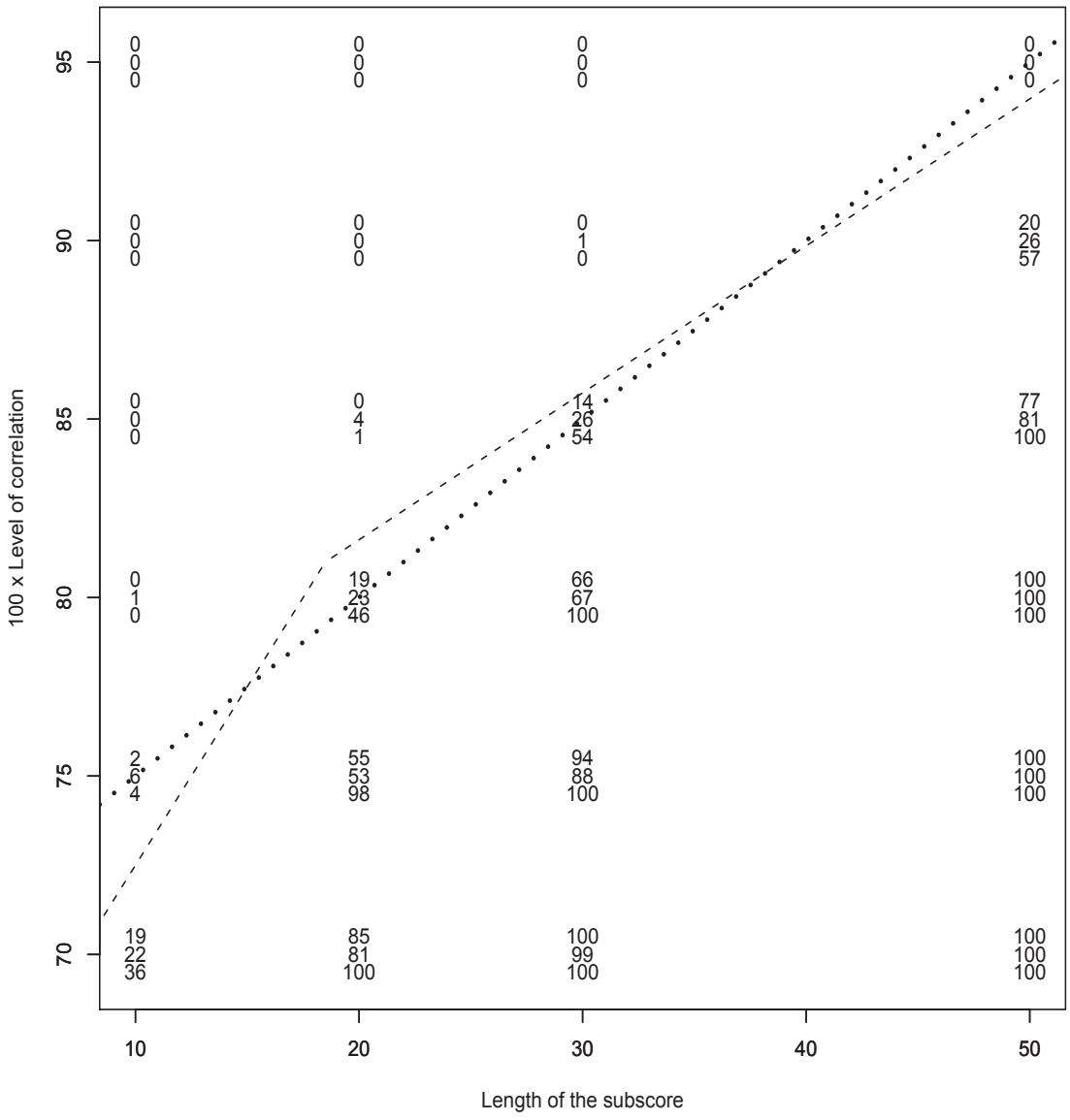

***Figure* 5. The overall percentage of subscores that had added value for sample size 1,000 versus subscore length and level of correlation.**

length of the subscores is 10, the augmented subscores have added value

- always for level of correlation 0.7,
- often for level of correlation between 0.75 and 0.85,
- sometimes for level of correlation 0.9, and

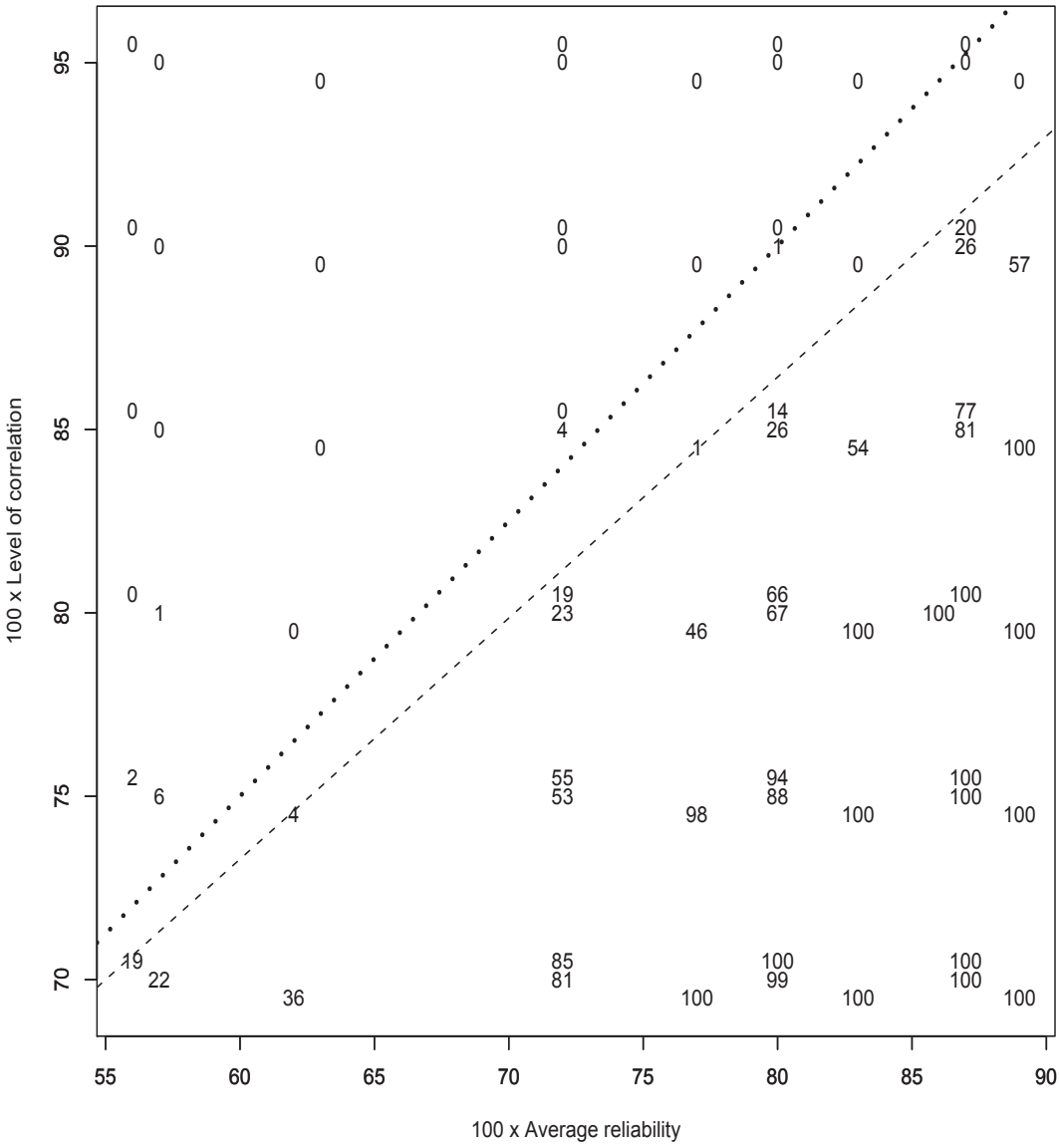

***Figure* 6. The overall percentage of subscores that had added value for sample size 1,000 versus average subscore reliability and level of correlation.**

– rarely for level of correlation 0.95.

- If the level of correlation is 0.9 or higher, subscores rarely have added value. Augmented subscores often have added value if the level of correlation is 0.9, but even they do not have any added value if the level is 0.95. This finding mostly agrees with the findings from

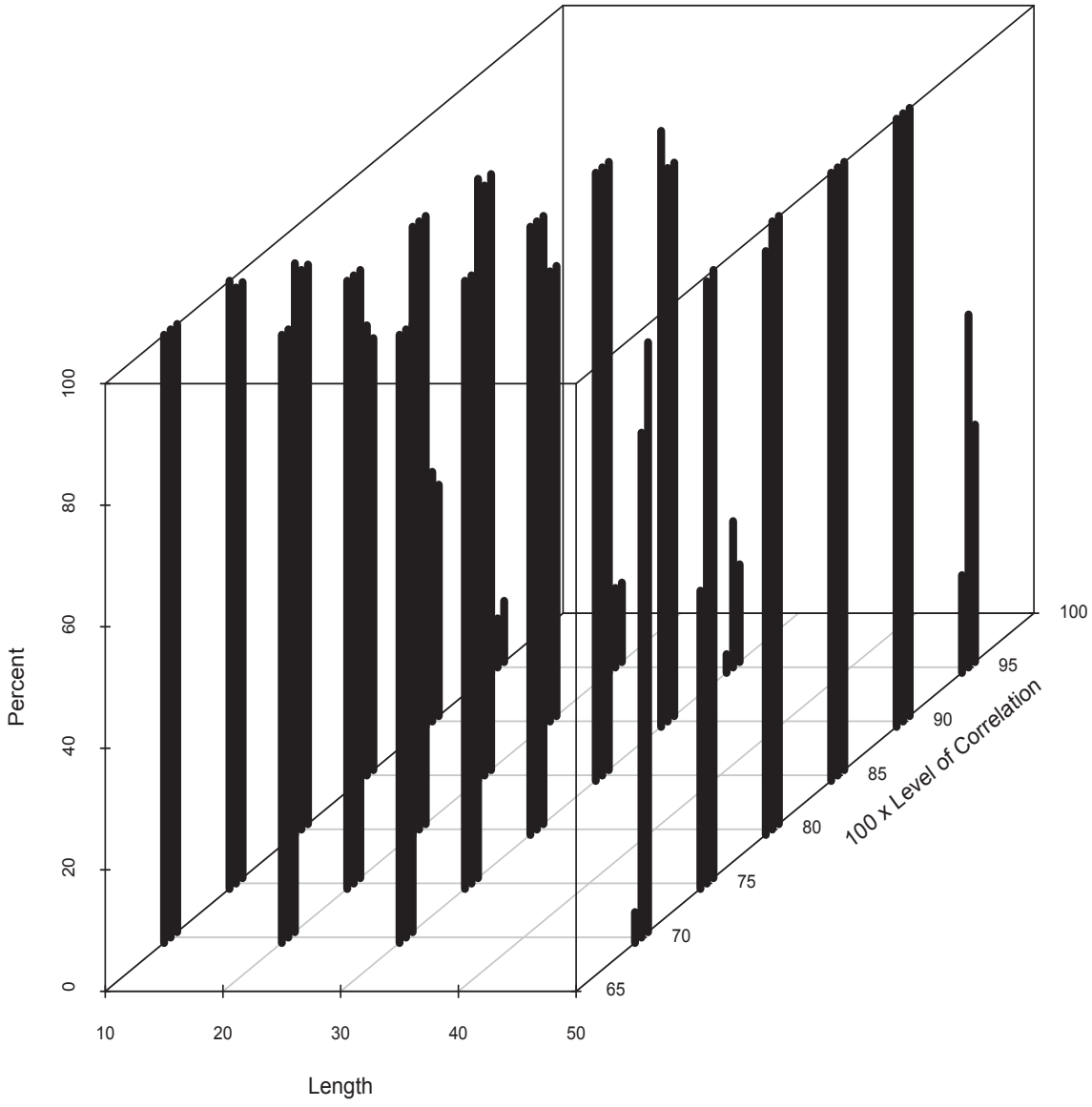

***Figure* 7. A three-dimensional scatterplot showing the overall percentage of augmented subscores that had added value for sample size 1,000 versus subscore length and level of correlation.**

Table 2, but seems 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

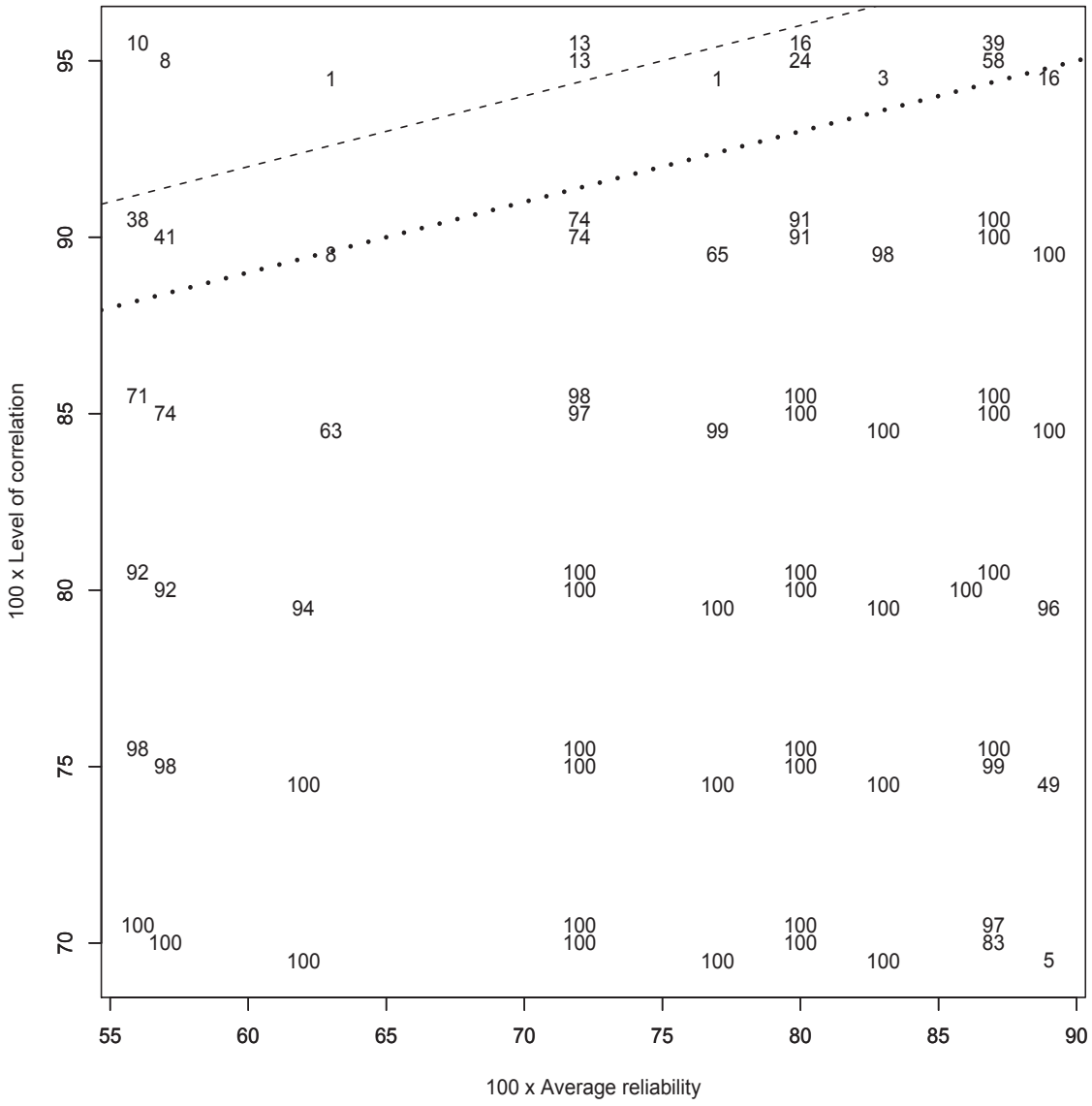

***Figure* 8. The overall percentage of augmented subscores that had added value for sample size 1,000 versus average subscore reliability and level of correlation.**

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 0.75, 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 0.85. Thus, there is an interaction between the length of the subscores and

the level of correlation.

- The dotted and dashed lines in Figures 5, 6, and 8 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 be long (consisting of at least 20 items) and sufficiently distinct from each other (with disattenuated correlations less than 0.85) to have any hope of having added value. On the other hand, it is much easier to have augmented subscores that have added value. In Figure 8, most of the percentages are higher than 50.

- The average disattenuated correlation (r_d in Table 3) 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 3 often change as this number changes.

- The PRMSE of the augmented subscores suggested by Wainer et al. (2001) is almost always very close to those suggested by Haberman (2008). The difference between them was almost always less than 0.01. Hence the results for the augmentation of Wainer et al. (2001) are not shown.

- As the level of correlation increases, $PRMSE_{sx}$ becomes closer to the total test reliability (because the augmented score becomes closer to the total test score).

4 Conclusions

Testing programs interested in reporting subscores often would like to know the properties their subscores should possess in order for them to have added value. In particular, they would like to know more about how long and how distinct the subscores should be in order for them to have added value. This paper is an attempt to provide guidance to these testing programs.

This paper first summarizes 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 obtain more information on the research problem, this paper performed a detailed and realistic simulation study 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 programs interested in reporting subscores. The most important finding is that it is not easy to have subscores that have added value. Based on our results, the subscores have to consist of at least 20 items and have to be sufficiently distinct from each other to have any hope of having added value. Several practitioners believe that short subscores may have added value if they are sufficiently distinct from each other. However, the results in this study provide evidence that are contrary to that belief. Subscores composed of 10 items[8] were not of any added value even for a realistically extreme (low) disattenuated correlation of 0.7. The practical implication of this finding is that the test developers have to work hard (to make the subscores long and/or distinct) if they want subscores that have added value.

Augmented subscores, on the other hand, were found to have added value more often. They mostly had added value as long as the disattenuated correlation between the subscores is less than 0.95. Even for a test length of 10, the augmented subscores were found to have added value when the disattenuated correlation was 0.85 or less. This finding should come as a good news to testing companies. Augmented subscores 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 subscores. However, this difficulty is more than compensated by the higher PRMSE (that is, greater precision) of the augmented subscore. Note that if a test has only a few short subscores, an augmented subscore may have added value, but should not be reported because its PRMSE, although substantially larger than $PRMSE_s$ and $PRMSE_x$, will still not be adequately high.

The several figures in this paper summarize the results in an easily understandable manner and may be used to provide guidance to testing companies. For example, if a testing program willing to report subscores can only afford to have 20-item subscore, Figures 1 and 5 suggest that it has to make sure that the average disattenuated correlations between the subscores is less than 0.80 (approximately) to achieve the goal. 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 3 and Figures 4 to 8. However, the results of the simulation study mostly agree with those in Table 2 that is based on analysis of operational data—which 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 simulation 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 paper, limited simulations were performed under different conditions 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 given by the so-called *bad items* described in Sinharay (2006).[9] The results for such data did not differ much from those reported in Table 3.

There are several related issues that can be examined in further research. For example, the simulation study considered only dichotomous items—it is possible to perform further simulations based on polytomous models.[1]0 It may be worthwhile to simulate data that mimic those from testing programs other than the three considered in this paper. This paper considers subscores that do not share common items (that is the most common phenomenon in practice; all of the tests shown in Table 2 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. One could consider subscores of unequal length and unequal pairwise correlation in a future simulation study; however, there will likely be too many cases to consider in such a study and summarizing the results will be a challenge. It is possible to consider other methods for determining when a subscore has added value. Such methods include the method of fitting beta-binomial models to the observed subscore distributions (Harris & Hanson, 1991) and factor analysis. However, the method of Harris and Hanson (1991) involves significance testing with a χ^2 statistic whose null distribution is not well-established (p. 5), and the factor analysis approach involves many issues such as the choice of items versus item parcels, the choice of exploratory versus confirmatory factor analysis, and the choice of proper test statistics 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, 2009) could be another possible candidate, but Haberman and Sinharay (2009) found the method to provide results similar to the CTT-based method (Haberman, 2008); hence the MIRT-based method was not considered here. This paper chose the CTT-based method suggested by Haberman (2008), 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.

References

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.

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

Haberman, S. J., & Sinharay, S. (2009). *How can multivariate item response item response theory be used in reporting of subscores?* (ETS Research Rep. No. RR-10-09). Princeton, NJ: ETS.

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 Rep. No. RR-08-45). Princeton, NJ: ETS.

Haladyna, S. J., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation and the health professions*, *24*(7), 349–368.

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

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

Reckase, M. D. (2007). Multidimensional item response theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Volume 26. Psychometrics* (pp. 607–642). Amsterdam: 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: ETS.

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. (2009). Providing subscale scores for diagnostic infor-

mation: A case study when the test is essentially unidimensional. *Applied Measurement in Education*, *23*(1), 63–86.

Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education*, *17*(2), 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.

Notes

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

[2]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.

[3]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.

[4]Changing 0.01 to other small values such as 0.02 or 0.03 did not affect our conclusions much.

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

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

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

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

[9]One of these bad items has an item response function that is a mixture of two logistic functions, and another bad item has an item response function that does not go to 1 as θ goes to ∞.

[10]Figures 6 and 8, that show reliability, instead of length, along the X-axis, will be comparable to similar plots made from polytomous item response data.

Appendix

Here we describe the methodology of Haberman (2008) and Haberman et al. (2009) that was used in this paper to determine whether and how to report examinee level subscores. The analysis involves the observed subscore s, the true subscore s_t, the observed total score x, and the true total score x_t. It is assumed that s_t, x_t, $s - s_t$, and $x - x_t$ all have positive variances. As usual in classical test theory, s and s_t have common mean $E(s)$, x and x_t have common mean $E(x)$, and the true scores s_t and x_t are uncorrelated with the errors $s - s_t$ and $x - x_t$. Let $\rho(a, b)$ denote the correlation between a and b. It is assumed that the true subscore s_t and true total score x_t are not collinear, so that $|\rho(s_t, x_t)|$ is less than 1. This assumption also implies that $|\rho(s, x)| < 1$. Haberman (2008) considered several approaches for estimation of the true score s_t.

In the first approach, s_t is estimated by the constant $E(s)$, so that the corresponding mean squared error in estimation is $E[s_t - E(s)]^2 = \sigma^2(s_t)$.

In the second, the linear regression

$$s_s = E(s) + \rho^2(s_t, s)[s - E(s)]$$

of s_t on the observed subscore s estimates s_t, and the corresponding mean squared error is $E(s_t - s_s)^2 = \sigma^2(s_t)[1 - \rho^2(s_t, s)]$, where $\rho^2(s_t, s)$ is the reliability of the subscore.

In the third approach, the linear regression

$$s_x = E(s) + \rho(s_t, x)[\sigma(s_t)/\sigma(x)][x - E(x)]$$

of s_t on the observed total score x estimates s_t, and the corresponding mean squared error is $E(s_t - s_x)^2 = \sigma^2(s_t)[1 - \rho^2(s_t, x)]$.

Haberman (2008) compared the last two approaches with respect to their PRMSE. Relative to using $E(s)$, the PRMSE corresponding to the use of s_s as the estimate of s_t is

$$\frac{\sigma^2(s_t) - \sigma^2(s_t)[1 - \rho^2(s_t, s)]}{\sigma^2(s_t)} = \rho^2(s_t, s),$$

which is the reliability of the subscore. Relative to using $E(s)$, the PRMSE corresponding to the use of s_x as the estimate of s_t is $\rho^2(s_t, x)$, which can be shown to satisfy the relation (Haberman, 2008)

$$\rho^2(s_t, x) = \rho^2(s_t, x_t)\rho^2(x_t, x), \tag{A1}$$

where $\rho^2(x_t, x)$ is the total score reliability. We describe the computation of $\rho^2(s_t, x_t)$ shortly.

Haberman (2008) argued on the basis of these results that the true subscore is better approximated by s_x (which is an estimate based on the total score) than by s_s (which is an estimate based on the subscore) if $\rho^2(s_t, s)$ is smaller than $\rho^2(s_t, x)$, and hence subscores should not be reported in that case.

The fourth approach consists of reporting an estimate of the true subscore s_t based on the linear regression s_{sx} of s_t on both the observed subscore s and the observed total score x. The regression is given by

$$s_{sx} = E(s) + \beta[s - E(s)] + \gamma[x - E(x)], \text{ where}$$

$$\gamma = \frac{\sigma(s)}{\sigma(x)}\rho(s_t, s)\tau, \quad \tau = \frac{\rho(x_t, x)\rho(s_t, x_t) - \rho(s, x)\rho(s_t, s)}{1 - \rho^2(s, x)}, \text{ and } \beta = \rho(s_t, s)[\rho(s_t, s) - \rho(s, x)\tau].$$

The mean squared error is then $E(s_t - s_{sx})^2 = \sigma^2(s_t)\{1 - \rho^2(s_t, s) - \tau^2[1 - \rho^2(s, x)]\}$, so that the PRMSE relative to $E(s)$ is

$$\rho^2(s_t, s_{sx}) = \rho^2(s_t, s) + \tau^2[1 - \rho^2(s, x)].$$

Computation of $\rho^2(s_t, x_t)$

The quantity $\rho^2(s_t, x_t)$ can be expressed as

$$\rho^2(s_t, x_t) = \frac{[\text{Cov}(s_t, x_t)]^2}{V(s_t)V(x_t)}.$$

The variances are computed by multiplying the observed variance by the reliabilities; for example,

$$V(s_t) = \rho^2(s_t, s) \times \text{Observed variance of } s.$$

The covariance $\text{Cov}(s_t, x_t)$ can be expressed, where s_{kt} denotes the true k-th subscore, as

$$\text{Cov}(s_t, x_t) = \text{Cov}(s_t, \sum_k s_{kt}) = \sum_k \text{Cov}(s_t, s_{kt}).$$

The right-hand side of the equation is the sum of the t-th row of C_T, the covariance matrix between the true subscores. The off-diagonal elements of C_T are the same as those of the covariance matrix between the observed subscores; the k-th diagonal element of C_T is obtained as

$$\text{variance of the } k\text{-th observed subscore} \times \text{ reliability of the } k\text{-th subscore}\cdot$$