Research Memorandum

Reporting Subscores: A Survey

Sandip Sinharay

Shelby J. Haberman

December 2008

ETS RM-08-18

Reporting Subscores: A Survey

Sandip Sinharay and Shelby J. Haberman

ETS, Princeton, New Jersey

December 2008

Abstract

Recently, there has been an increasing level of interest in subscores for their potential diagnostic value. As a result, there is a constant demand from test users for subscores. Haberman (2008) and Haberman, Sinharay, and Puhan (2006) suggested methods based on classical test theory to examine whether subscores provide any added value over total scores. This paper applied the above mentioned methods to recent data sets from a variety of operational tests. The results indicate that subscores provide added value for only a handful of tests.

Key words: KR-20, proportional reduction in error, reliability

Acknowledgments

The authors thank Dan Eignor, Judit Antal, Rick Morgan, Gautam Puhan, Lora Monfils, Yu-Li Zhang, and Raymond Mapuranga for useful advice; Carmen Payton, Natalie Hatrak, Diane Cruz, and Chantal Delaney for their help with the data, and Amanda McBride for help with proofreading.

1 Introduction

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 on 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 [AERA, APA, & NCME] 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 (AERA, APA, & NCME, 1999) demands that if a test provides more than one score, the distinctiveness of the separate scores should be demonstrated. Several researchers, such as Tate (2004) and Wainer et al. (2001), also emphasized the importance of ensuring reasonable subscore performance.

It is apparent then that the quality of the subscores must be assessed before considering score reporting at the subscore level. Just as inaccurate information at the total test-score level can lead to inaccurate pass and fail decisions with damaging consequences to both the testing programs and test takers, inaccurate information at the subscore level can also lead to incorrect instructional and remedial decisions, resulting in large and needless expense for states or institutions.

Haberman (2008) and Haberman, Sinharay, and Puhan (2006) suggested methods based on classical test theory to examine whether subscores provide any added value over total scores at an individual level and at the level of institutions that the examinees belong to. Both of these papers, as well as Sinharay, Haberman, and Puhan (2007) and Puhan, Sinharay, Haberman, and Larkin (2008), showed examples, using data from several operational tests, of subscores that were of interest, but did not offer any added value over the total scores.

This study considered four well-known tests, several of which are large scale and have high stakes. Data from several forms of each of the four tests were analyzed. For each test, an analysis was performed using the methods suggested by Haberman (2008) and Haberman et al. (2006) to determine whether subscores have any added value over the total score at the individual level and, if applicable, at an aggregate level (such as the institutions or schools which to the examinees belong). This report recommends the optimum statistical approach to report subscores for each test, where applicable. Another goal of this report is to gather knowledge regarding when subscores are of added value; such knowledge should be useful to test developers.

Section 2 briefly describes the methodology involved. Sections 3 to 6, respectively, discuss the results obtained for the four tests. For confidentiality purposes, this report denotes the tests as Test A, Test B, Test C, and Test D, and the subscores as Subscore 1, Subscore 2, Subscore 3, and Subscore 4. Discussions and conclusions are provided in section 7.

2 Methodology

This section describes the approach of Haberman (2008) and Haberman et al. (2006) to determine whether and how to report subscores. We first describe the methods used at the examinee level. We then describe the methods used at an institutional level. Further details of the methods are provided in the appendix.

2.1 Examinee-Level Analysis

Consider the following estimates of the subscore:

- An estimate s_s based on the observed subscore
- An estimate s_x based on the observed total score
- An estimate s_{sx} based on the observed subscore and the observed total score. This is an example of an augmented subscore (Wainer et al., 2001). We will often refer to the procedure by which s_{sx} is obtained as the *Haberman augmentation.*

It is also possible to consider an augmented estimate s_{aug} based on the observed subscore of interest and the other observed subscores (Wainer et al., 2001). However, this augmentation provides results that are very similar to those of Haberman augmentation. Hence we do not provide any results for s_{aug} in this paper.

Haberman (2008) and Haberman et al. (2006) reformulated the problem of subscore reporting as one of estimating the true subscore s_t by one among the estimates s_s, s_x, and s_{sx}. The tool used to compare the estimates is the proportional reduction in mean squared error (PRMSE), where the larger the PRMSE, the more accurate is the estimate.[1] We 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 equal to the reliability of the subscore. Our strategy in general will be the following:

- If $PRMSE_s$ is less than $PRMSE_x$, we will declare that the subscore *does not provide added value over the total score* because the observed total score will provide more accurate diagnostic information than the observed subscore in that case. Sinharay et al. (2007) discussed why this strategy is reasonable and how it 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 we will recommend the use of s_{sx} only if it leads to substantially larger PRMSE compared to both s_s and s_x.

Occasionally, as we will see later for Tests A and D, we will encounter tests for which we will declare that the subscores are not of added value even before computing the PRMSE. These will mostly be tests for which subscores are based on very few items, or are highly correlated, or both.

Cronbach's α is used to estimate the reliabilities of subscores and total scores. Haberman (2008) and Haberman et al. (2006) showed that a subscore will have more chance of providing added value when it is reliable and distinct from the other subscores.

2.2 Institutional-Level Analysis

As with an individual-level analysis, we consider the following estimates of the average subscore for institutions:

- An estimate s_{Is} based on the average observed institutional-level subscore
- An estimate s_{Ix} based on the average observed institutional-level total score
- An estimate s_{Isx} based on the average observed institutional-level subscore and the average observed institutional-level total score

Our strategy will be the same as that in the individual-level analysis, that is, we will determine whether institutional-level subscores have added value over institutional-level total scores based on the PRMSEs (Haberman et al., 2006). Here, the values of the PRMSE will depend on the institution size. We consider three sizes: small, moderate, and large.

3 Test A

3.1 Data

Test A is a comprehensive test that uses scenario-based tasks to measure both cognitive and technical skills. All forms of Test A use 15 tasks administered via computer, and each task yields scores from two or more items. Each item has possible scores of 0, 0.5, and 1, with 1 being the highest score and 0 being the lowest score. Scoring is performed via computer, and the rules for assigning scores are somewhat complex. These rules lead in some cases to item scores for several items within a task that have restrictions on possible combinations of values. Each of the seven performance/skill areas is measured by 2 tasks giving 14 tasks in total, and the 1 remaining task is used to measure two performance/skill areas.

We analyzed data from five recently administered forms of Test A. The sample sizes were 603, 581, 1,050, 1,029, and 649. In these test forms, several items are often based on the same stimulus in a task, causing inter-item dependence (so that if an analysis is done at the item level, the reliability figures will be incorrect unless some adjustment is made for the inter-item dependence). Hence we performed our analysis based on the scores of examinees on the 15 tasks. The score on a task is the sum of the scores on the items based on the task. Fourteen of these tasks have items on only one skill area. One of these tasks, however, has a few items that are supposed to measure one skill area and a few other items that are supposed to measure another skill area. As the theory of Haberman (2008) and Haberman et al. (2006) does not yet apply to items/tasks that capture multiple skills, we partition the task into two different tasks; the items that measure one skill area are treated as part of one task while those measuring the other skill area are treated as part of another task.

We first performed an analysis under the assumption that each of the above mentioned seven performance areas correspond to a subscore. From the description earlier, each of the seven subscores is based on either two and three tasks. We then performed another analysis with four subscores that are also of interest. Three of these subscores are obtained by pooling two subscores each from the seven-subscore analysis, whereas the fourth subscore is the same as one in the seven-subscore analysis. Each of these four subscores is based on between three and five tasks.

3.2 Results From Examinee-Level Analysis

For each test form, the reliability of each subscore is given in Tables 1 (seven-subscore analysis) and 2 (four-subscore analysis). The correlation between the subscores for the Form 3 are shown in Table 3 (seven-subscore analysis) and Table 4 (four-subscore analysis). The correlations are similar for the other forms and are not shown. The total test reliability values are 0.86, 0.80, 0.88, 0.84, and 0.85, respectively, for the five forms.

Table 1
Reliability for the Five Forms of Test A: Seven-Subscore Analysis

Subscore	Form 1	Form 2	Form 3	Form 4	Form 5
1	0.59	0.40	0.62	0.34	0.42
2	0.23	0.18	0.21	0.49	0.51
3	0.37	0.34	0.43	0.48	0.52
4	0.30	0.26	0.30	0.25	0.25
5	0.50	0.42	0.54	0.52	0.57
6	0.43	0.32	0.44	0.47	0.41
7	0.62	0.47	0.63	0.43	0.48

Table 2

Reliability for the Five Forms of Test A: Four-Subscore Analysis

Subscore	Form 1	Form 2	Form 3	Form 4	Form 5
1	0.62	0.53	0.67	0.48	0.55
2	0.55	0.50	0.57	0.65	0.68
3	0.66	0.57	0.68	0.60	0.60
4	0.50	0.42	0.54	0.52	0.57

Table 3

Correlations for Form 3 of Test A: Seven-Subscore Analysis

	1	2	3	4	5	6	7
1	1.00	0.49	0.57	0.48	0.67	0.59	0.63
2	**1.38**	1.00	0.46	0.42	0.49	0.46	0.48
3	**1.11**	**1.54**	1.00	0.43	0.55	0.53	0.55
4	**1.10**	**1.69**	**1.18**	1.00	0.46	0.41	0.45
5	**1.16**	**1.46**	**1.14**	**1.13**	1.00	0.54	0.60
6	**1.12**	**1.52**	**1.23**	**1.11**	**1.11**	1.00	0.56
7	**1.01**	**1.33**	**1.06**	**1.03**	**1.03**	**1.06**	1.00

Note. The simple correlations are shown above the diagonal, and the disattenuated correlations are shown in bold font below the diagonal.

Table 4

Correlations for Form 3 of Test A: Four-Subscore Analysis

	1	2	3	4
1	1.00	0.67	0.70	0.69
2	**1.08**	1.00	0.67	0.61
3	**1.04**	**1.08**	1.00	0.64
4	**1.15**	**1.10**	**1.05**	1.00

Note. The simple correlations are shown above the diagonal, and the disattenuated correlations are shown in bold font below the diagonal.

The reliabilities in the seven-subscore case range from extremely low (such as 0.18 and 0.23) to rather low and those in the four-subscore case are also low.

The examinee-level analysis of Haberman (2008) involves the computation of C_T, the covariance matrix between the true subscores. The off-diagonal elements of C_T are the same as those of C_O, the covariance matrix between the observed subscores. Each diagonal element of C_T is obtained by multiplying the corresponding diagonal of C_O by the reliability of the corresponding subscore. The covariance matrix C_T was not positive semi-definite for any of these data sets, primarily because the reliability values were low and the correlations were high.[2]

This is equivalent to the phenomenon that for several pairs of subscores, the correlation after correcting for attenuation (which is the simple correlation between a pair of subscores divided by the square root of the product of the reliabilities of the two subscores) is larger than 1. Occasionally, the disattenuated correlations were as large as 1.7 for the 7-subscore case and 1.2 for the 4-subscore case.[3] Hence it is clear that the subscores are far from satisfying the criteria of reliability and distinctness as demanded by Standards 1.12 and 5.12 of the *Standards for Educational and Psychological Testing* (AERA, APA,& NCME, 1999) and we concluded without any further analyses that the examinee-level subscores do not provide any added value for Test A.

3.3 Results From Institutional-Level Analysis

The number of institutions were 18, 17, 37, 41, and 21, respectively, for the five forms. The median of the number of examinees from an institution was 18.5, 21, 16, 15, and 16 for the test forms, respectively (and the maximum number of examinees from an institution over all the forms was 197).

Tables 5 and 6 provide the values of 100xPRMSE for the five forms of Test A for the different estimates and for different institution sizes. For each form, three columns of PRMSEs are shown. The column with the heading S contains PRMSEs for the estimate based on the average institutional subscore. The column with the heading T contains PRMSEs for the estimate based on the average institutional total score. The column with the heading ST contains PRMSEs for the estimate based on the average institutional subscore and average institutional total score.

Table 7 shows 100 times the reliability of the prediction of the institutional-level true total score from the institutional-level average observed total score.

Tables 5 to 7 show that the institutional-level subscores often have lower PRMSE compared to the institutional-level total scores, even for institutions with size 100. Hence

Table 5

100 × Proportional Reduction in Mean-Squared Errors (PRMSEs) for Institutional Subscores for the Five Forms of Test A: Seven-Subscore Analysis

Institution	Skill		1			2			3			4			5	
size		S	T	ST	S	T	ST	S	T	ST	S	T	ST	S	T	ST
30	1	75	81	81	52	75	75	80	79	81	72	86	86	79	87	87
	2	61	72	74	44	74	74	54	76	76	68	86	86	81	87	87
	3	71	82	82	71	71	76	71	83	83	78	86	86	79	85	86
	4	64	65	72	58	65	69	65	71	75	73	81	83	74	76	81
	5	74	72	77	46	75	75	76	77	79	81	86	86	82	87	87
	6	77	82	82	65	75	75	78	83	83	72	86	86	77	87	87
	7	80	81	83	72	70	76	81	82	83	74	81	83	68	87	87
50	1	83	87	87	65	83	83	87	85	88	81	91	91	87	92	92
	2	72	78	81	57	83	83	66	82	82	78	91	91	87	92	92
	3	81	88	88	81	79	83	81	89	89	86	91	91	86	90	90
	4	75	70	80	69	73	77	76	76	82	82	86	89	83	80	87
	5	83	78	84	58	83	83	84	83	86	87	91	91	88	92	92
	6	84	88	88	75	83	83	86	89	89	81	91	91	85	92	92
	7	87	87	89	81	78	84	88	88	89	83	86	88	78	92	92
100	1	91	92	93	79	91	91	93	90	93	90	95	95	93	96	96
	2	84	83	88	73	90	90	79	86	88	88	95	95	93	96	96
	3	89	94	94	89	87	91	89	94	94	92	95	95	93	93	94
	4	86	74	88	82	79	86	86	80	89	90	90	93	91	83	92
	5	91	83	91	74	91	91	91	88	92	93	95	95	94	96	96
	6	92	94	94	86	91	91	92	94	94	90	95	95	92	96	96
	7	93	92	94	90	85	91	93	93	94	91	90	93	88	96	96

Note. S = estimate based on the average institutional subscore, T = estimate based on the average institutional total score, ST = estimate based on the average institutional subscore and total score.

we do not recommend reporting institutional-level observed average subscores for Test A. An augmented estimate based on the institutional-level observed average subscore and the institutional-level observed average total score has higher PRMSE compared to either of them alone and may be reported.

Table 6
100 × Proportional Reduction in Mean-Squared Errors (PRMSEs) for Institutional Subscores for the Five Forms of Test A: Four-Subscore Analysis

Institution	Skill		1			2			3			4			5	
size		S	T	ST	S	T	ST	S	T	ST	S	T	ST	S	T	ST
30	1	77	79	80	61	75	75	80	79	81	79	86	86	83	87	87
	2	73	81	81	67	75	75	72	81	81	80	85	85	84	87	87
	3	82	82	82	77	70	77	83	83	83	79	86	86	80	87	87
	4	74	72	77	46	75	75	76	77	79	81	86	86	82	87	87
50	1	85	85	87	72	83	83	87	85	88	86	91	91	89	92	92
	2	82	87	87	77	83	83	81	87	87	87	90	90	90	92	92
	3	88	88	88	85	78	85	89	89	89	86	91	91	87	92	92
	4	83	78	84	58	83	83	84	83	86	87	91	91	88	92	92
100	1	92	91	93	84	91	91	93	90	93	93	95	95	94	96	96
	2	90	93	93	87	91	91	90	92	92	93	94	95	95	96	96
	3	94	94	94	92	85	92	94	94	94	92	95	95	93	96	96
	4	91	83	91	74	91	91	91	88	92	93	95	95	94	96	96

Note. S = estimate based on the average institutional subscore, T = estimate based on the average institutional total score, ST = estimate based on the average institutional subscore and total score.

Table 7
100 × Reliability of the Prediction of the Institutional-Level True Total Score From the Institutional-Level Average Total Score for Test A

Institution size	Form 1	Form 2	Form 3	Form 4	Form 5
30	82	75	83	86	87
50	88	83	89	91	92
100	94	91	94	95	96

4 Test B

4.1 Data

Students take Test B, a battery of tests, to demonstrate their mastery of specific subject areas. We considered two language tests under Test B—denoted as Tests B1 and B2—and analyzed data from one recent administration each for the two tests. Each of them reports two subscores. The sample sizes were 2,820 and 7,376, respectively. We examined if the two reported subscores for Tests B1 and B2 offer any added value over the total score.

No institutional-level information was available—so we performed only an examinee-level analysis for Tests B1 and B2. Several items in these tests are based on a common stimulus, which leads to inter-item dependence. We pooled the items based on a common stimulus.

4.2 Results

The values of the PRMSEs are provided in Table 8. The table also shows the lengths of the subscores, raw score range, and correlations between the subscores. The total test reliabilities are 0.91 for Test B1 and 0.93 for Test B2.

Table 8
Proportional Reduction in Mean-Squared Errors (PRMSEs) for Tests B1 and B2

	Test B1		Test B2	
	1	2	1	2
Length	15	23	23	26
Raw score range	0–28	0–59	0–30	0–55
Correlation	1.00	0.77	1.00	0.68
	0.90	1.00	**0.75**	1.00
PRMSE_s	0.84	0.86	0.91	0.88
PRMSE_x	0.83	0.89	0.78	0.85
PRMSE_{sx}	0.88	0.90	0.92	0.89

Note. In the correlation matrix, the simple correlations are shown above the diagonal, and the attenuated correlations are shown in bold font below the diagonal.

Table 8 shows that the two subscores do not offer any added value over the total score for Test B1, while they do offer added value for Test B2. For Test B1, the first augmented subscore leads to a considerable gain in PRMSE while the second augmented subscore leads to a modest gain in PRMSE—so our recommendation would be to report augmented subscores for Test B1. For Test B2, the augmentation leads to only a modest gain for both the subscores—so reporting the subscores themselves seems justified.

5 Test C

5.1 Data

Test C, a battery of tests, gauges achievement in several disciplines. Each test under Test C is intended for students who have majored in or have extensive background in that specific area. We considered the two titles under Test C with the largest volumes—we denote them here as Tests C1 and C2. We analyzed data from two recently administered test forms for each of these two tests. The sample sizes were 4,242 and 3,870 for the two forms of Test C1 and 1,932 and 1,942 for the two forms of Test C2. In Test C1, which has approximately 205 multiple-choice (MC) items, the examinees receive two subscores. Some questions(about 17%) are not part of a reported subscore but contribute to the total reported score on Test C1. In Test C2, which has approximately 200 MC items, the examinees receive three subscores. We performed a three-subscore analysis for both these tests.

No institutional-level data were available—so we performed only an examinee-level analysis for Tests C1 and C2. Several items were based on a common stimulus, which led to inter-item dependence. We pooled the items based on a common stimulus. On a very few occasions, items based on a common stimulus contributed to two different subscores. On these occasions, we pooled the items that measure one subscore and separately pooled the items that measure the other subscore.

5.2 Results

The values of the PRMSEs are provided in Table 9. The table also shows the lengths of the subscores, raw score range, and correlations between the subscores. The total test reliabilities are 0.95 for both the forms of Test C1 and 0.94 and 0.95 for the two forms of Test C2.

Figure 1 shows, for nine examinees with total scores on Test C1 that span the whole range, the standardized observed subscores and the standardized estimated subscores obtained by Haberman augmentation for Form 1. The observed subscores are shown using vertical bars and the augmented subscores are shown using points joined by dashed lines. Figure 2 shows a similar plot for examinees on Test C2.

Table 9

Proportional Reduction in Mean-Squared Errors (PRMSEs) for Tests C1 and C2

	Test C1 Form 1			Test C1 Form 2			Test C2 Form 1			Test C2 Form 2		
	1	2	3	1	2	3	1	2	3	1	2	3
Length	82	86	32	80	87	25	49	50	48	49	51	44
Raw score range	0-82	0-88	0-35	0-82	0-88	0-35	0-66	0-67	0-67	0-67	0-67	0-66
Correlation	1.00	0.80	0.77	1.00	0.77	0.74	1.00	0.72	0.60	1.00	0.79	0.66
	0.90	1.00	0.75	**0.87**	1.00	0.74	**0.82**	1.00	0.76	**0.90**	1.00	0.78
	0.91	**0.90**	1.00	**0.89**	**0.90**	1.00	**0.68**	**0.88**	1.00	**0.75**	**0.91**	1.00
PRMSE_s	0.91	0.88	0.78	0.90	0.88	0.78	0.89	0.85	0.87	0.90	0.86	0.87
PRMSE_x	0.90	0.89	0.87	0.89	0.88	0.85	0.78	0.89	0.79	0.84	0.92	0.82
PRMSE_{sx}	0.93	0.92	0.89	0.92	0.91	0.88	0.91	0.91	0.89	0.91	0.93	0.90

Note. In the correlation matrix, the simple correlations are shown above the diagonal, and the disattenuated correlations are shown in bold font below the diagonal.

Table 9, and Figures 1 and 2 lead us to make the following conclusions:

- Test C1: The subscores barely provide any added value over the total score—the PRMSEs for the subscores in Table 9 never exceeds those for the total score by more than 0.01. The three standardized subscores (either observed or augmented) are often close to each other in Figure 1, which shows that the subscores are not too well-separated and should not be reported. Augmented subscores could be reported as the PRMSEs for them are somewhat larger than those for the subscores and the total score.

- Test C2: Though the second subscore does not seem to offer any added value over the total score, the other two subscores provide substantial added value over the total score—the PRMSEs for these two subscores are substantially larger than the PRMSE for the total score. Note the comparatively lower correlations between these two subscores. The three standardized subscores (either observed or augmented) in Figure 2 are much more separated from each other than those in Figure 1. A practical solution in this situation may be to report augmented subscores, which have PRMSEs larger than those for the subscores and the total score.

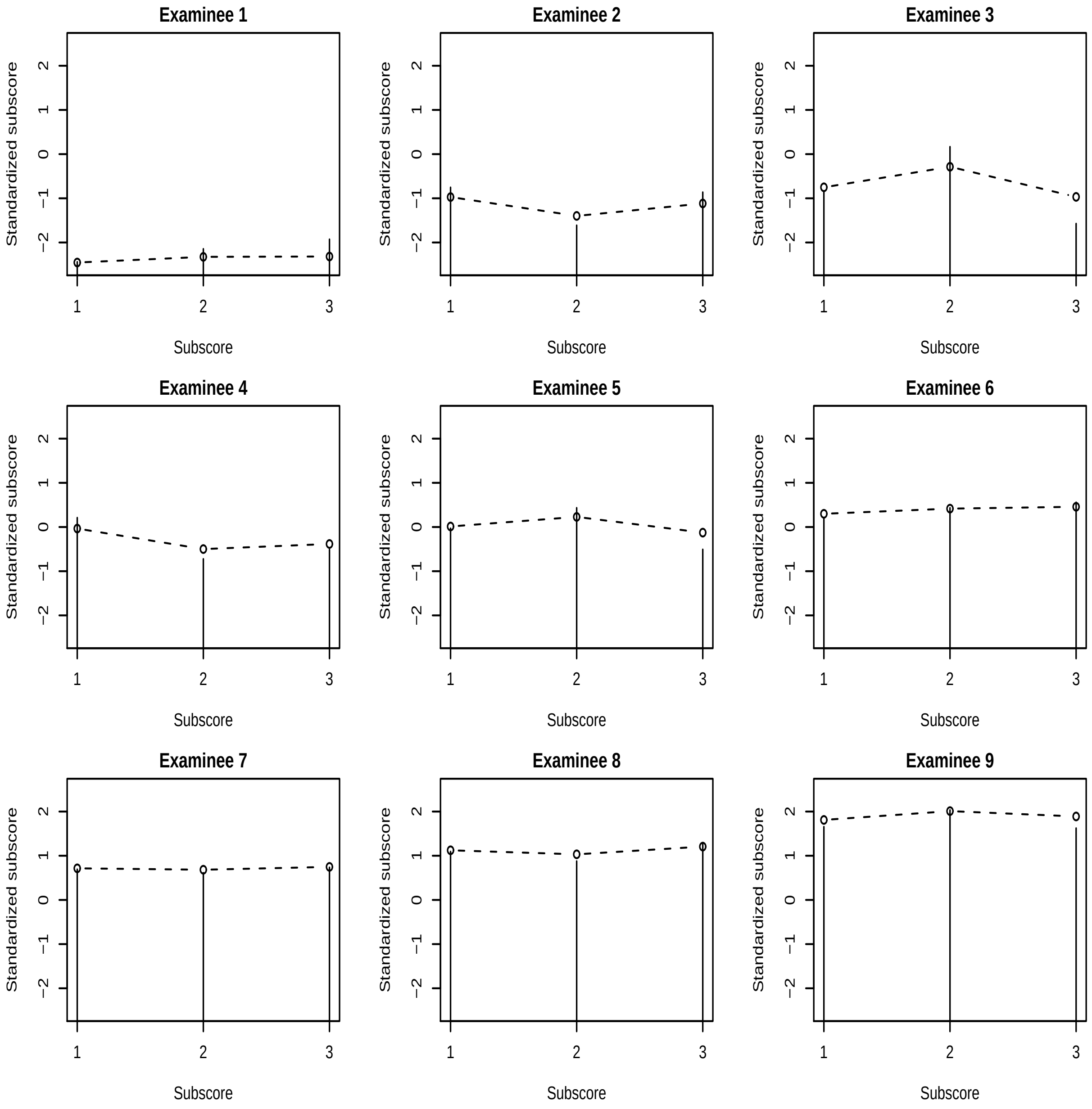

***Figure* 1 Observed and augmented subscores (both standardized) for nine examinees that took Form 1 of Test C1.**

6 Test D

6.1 Data

Test D, a battery of tests, measures school and individual student progress. We considered two assessments from Test D—Tests D1 and D2. Here, we report results from

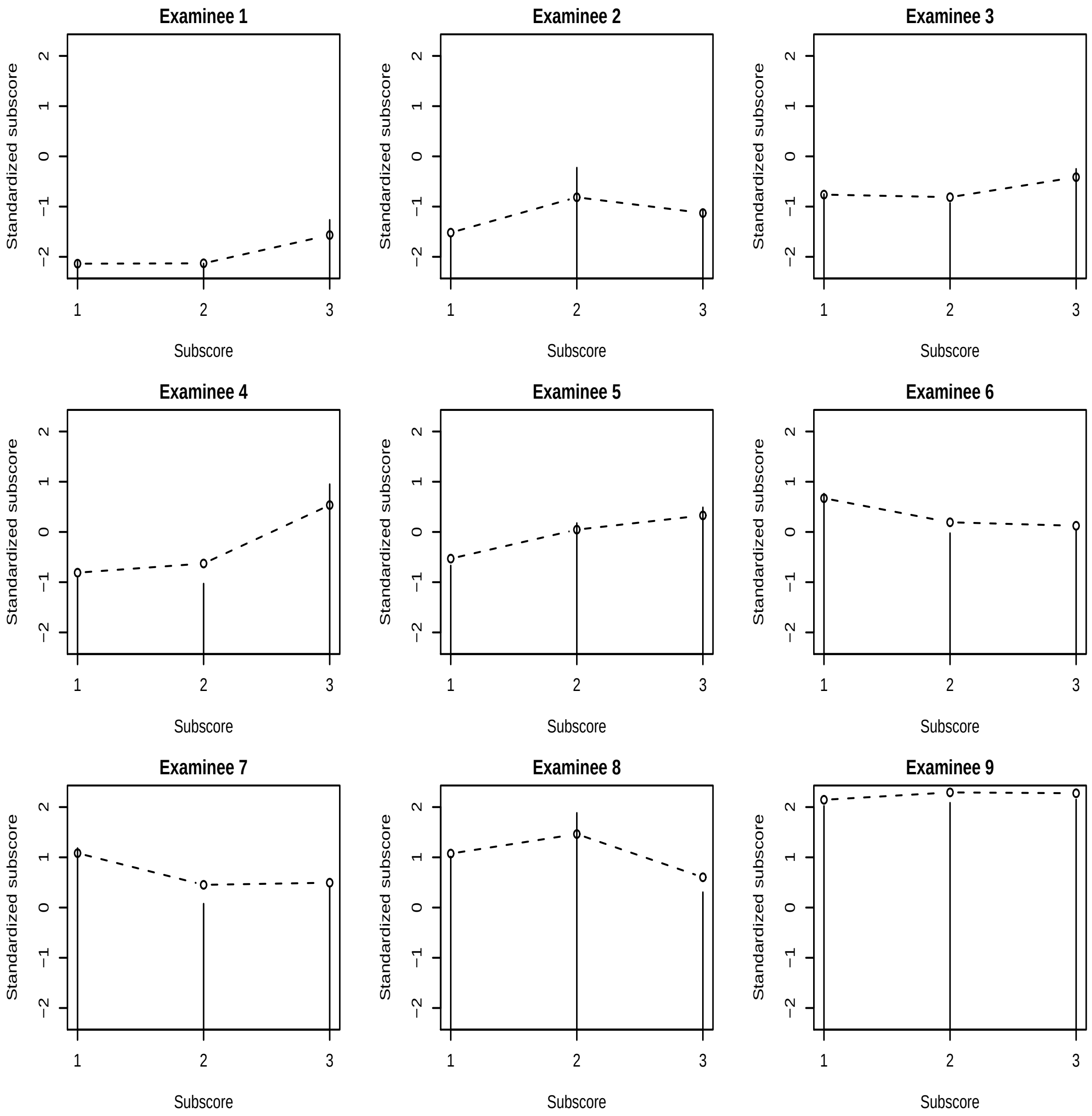

***Figure* 2 Observed and augmented subscores (both standardized) for nine examinees that took Form 1 of Test C2.**

two recent forms each of Tests D1 and D2. The resulting sample sizes were 6,563 and 55,390 for the two forms of D1, and 7,362 and 49,882 for the two forms of D2. In the 50-item (both MC and constructed-response [CR] items) Test D1, there are four subscores. In the 55-item Test D2 (both MC and CR items), there are six subscores. These subscores are not reported on the individual student score reports, but are included in the student data files sent to the schools.

We performed both an examinee-level analysis and a school-level analysis (though school-level scores are not reported operationally). Several items are based on a common stimulus, which leads to inter-item dependence. The different items based on common stimuli often contributed to different subscores. As the theory of Haberman (2008) and Haberman et al. (2006) does not yet apply to items contributing to multiple skills, we ignored the stimulus-level dependence and performed our analysis on the item scores. However, if the subscores are found not to have added value in this analysis, chances are extremely low that they will be found to have added value in an analysis that takes into account the stimulus-level dependence.

6.2 Results From Examinee-Level Analysis

Table 10 shows the lengths, reliabilities, and possible raw score ranges of the subscores and the correlations between the subscores. The total test reliabilities are 0.92 and 0.90 for Test D1 and 0.94 and 0.92 for Test D2.

As in Test A, the covariance matrix between the true subscores was not positive semi-definite for either of the two data sets from Test D1 or for either of the two data sets from Test D2, primarily because the reliability values were modest and the correlations were high. As a result, Table 10 has several disattenuated correlations larger than 1.[3] Hence it is clear that the subscores for Test D are far from satisfying the criteria of reliability and distinctness as demanded by Standards 1.12 and 5.12 of the *Standards for Educational and Psychological Testing* (1999). So we concluded that the individual-level subscores do not provide any added value for Test D.

6.3 Results From School-Level Analysis

The number of schools were 187, 253, 190, and 248, respectively, for the four data sets for Tests D1 and D2. The median of the number of examinees from a school was 10, 222, 15, and 173, respectively (and the maximum of the number of examinees from a school across all four data sets was 799).

Table 10

Reliabilities and Correlations for Tests D1 and D2

Form		Subscores					
		1	2	3	4	5	6
Test D1 Form 1							
	Length	14	12	10	14		
	Raw score range	0-16	0-14	0-16	0-14		
	Correlation	1.00	0.79	0.70	0.75		
		1.06	1.00	0.69	0.71		
		0.95	**0.99**	1.00	0.72		
		0.96	**0.96**	**0.97**	1.00		
	Reliability	0.77	0.71	0.69	0.78		
Test D1 Form 2							
	Length	14	12	10	14		
	Raw score range	0-16	0-14	0-16	0-14		
	Correlation	1.00	0.76	0.65	0.65		
		1.06	1.00	0.66	0.64		
		0.90	**0.94**	1.00	0.66		
		0.90	**0.91**	**0.93**	1.00		
	Reliability	0.74	0.70	0.70	0.71		
Test D2 Form 1							
	Length	10	9	10	10	6	10
	Raw score range	0-16	0-12	0-13	0-13	0-9	0-13
	Correlation	1.00	0.73	0.73	0.76	0.74	0.79
		0.99	1.00	0.72	0.73	0.68	0.73
		0.98	**1.01**	1.00	0.72	0.69	0.73
		0.99	**0.99**	**0.98**	1.00	0.71	0.74
		1.02	**0.98**	**0.99**	**0.98**	1.00	0.75
		1.02	**0.99**	**0.98**	**0.96**	**1.03**	1.00
	Reliability	0.77	0.71	0.71	0.77	0.68	0.77
Test D2 Form 2							
	Length	10	9	10	10	6	10
	Raw score range	0-16	0-12	0-13	0-13	0-9	0-13
	Correlation	1.00	0.71	0.69	0.69	0.67	0.71
		1.04	1.00	0.72	0.69	0.65	0.70
		0.99	**1.05**	1.00	0.70	0.64	0.69
		1.04	**1.07**	**1.07**	1.00	0.65	0.67
		1.10	**1.09**	**1.05**	**1.13**	1.00	0.66
		1.01	**1.03**	**0.98**	**1.02**	**1.08**	1.00
	Reliability	0.70	0.67	0.70	0.63	0.53	0.70

Note. Test D1 has four subscores while Test D2 test has six. In the correlation matrix, the simple correlations are shown above the diagonal, and the disattenuated correlations are shown in bold font below the diagonal.

Table 11
100 × Proportional Reduction in Mean-Squared Errors (PRMSEs) for School Subscores for Tests D1 and D2

School	Skill	Test D1 Form 1			Test D1 Form 2			Test D2 Form 1			Test D2 Form 2		
size		S	T	ST	S	T	ST	S	T	ST	S	T	ST
10	1	83	84	85	65	70	70	86	88	88	73	76	76
	2	81	84	84	63	71	71	83	87	87	71	76	76
	3	79	83	84	68	68	70	82	86	87	71	75	75
	4	78	84	84	62	70	70	87	88	89	69	76	76
	5	-	-	-	-	-	-	85	87	88	67	75	75
	6	-	-	-	-	-	-	85	88	88	70	75	76
30	1	93	93	94	85	86	87	95	95	95	89	89	90
	2	93	93	94	84	87	87	94	94	95	88	89	90
	3	92	93	93	86	84	87	93	93	95	88	88	89
	4	92	93	94	83	86	87	95	95	96	87	89	90
	5	-	-	-	-	-	-	94	93	95	86	88	89
	6	-	-	-	-	-	-	95	95	95	88	89	90
100	1	98	97	98	95	94	95	98	97	99	96	95	97
	2	98	97	98	95	95	96	98	96	98	96	95	97
	3	97	96	98	95	92	96	98	96	98	96	94	96
	4	97	97	98	94	94	95	99	97	99	96	95	96
	5	-	-	-	-	-	-	98	96	98	95	94	96
	6	-	-	-	-	-	-	98	97	98	96	95	96

Note. S = estimate based on the average institutional subscore,
T = estimate based on the average institutional total score,
ST = estimate based on the average institutional subscore and total score.

Table 11 provides the values of 100 × PRMSE for the four data sets for schools sizes 10, 30, and 100. For each test/form combination, three columns of PRMSEs are shown. The column with the heading S contains the PRMSE for the estimate based on the average school subscore. The column with the heading T contains the PRMSE for the estimate based on the average school total score. The column with the heading ST contains the PRMSE for the estimate based on the average school subscore and average school total score.

Table 12 shows 100 times the reliability of the prediction of the school-level true total score from the school-level average observed total score.

Table 11 shows that the school-level subscores often have lower PRMSE compared to the school-level total scores, even for schools with size 100. In cases where the school-level

Table 12
100 × Reliability of the Prediction of the School-Level True Total Score From the School-Level Average Total Score for Tests D1 and D2

School size	Test D1 Form 1	Test D1 Form 2	Test D2 Form 1	Test D2 Form 2
10	85	71	89	77
30	94	88	96	91
100	98	96	99	97

subscores have higher PRMSE, the difference is small. Hence it seems that reporting school-level average subscores would not be justified for these tests. An augmented estimate based on the school-level observed average subscores and the school-level observed average total score has higher PRMSE compared to either of them alone and could be reported. Table 12 shows that total scores could be reported for schools with at least 30 examinees.

7 Discussion and Conclusion

Although we considered a wide variety of tests, the subscores were found to have added value over total scores for few of these tests. Table 13 summarizes our findings for the tests. We found that the subscores provided no added value for Tests A and D—even augmentation did not provide a meaningful way to report subscores for these assessments (because the augmented subscores replicate the total score and hence are very close to each other for these tests). This finding is similar to that of Puhan, Sinharay, Haberman, and Larkin (2008), who analyzed data from several tests that report subscores and found that the subscores did not provide added value for those tests.

The results also suggest that any possible use of subscores is most likely to succeed with subscores based on a large number of items. Subscores that are based on only a few items (for example, Tests A or D) generally do not provide any added value. On the other hand, the subscores that are of added value (for example, Test C2) are mostly based on a large number of items.

However, the prediction of when a subscore will provide added value is not trivial, which is a finding that will affect testing programs interested in reporting subscores.

Table 13
Summary of Our Findings

Test	Finding	Note
Test A	Subscores have added value at neither the examinee level nor the institutional level.	
Test B	Subscores have added value for Test B1 and do not have added value for Test B2.	Augmented subscores can be reported.
Test C	Subscores barely have added value for Test C1 and two out of three subscores have added value for Test C2.	Augmented subscores can be reported.
Test D	Subscores have added value at neither the examinee level nor the school level.	

Although it is reasonably clear that subscores based on a few items or a few tasks are unlikely to prove useful, it does not follow that subscores based on many items or many tasks are always useful (for example, the two subscores for Test C1, both based on 80+ items, hardly provide any added value over the total score). Hence it is not possible, based on the findings here, to make clear recommendations for test developers regarding when subscores of a test will have added value. Sinharay (in press) provides further insight on this issue. No substitute exists for analysis of test data.

Augmentation seems to be the preferable way to report subscores for several of these tests. The PRMSE for an augmented subscore is always larger than that for both the observed subscore and the total score and substantially so for several of these tests. A problem with augmented subscores is that it may be difficult to explain them to clients. However, we are of the opinion that the statistical superiority of augmented subscores more than makes up for the difficulty in explaining them.

An important issue with reporting subscores is that equating and/or scaling of subscores would seem to be essential for most practical applications. In typical cases, equating is feasible for the total score but not for subscores (for example, if an anchor test is used to equate the total test, only a few of the items will correspond to a particular subscore so that an anchor test equating of the subscore is not feasible). However, this issue must be considered case by case. In many cases in which subscores are based on a large number of items, equating is already done or can readily be accomplished. In some cases in which augmentation is involved, linking can be accomplished by techniques currently associated with AP® examinations. In AP examinations, the MC score in a new form is equated to the MC score in an old form using an anchor test and then the MC + CR score on the new form is linked to the MC score on the new form (the MC score on the old form was linked to the MC + CR score on the old form); the linking is satisfactory because of the high correlation between the MC and MC + CR scores. Similarly, if the total score on a new form of a test is equated to that on an old form, it will be possible most often to link an augmented subscore in the new form to the total score in the new form. The usually high correlation between an augmented subscore and the total score will lead to satisfactory linking.[4] The important point is that appropriate methods of linking must be considered in a decision on if and how to report subscores. Puhan and Liang (in press) examined linking of subscores.

Subscores must be reported on some established scale. A temptation exists to make this scale comparable to the scale for the total score or to the fraction of the scale that corresponds to the relative importance of the subscore, but these choices are not without difficulties given that subscores and total scores typically differ in reliability. In addition, if the subscore is worth reporting at all, then the subscore presumably does not measure the same construct as the total score.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2009). *Standards for educational and psychological testing* (1999). Washington DC: American Educational Research Association.

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

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

Haberman, S. J., Sinharay, S., & Puhan, G. (2005). *Subscores for institutions.* (ETS Research Rep. No. RR-06-13). Princeton, NJ: ETS.

Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation and the Health Professions, 24*(7), 349–368.

Puhan, G., & Liang, L. (in press). *Equating subscores under the nonequivalent groups with anchor test (NEAT) design.* Princeton, NJ: ETS.

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2008, March). *Comparison of subscores based on classical test theory.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Sinharay, S. (in press). *When can subscores be expected to have added value? Results from operational and simulated data.* 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.

Stricker, L. J., Rock, D. A., & Lee, Y.-W. (2005). *Factor structure of the LanguEdge test across language groups* (TOEFL Monograph Series No. MS-32). Princeton, NJ: ETS.

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., et al. (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 Associates.

Notes

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

[2] We could have applied the approach of Bock and Petersen (1975) to obtain a corrected covariance matrix between the true subscores that is positive semi-definite. However, the approach consists in setting one or more linear combinations of the true subscores equal to zero, which is equivalent to assuming that some of the subscores provide no added value.

[3] Note that our estimated reliability, the Cronbach's α, provides a lower bound, so that the actual reliability may be a little higher than this estimate, which will make the actual correlation after correcting for attenuation a little lower. However, the values of the attenuated correlation were too high to attribute the values obtained to this phenomenon.

[4] Of course, if this correlation is too high, that will mean that the augmented subscore is virtually the same as the total score and then the augmented subscore will not provide any added value.

Appendix

Here, we describe the methodology of Haberman (2008) and Haberman et al. (2006), which was used in this paper to determine whether and how to report examinee-level subscores. The examinee-level 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 $x - x_t$. 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 prediction of the true score s_t.

In the first approach, s_t is predicted by the constant $E(s)$, so that the corresponding mean-squared error 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 predicts 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 predicts 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 three 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 $\rho^2(s_t, s)$, 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

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

where

$$\gamma = \frac{\sigma(s)}{\sigma(x)}\rho(s_t, s)\tau,$$

$$\tau = \frac{\rho(x_t, x)\rho(s_t, x_t) - \rho(s, x)\rho(s_t, s)}{1 - \rho^2(s, x)},$$

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

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

The variances in the denominator are computed by multiplying the corresponding observed variances by the reliabilities; for example,

$$V(s_t) = \rho^2(s_t, s) \times V(s).$$

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

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

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

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