Subscores Based on Classical Test Theory: To Report or Not to Report

Sandip Sinharay, Shelby Haberman, and Gautam Puhan, *Educational Testing Service*

There is an increasing interest in reporting subscores, both at examinee level and at aggregate levels. However, it is important to ensure reasonable subscore performance in terms of high reliability and validity to minimize incorrect instructional and remediation decisions. This article employs a statistical measure based on classical test theory that is conceptually similar to the test reliability measure and can be used to determine when subscores have any added value over total scores. The usefulness of subscores is examined both at the level of the examinees and at the level of the institutions that the examinees belong to. The suggested approach is applied to two data sets from a basic skills test. The results provide little support in favor of reporting subscores for either examinees or institutions for the tests studied here.

Keywords: institutional-level subscore, mean squared error, reliability

Educational and psychological tests often have different subsections based on content categories or blueprints. For example, a test on mathematics knowledge may have subsections on algebra and geometry. Similarly, a test of general ability can have subsections on mathematics, reading, and writing. Scores assigned to these subsections are commonly known as *subscores*. Subscores can be reported both at an individual examinee level and at an aggregate level. Failing candidates want to know their strengths and weaknesses in different content areas to plan future remedial studies. States and academic institutions such as colleges and universities want a summary of performance for their graduates to better evaluate their training and focus on areas that need remediation (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. Although many tests are designed to cover a broad domain and the total test score is considered as a composite of different abilities measured by different subsections, a subsection with fewer items than the total test may not be able to precisely measure a unique ability.

Haberman (2005) argued that a subscore may be considered useful only when it provides a more accurate measure of the construct being measured than is provided by the total score. Wainer et al. (2001) suggested that a test used for diagnostic purposes must yield scores that are reliable both for the total test and for the subscores associated with specific subsections or content areas. Finally, Tate (2004) has emphasized the importance of ensuring reasonable subscore performance in terms of high reliability and validity to minimize incorrect instructional and remediation decisions. The Standard 5.12 of the *Standards for Educational and Psychological Testing* (1999) states that

> 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, the Standard 1.12 of the *Standards for Educational and Psychological Testing* (1999) demands that if a test provides more than one score, the distinctiveness of the separate scores should be demonstrated.

From the above review, it is apparent that the quality of the subscores, whether at examinee level or at an aggregate level, must be assessed before reporting them. It also serves as an important reminder of the following: *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 remediation decisions resulting in large and needless expenses for states or institutions.*

The primary purpose of this article is to demonstrate the application of a statistical measure described in Haberman (2005) and Haberman, Sinharay, and Puhan (2006) for evaluating when subscores have any added value over the total scores. The measure is based on classical test theory (CTT). Two data sets from an operational test for which there is great interest in the reporting

Dr. Sandip Sinharay, Educational Testing Service, MS-12T, Rosedale Road, Princeton, NJ 08541; ssinharay@ets.org.

Note: Any opinions expressed in this paper are those of the authors and not necessarily of ETS.

of subscores were used. This test uses classical methods to score, scale, and equate scores. First, an analysis is described to assess whether examinee-level subscores have added value. Then, a similar analysis is described to determine whether institutional-level subscores have added value. Because few studies (with the exception of Longford, 1990) examine subscores at an aggregate level, this analysis may make an important contribution to the literature. The statistical measure employed is conceptually similar to test reliability and hence practitioners will find the measure appealing. All the computations involved are quite simple and use popular software programs, so that their operational implementation is straightforward. The data analyses show that subscores do not have added value, either at the examinee level or at the institutional level, for the data sets considered. Thus this article serves another purpose—that of demonstrating that there is a need to be careful about reporting subscores.

The next section reviews current approaches for reporting subscores. The following two sections describe the data sets and then the methodology. The methods are employed to two data sets in the penultimate section. Discussion and conclusions are provided in the last section.

Current Approaches for Reporting Subscores

Several researchers have examined the issue of reporting subscores and of evaluating if subscores are of added value given total scores.

Factor analysis, usually on the tetrachoric correlation matrix, has often been used to determine the number of subscores in a test. One could use a confirmatory factor analysis model (if there are some prior ideas about the subscore composition), or an exploratory factor analysis model to explore the number of subscores. For example, Grandy (1992) employed a nine-factor model, one factor for each subtest, in a confirmatory factor analysis of National Teacher Examination (NTE) Core Battery data from its November 1982 administration, but factors correlated too highly to justify nine different constructs. Additional exploration by Grandy (using both confirmatory and exploratory factor anaysis) suggested that there were three distinct factors (i.e., general academic skills, mathematics, and essay writing).

Harris and Hanson (1991) employed a method that fits 4-parameter beta-binomial distributions to the observed subscore distributions to determine if subscores are of added value given total scores. If the bivariate distribution of subscores (as they considered tests with two subscores) computed under the assumption that their true scores are functionally related provides an adequate fit to the observed bivariate distribution of the subscores, that will provide support for the assumption that the true subscores are functionally related and hence do not provide any added value. They found the subscores to be of no added value for English and Mathematics tests from the P-ACT+ examination.

Wainer, Sheehan, and Wang (2000) and Wainer et al. (2001) suggested the idea of an augmented subscore, that is, an estimated subscore based on all the available observed subscores so that, for example, an estimated algebra subscore of a student will be based on not only the observed algebra score of the student, but also on his/her other observed subscores, like those on arithmetic and geometry. Augmented subscores are more stable than the observed subscores themselves, especially when the observed subscores are based on a few items. Wainer et al. (2000) declared that "all subscores are measuring the same thing" (in other words, the subscores do not have any added value) if the reliabilities of the augmented subscores are all equal and are equal to that of the total score. They found subscores from an 150-item (all multiple choice) Education in the Elementary School Assessment (EES), one of the PRAXIS^{TM} examinations for teacher licensure, of no added value and commented that "We learned that the test's items were fiercely unidimensional, and so any set of subscales that were chosen would yield essentially the same information." They found that augmented subscores were reliable enough for diagnostic purposes; however, they also found (see their figure 8) that for any examinee, the standardized augmented subscores were all virtually the same, in which case the subscores did not serve any useful diagnostic purpose. Wainer et al. (2001) found the six subscores to have no added value for a 100-item (multiple choice) 1994 American Production and Inventory Control Society (APICS) Certification Examination, but found the four subscores to have some added value for a 26-item (constructed response) North Carolina Test of Computer Skills.

Yen (1987) proposed the objective performance index (OPI), which is based on fitting a unidimensional item response theory (IRT) model to test data, and is an estimate of the true subscore. OPI combines information from observed subscore and observed total score. This approach, because of the use of a unidimensional IRT model that may not adequately describe the data, may not provide accurate results when the data are truly multidimensional, as should be true when subscores have added value.

Though the focus of this article is on subscores based on CTT, we acknowledge the wealth of research on subscores based on multidimensional IRT. Researchers de la Torre and Patz (2005) applied a multidimensional IRT (MIRT) model using the Markov chain Monte Carlo (MCMC) algorithm to data from tests that measure multiple correlated abilities. This method can be used to estimate subscores, although the subscores, components of an IRT model, will be not in the raw score scale, but in the θ-scale. Nonetheless, this approach provided results very similar to those of Wainer et al. (2001). Other researchers, for example, Yao and Boughton (2007) also examined subscore reporting based on a MIRT model and the MCMC algorithm. Use of the MCMC algorithm is time-consuming and hence appears unsuitable for many applications to testing programs with short score-reporting deadlines. Nonetheless, it is possible to fit a MIRT model using more efficient algorithms such as the EM or stabilized Newton–Raphson algorithm, so that more practical application of MIRT to subscores is possible.

Researchers have also compared different approaches for reporting subscores. For example, Dwyer, Boughton, Yao, Steffen, and Lewis (2006), who compared four methods—raw subscores, OPI, augmentation, MIRT-based subscores—found the MIRT-based methods and augmentation methods to provide the best estimates of subscores overall.

Longford (1990) first recognized the importance of studying if subscores at an aggregate level are of added value

over aggregated total scores. He studied reporting of subscores at the college level using a multilevel variance component analysis on the data from the pilot stage of development of a test, and also found college-level subscores to be of little added value for one of the tests considered.

Our approach has the advantage of simplicity in that it does not require fitting a statistical model and requires only simple summary statistics of the subscores for the computations. Among the methods discussed above, only the augmentation method of Wainer et al. (2000) shares the advantage. Also, an investigator using our approach only has to compare two numbers to determine if a subscore is of added value—so the approach is very objective, without the scope of an investigator's personal judgment contaminating the outcome. Another advantage of the measure suggested is that it is conceptually very close to test reliability—so the measure will be intuitively appealing to the practitioners. This article has the added benefit over any above-mentioned work, other than that of Longford (1990), that this also suggests a straightforward approach to determine whether subscores at an aggregate level have added value or not.

Data

There is a constant demand for subscores, both at individual level and at institutional level, for a basic skills test, and the test administrators wanted to find out if reporting of subscores was justified for this test. Data from two recent forms (referred to as test form 1 and test form 2) of the test were used to examine if the subscores are of added value. The test has six operational subscore categories, namely (i) reading skills, (ii) reading application, (iii) mathematics skills, (iv) mathematics application, (v) writing skills, and (vi) writing application. Although the results for six subscores were of primary interest, we also examine the results for the case with the three subscores writing, mathematics, and reading obtained by pooling the skills and application portions of each of the three content areas.

The test takers are primarily prospective or practicing teacher's aides in the classroom. A majority of the test takers (approximately 80%) have either earned a high school diploma or have completed less than 2 years in college (i.e., completed some course work but did not get a college degree). Among the remaining test takers, approximately 7% completed the General Educational Development (GED) test and 10% had less than an Associate's degree (i.e., completed some course work but did not get an associate degree).

For each of the test forms, about 25% of the examinees did not report the names of their institutions (i.e., school districts). As a consequence, these examinees were removed from the analysis of institutional-level subscores. The precise effect of this omission cannot be readily determined. After removing these examinees, the number of examinees for the two test forms were 3,240 and 2,331 respectively. The respective number of institutions were 712 and 653. The number of students in an institution (*institution size*) ranged from 1 to 90 in these data, with the median size being 2 for both test forms, the 75th percentile being 4 for both test forms, and the 95th percentile being 16 and 14 for the respective test forms. Given current reporting standards for institutions that typically require at least 10 examinees for reporting subscores (ETS, 2006), the number of institutions for which any subscore report is possible is limited unless reports combine more than one administration of the test.

The Methodology

This section first describes the methodology for examinee-level subscores before proceeding to institutional-level subscores.

Examinee-Level Analysis

As is typical with classical test theory (CTT), let us denote the observed subscore as s, the true subscore as s_t, the observed total score as x, and the true total score as x_t. It is assumed that s_t, x_t, $s_e = s - s_t$, and $x_e = x - x_t$ all have positive variances, $\sigma^2(s_t)$, $\sigma^2(x_t)$, $\sigma^2(s_e)$, and $\sigma^2(x_e)$ respectively. The expectations of both s and s_t are equal to $E(s)$, the expectations of both x and x_t are equal to $E(x)$, and the true scores s_t and x_t are uncorrelated with the errors s_e and x_e.

While reporting a subscore for an examinee, the goal from a CTT perspective is to predict the "true subscore" s_t from the observed score of the examinee in different parts (including the part to which the subscore of interest belongs) of the test. It is then imperative that a necessary condition to be met for subscores to have added value is that *the true subscore should be predicted better by a predictor based on the observed subscore than by a predictor based on the total score*. Let us refer to this condition as Condition 1. If this condition is not satisfied, then instructional or remediational decisions based on subscores will have the undesirable property that they will lead to more errors than those based on total scores.

Predictors of True Subscore

It is possible to think of the following two predictors of the "true subscore" based on the observed subscore:

(i) The observed subscore s itself.

(ii) The predictor based on a regression of the true subscore on the observed subscore (which is Kelley's formula applied to the subscore). The predictor, obtained after some algebra (Haberman, 2005), is given by

$$s_s = E(s) + \rho^2(s_t, s)[s - E(s)], \quad (1)$$

where $\rho^2(s_t, s)$ is the reliability of the subscore. In an application of the above formula, $E(s)$ is estimated by the average observed subscore over all the examinees and $\rho^2(s_t, s)$ is estimated by the KR-20 approach (Kuder & Richardson, 1937) (the test we considered here contains dichotomous items only; however, the approaches discussed in this paper apply to tests with other item types as well).

The predictor of the "true subscore" based on the observed total score is a regression of the "true subscore" on the observed total score, which, after some algebra, is given by

$$s_x = E(s) + \rho(s_t, x)[\sigma(s_t)/\sigma(x)][x - E(x)], \quad (2)$$

where $\rho(s_t, x)$ is the correlation between the true subscore s_t and the observed total score x. In an application of the above formula, $\sigma(s_t) = \sigma(s)\sqrt{\rho^2(s_t, s)}$ is computed using the values of the observed variance of the subscore and estimated reliability, $\sigma(x)$ is the observed standard deviation (SD) of the total score, and $\rho(s_t, x)$ is computed using the formula

$$\rho(s_t, x) = \sqrt{\rho^2(s_t, x_t)\rho^2(x_t, x)}, \quad (3)$$

where $\rho^2(x_t, x)$, the total score reliability, is computed using the KR-20 approach (Kuder & Richardson, 1937), and the computation of $\rho^2(s_t, x_t)$ is described in Haberman (2005).

For example, consider a 25-item reading subtest whose mean score and reliability are 18 and .80 respectively. Suppose that the total test (that has writing and mathematics subtests besides the reading subtest, and a total of 75 dichotomous items) has mean and reliability of 50 and .90 respectively. Suppose further that the SD of the reading subscore and the total test score are 5 and 15, respectively, and that $\rho^2(s_t, x)$ is .90. Then $\sigma(s_t) = 5\sqrt{\rho^2(s_t, s)} = 5\sqrt{0.80} = 4.47$. Consider an examinee who obtained a reading subscore of 24 and a total test score of 60. For the examinee, the predictor s of his/her true reading subscore will be 24, the predictor s_s will be

$$18 + .80(24 - 18) = 22.8,$$

and the predictor s_x will be

$$18 + \sqrt{.90}\frac{4.47}{15}(60 - 50) = 20.8.$$

For the above-mentioned Condition 1 to be satisfied, at least one among s and s_s has to be a better predictor of the true subscore than s_x. A natural question is "What criterion should we use to judge that a predictor is better than another?" An answer to the question is discussed below.

Criterion for Comparing Predictors of True Subscore

Haberman (2005) suggested the use of mean squared error (MSE) of a predictor as the criterion in this situation. The MSE is a popular criterion for comparing the performance of competing estimators. The MSE for a predictor in this context measures the average squared error in predicting the true subscore by the predictor. Practically, larger MSE would lead to more error in instructional and remedial decisions.

For the predictor s above, the MSE is

$$E(s - s_t)^2 = E\left(s_e^2\right) = \sigma^2(s_e), \quad (4)$$

that is, the subscore error variance.

For the predictor s_s, the MSE can be shown to be

$$E(s_s - s_t)^2 = \sigma^2(s_t)\left[1 - \rho^2(s_t, s)\right], \quad (5)$$

and, for the predictor s_x, the MSE can be shown to be

$$E(s_x - s_t)^2 = \sigma^2(s_t)[1 - \rho^2(s_t, x)]. \quad (6)$$

Haberman (2005) also suggested a measure based on MSEs that is conceptually very close to the test reliability. Consider the trivial predictor $E(s)$ that predicts the true subscore of every examinee by the same number, the average subscore over all examinees. The MSE for this trivial predictor can be shown to be $\sigma^2(s_t)$. Now calculate the proportional reduction of mean squared error (PRMSE) for the three predictors s, s_s, and s_x, compared to the MSE for the trivial predictor. For example, for the predictor s_s, the PRMSE is given by

$$\frac{\text{MSE for the trivial predictor} - \text{MSE for } s_s}{\text{MSE for the trivial predictor}}, \quad (7)$$

which is equal to $\rho^2(s_t, s)$, the subscore reliability. Thus, when a true subscore is predicted by its regression on the observed subscore, the PRMSE criterion is identical to the subscore reliability and hence the criterion should be appealing to the psychometric community. Note that smaller MSE is equivalent to larger PRMSE and hence a predictor with a larger PRMSE is preferable to one with a smaller PRMSE.

The PRMSE for the predictor s can be shown to be equal to $2 - 1/\rho^2(s_t, s)$, which can be shown to be always less than or equal to $\rho^2(s_t, s)$, the PRMSE for the predictor s_s. Hence we will no longer consider the predictor s in this article. The PRMSE for the predictor s_x can be shown to be equal to $\rho^2(s_t, x)$.

The above discussion implies that for subscores to have added value, the PRMSE has to be larger for s_s than for s_x, that is, $\rho^2(s_t, s)$ has to be larger than $\rho^2(s_t, x)$. This is justifiable from the viewpoint of correlation as well; for a subscore to have added value, it is reasonable to expect that the correlation between the true subscore and the observed subscore should be larger than the correlation between the true subscore and the observed total score.

Equations 3, 5, and 6 suggest that a subscore will be favored as the subscore reliability increases (this will happen when the subscore is based on more number of items), the total score reliability decreases, and the correlation between true subscore and true total score decreases (which will happen when the subtests measure very different skills).

Institutional-Level Analysis

At the institutional level, the above analyses can be modified by decomposition of total scores and subscores into institutional and individual components. Thus subscore s has the decomposition $s = s_I + s_E$, where s_I, the component for the institution, is the same for each examinee in an institution and has mean $E(s)$ and variance $\sigma^2(s_I) > 0$. The component s_E above is an examinee-specific effect that should not be confused with s_e. The score x has the decomposition $x = x_I + x_E$, where x_I, the component for the institution, is the same for each examinee in an institution and has mean $E(x)$ and variance $\sigma^2(x_I) > 0$. The residual examinee subscore $s_E = s - s_I$ within institution has mean 0, variance $\sigma^2(s_E) > 0$, and is uncorrelated with the institutional means s_I and x_I. The residual examinee total score $x_E = x - x_I$ within institution has mean 0, variance $\sigma^2(x_E) > 0$, and is uncorrelated with s_I and x_I. Denote the average observed subscore and the average observed total score for an institution as $\bar{s}$ and $\bar{x}$ respectively. We use an approach similar to that used for examinee-level subscores to determine whether institutional-level subscores have added value.

The predictor of institutional-level true subscores based on the observed subscores is a regression of the institution's true subscore on the institution's average observed subscore, and is given by

$$s_{Is} = E(s) + \rho^2(s_I, \bar{s})[\bar{s} - E(s)]. \quad (8)$$

The predictor of institutional-level true subscores based on the observed total scores is a regression of the institution's true subscore on the institution's average observed total score, and is given by

$$s_{Ix} = E(s) + \rho(s_I, \bar{x})[\sigma(s_I)/\sigma(\bar{x})][\bar{x} - E(x)]. \quad (9)$$

Consider, for example, a 25-item reading subtest whose average score (averaged over all examinees) is 18.

Suppose the total test (that has writing and mathematics subtests besides the reading subtest, and a total of 75 dichotomous items) has an average score (averaged over all examinees) of 50. Consider an institution whose average observed reading subscore is 24 and average observed total test score is 60. Suppose further that $\rho^2(s_I, \bar{s})$ is found to be equal to .90, $\rho^2(s_I, \bar{x})$ is equal to .92, $\sigma(s_I)$ is equal to 1.0 and $\sigma(\bar{x})$ is equal to 3.0. Note that these values are computed using techniques from multivariate analysis of variance technique (for details, see Haberman et al. 2006). For the institution, the predictor s_{Is} predicts the true institutional reading subscore to be

$$18 + .90(24 - 18) = 23.4,$$

while the predictor s_{Ix} predicts the true institutional reading subscore to be

$$18 + \sqrt{.92}\frac{1}{3}(60 - 50) = 21.2.$$

As with examinee-level subscores, we will use the MSE criterion to compare the performance of the predictors. Haberman et al. (2006) showed that the MSE for s_{Is} is $\sigma^2(s_I)[1 - \rho^2(s_I, \bar{s})]$ whereas that for s_{Ix} is given by $\sigma^2(s_I)[1 - \rho^2(s_I, \bar{x})]$.

The trivial predictor of the true subscore for any institution is a constant $E(s)$. The corresponding MSE is $\sigma^2(s_I)$. Hence, relative to use of $E(s)$, the PRMSE for s_{Is} is the institutional subscore reliability $\rho^2(s_I, \bar{s})$ while the PRMSE for s_{Ix} is $\rho^2(s_I, \bar{x})$.

Hence, for the institutional-level subscores to have added value, the PRMSE for s_{Is} has to be larger than that for s_{Ix}, that is, $\rho^2(s_I, \bar{s})$ has to be larger than $\rho^2(s_I, \bar{x})$. Haberman et al. (2006) discussed in detail the use of multivariate analysis of variance to estimate $\rho^2(s_I, \bar{s})$ and $\rho^2(s_I, \bar{x})$, which depend on n, the sample size for the institution.

Results

This section employs the methodology described above to the two data sets mentioned earlier. First, we assess whether the examinee-level subscores are of added value for these data before proceeding to assess whether the institutional-level subscores are of added value.

Examinee-Level Subscores

The reliability of the total score for both the test forms was .94. The first two rows of Tables 1 and 2 display estimated values of the PRMSEs $\rho^2(s_t, s)$ and $\rho^2(s_t, x)$ as percentages rather than proportions. The other rows of the tables will be described later.

Table 1 shows the results for the analysis with six subscores while Table 2 shows the results for three subscores. Remember that for subscores to have added value, $\rho^2(s_t, s)$ has to be larger than $\rho^2(s_t, x)$. However, Tables 1 and 2 indicate that the estimated values of $\rho^2(s_t, x)$ are substantially higher than those of $\rho^2(s_t, s)$, so that no evidence supports reporting of individual level subscores. Note that the subscore reliability is quite high for Table 2 (i.e., for three subscores) so that reporting subscores for the three-subscore case will not cause much harm to the examinees; however, it is also true that the three subscores are redundant for this case because the total score leads to less error than the subscores.

The 6×6 correlation matrix of the individual subscores, considering six subscores, for the two test forms is given as Table 3.

Eigenvalues computed from the estimated 6×6 correlation matrix of the individual subscores (shown in the last row of Table 3) strongly suggest that a single composite score exists that places nearly equal weight to each subscore, so that the composite is very close to the total score. The eigenvalues also suggest that each subscore can be well approximated by use of a linear transformation of the composite score.

For the division of the total score into three subscores, Table 4 provides the estimated 3×3 correlation matrix of the individual subscores for the two test forms. Eigenvalues computed from the 3×3 correlation (also shown in Table 4) suggest the presence of a single composite score.

It is also possible to use another approach to show that the subscores have little added value for these data by computing simple correlations. This approach is related to the development of

Table 1. Estimated Percent Reduction (100 × proportional reduction) in Mean Squared Error for Six Subscores

		Test form 1						Test Form 2					
		Subscore						Subscore					
	N	**1**	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**	**4**	**5**	**6**
$\rho^2(s_t, s)$	-	77	71	77	73	75	74	78	75	79	58	76	75
$\rho^2(s_t, x)$	-	84	91	83	88	81	81	88	91	86	83	83	83
$\rho^2(s_I, \bar{s})$	10	73	74	74	73	69	64	67	63	69	64	45	68
	30	89	89	89	89	87	84	86	84	87	84	71	82
	100	97	97	97	96	96	95	95	95	96	95	89	94
$\rho^2(s_I, \bar{x})$	10	80	80	79	80	80	80	73	73	73	73	73	73
	30	92	92	91	92	92	92	89	89	89	89	89	89
	100	98	98	96	98	98	98	96	96	96	96	96	96

Note. The six subscores correspond respectively to: (i) reading skills, (ii) reading application, (iii) mathematics skills, (iv) mathematics application, (v) writing skills, and (vi) writing application. The first two rows of numbers provide PRMSEs at examinee level; the last six rows provide PRMSEs at institutional level. The institutional size is denoted by n.

Table 2. Estimated Percent Reduction (100 × proportional reduction) in Mean Squared Error for Three Subscores

		Test Form 1			Test form 2		
		Subscore			Subscore		
	n	1	2	3	1	2	3
$\rho^2(s_t, s)$	-	85	85	84	87	84	85
$\hat{\rho}^2(s_t, x)$	-	87	86	87	90	85	87
$\rho^2(s_I, \bar{s})$	10	77	76	72	70	72	59
	30	91	91	89	88	89	81
	100	97	97	96	96	96	93
$\rho^2(s_I, \bar{x})$	10	80	80	80	73	73	73
	30	92	92	92	89	89	89
	100	98	98	98	96	96	96

Note. The three subscores correspond respectively to: (i) reading, (ii) mathematics, and (iii) writing. The first two rows of numbers provide PRMSEs at examinee level; the last six rows provide PRMSEs at institutional level. The institutional size is denoted by n.

the Cronbach α coefficient in terms of split halves of tests (Lord and Novick, 1968, pp. 93–94). We partitioned the first test form (denoted as X) into two subforms: X1 and X2. The subform X1 consists of the 38 odd-numbered items in form X whereas the subform X2 consists of the 37 even-numbered items in form X. Their composition (in terms of the number of items corresponding to each subscore) and difficulty are very similar to that of X, X1 and X2 can be treated as parallel test forms. We denote the scores on reading, mathematics, and writing parts in X1 as R1, M1, and W1, and in X2 as R2, M2, and W2. Figure 1 shows the partitioning.

If the subscores are of added value, we expect that the correlation coefficient between R1 and R2 should be greater than that between R1 and X2, and greater than that between R2 and X1. However, the correlations, shown in Table 5, show the exactly opposite pattern. For example, the correlation between R1 and R2 is .77 while that between R1 and X2 is .80 and that between R2 and X1 is 0.81. Thus, the reading score of an examinee on a test form can be predicted better by her *total* score on a parallel form rather than by her *reading* score on a parallel form. The same is true for the writing and mathematics subscores as well. Though half-tests are less favorable to subscores than the original test, this study of correlations provides a direct illustration of the phenomenon with subscores that do not have added value.

Thus, the demand of Standard 1.12 of the *Standards for Educational and Psychological Testing* (1999) to demonstrate the distinctiveness of the subscores cannot be satisfied for these data. The above results should not come as a surprise, as other studies also found subscores to have little added value for other tests. Harris and Hanson (1991) found subscores of little added value for the English and Mathematics tests from the P-ACT+ examination. Haberman (2005) found subscores to be of no added value for the SAT™ I verbal and mathematics examinations.

Institutional-level Subscores

The value of the PRMSE for institutional-level subscores depends on the institution size n. We performed computations for institution sizes of 10, 30, and 100. Because the maximum institution size observed for our data is 90, the upper bound of 100 appeared reasonable for the application. The lower bound of 10 was chosen following the above-mentioned reporting standards for the test concerned.

Rows 3–8 of Tables 1 and 2 show the PRMSEs (as percentages) for institution sizes of 10, 30, and 100. The computations involved multivariate analysis

Table 3. Estimated Correlations and Eigenvalues for the Six Subscores (Reading Skills, Reading Application, Mathematics Skills, Mathematics Application, Writing Skills, and Writing Application) for the Two Test Forms

	Test form 1						Test Form 2					
	Subscore						Subscore					
Subscore	1	2	3	4	5	6	1	2	3	4	5	6
1	-	.74	.63	.66	.64	.66	-	.76	.70	.59	.67	.69
2	-	-	.65	.67	.66	.66	-	-	70	.57	.69	.69
3	-	-	-	.73	.63	.59	-	-	-	.67	.66	.64
4	-	-	-	-	.63	.64	-	-	-	-	.55	.54
5	-	-	-	-	-	.64	-	-	-	-	-	.69
Eigenvalues	4.3	.45	.39	.36	.26	.26	4.3	.53	.36	.32	.28	.24

Table 4. Estimated Correlations and Eigenvalues for the Three Subscores (Reading, Mathematics, and Writing) for the Two Test Forms

	Test form 1			Test form 2		
	Subscores			Subscores		
Subscore	1	2	3	1	2	3
1	-	.75	.77	-	.76	.79
2	-	-	.74	-	-	.73
Eigenvalues	2.50	.27	.24	2.52	.28	.21

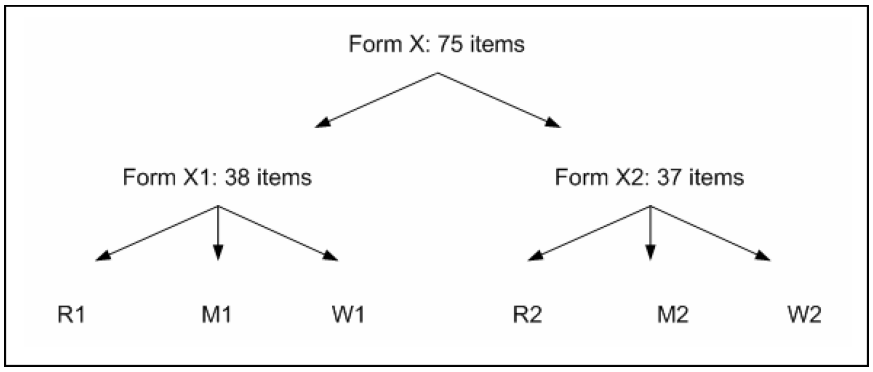

FIGURE 1. Construction of two parallel test forms X1 and X2 from the original test form X.

of variance, using the publicly available SAS software. The tables show that the estimated $\rho^2(s_I, \bar{s})$'s are mostly smaller than the estimated $\rho^2(s_I, \bar{x})$'s, i.e., the institutional-level subscores do not appear to have added value, even for large institution size. It is important to note that the estimated $\rho^2(s_I, \bar{s})$'s are substantially smaller than the estimated $\rho^2(s_I, \bar{x})$'s for institution size 10, which is the lower bound in operational reporting standard for the test. The observed institutional subscore means come close to being of added value for three subscores and at least 100 examinees. Also note that the institutional-level subscores, though redundant given institutional-level total scores, are quite reliable for institution size of 100, so that reporting them will not cause much harm.

The finding that institutional-level subscores are of little added value is not a surprise either. Longford (1990) found college-level subscores to be of little added value for a test he considered.

Discussion and Conclusion

This article analyzes data from an assessment for which there is a constant demand for subscores, and shows that the intended subscores, both at the examinee level and institutional level, do not have any added value for the data. Thus, this article demonstrates that although there is an increasing demand for subscores, the test administrators should exercise caution before reporting subscores and thoroughly evaluate if subscores have any added value over the total scores. As described earlier, inaccurate information at the subscore level can result in negative consequences such as large and needless expenses for individuals, states and institutions. Moreover, where subscores are used either wholly or in conjunction with total test scores to make admissions or hiring decisions, the negative consequences of inaccurate information at the subscore level is even more severe. According to the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), validity refers to the degree to which evidence and theory supports the interpretation of test scores entailed by proposed uses of test scores (1999, p.9). In this context, Kane (2006) points out that the validation of a proposed use of test scores is to evaluate the claims that are made on the basis of these test scores. Therefore, if the claim of test administrators and/or users is that subscores are useful for identifying strengths and weaknesses of test takers in different sub categories (e.g., reading, math, and writing), then evidence should be provided to support such claim/s. Results of this study do not support this claim for the basic skills test considered here—subscores for the test did not provide any added value over the total test score.

Subscores, if defined in raw score units (as in this study), are not directly comparable across different forms of the test. Therefore an important issue with reporting subscores, both for individuals and institutions, is that they have to be equated and/or scaled. In typical cases, equating is possible for the total score but not for subscores. For example, if an anchor test is used to equate the total score, only a few of the anchor test items will correspond to a particular subscore, so that anchor test equating of the subscore will probably not be feasible. When a subscore has few possible score points, appropriate scaling may be a problem. In addition, the possibility exists that a scaling that may be adequate for individuals may be far from satisfactory if applied to institutions.

It is possible to consider at least two additional predictors for both individual-level and institutional-level subscores: (i) a predictor based on both the subscore and the total score (Haberman, 2005), and (ii) an augmented score (Wainer et al., 2001), that is, a predictor based on the subscore of interest and other subscores. For the data employed in the exam-

Table 5. Estimated Correlation Coefficients Between Scores on Test Forms X1 and X2

R1 & R2	0.77	M1 & M2	.72	W1 & W2	.77
R1 & X2	0.80	M1 & X2	.76	W1 & X2	.79
R2 & X1	0.81	M2 & X1	.77	W2 & X1	.79

Note: For notational convenience, the scores on the test forms X1 and X2 are denoted as X1 and X2, respectively, in this table.

ples here, these predictors improved somewhat for the individual-level subscores, but did not improve for the institutional-level subscores (see the details in Haberman et al., 2006). One problem with these two predictors is that they are not easy to explain to the public, who are the end-users of subscores.

Our method has the limitation that it is based on CTT and hence it will not have much appeal to a test for which an IRT model or a cognitive diagnostic model (see, e.g., Haberman & von Davier, 2006) is employed to score examinees. Also, this paper examined only a specific basic skills test and referred to a few other tests—subscores were found to be not of added value for any of these. However, it is possible that for other tests, the results are different. For example, Wainer et al. (2001) and Haberman (2005) each found one test for which subscores were somewhat useful. As Haberman (2005) concludes, "Subscores are most likely to have value if they have relatively high reliability by themselves and if the true subscore and true total score have only a moderate correlation. Both conditions are important." Therefore, it seems more likely for subscores to be useful for tests with reasonably large number of items in each subcategory (which may ensure high subscore reliabilities) and composed of distinct subcategories (which may ensure moderate but not very high correlation of the subscores to the total score). An example that may fit the above description is the Praxis Fundamental Subjects Content Knowledge test (e.g., Grant, 2003) which has 25 items in each of its four distinct subcategories (i.e., English, Mathematics, Social Sciences, and Science).

References

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

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional IRT in test scoring. *Journal of Educational and Behavioral Statistics*, *30*, 295–311.

Dwyer, A., Boughton, K. A., Yao, L., Steffen, M., & Lewis, D. (2006). A comparison of subscale score augmentation methods using empirical data. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Fransisco, CA.

Educational Testing Service (2006). Interpreting the Praxis™ Institutional Summary Report. Princeton, NJ. Retrieved August 15, 2006, from http://www.ets.org/Media/Tests/PRAXIS/pdf/interpreting_praxis_score_reports_51657.pdf.

Grandy, J. (1992). Construct Validity Study of the NTE Core Battery Using Confirmatory Factor Analysis (ETS RR 92-03). Princeton, NJ: Educational Testing Service.

Grant, M. (2003). *Fundamental subjects: Content knowledge (0511), test analysis, form 3ypx1* (Report No. SR-2003-62). Princeton, NJ: ETS.

Haberman, S. J. (2005). *When can subscores have value?* (ETS RR-05-08). Princeton, NJ: Educational Testing Service.

Haberman, S. J., Sinharay, S., & Puhan, G. (2006). *Subscores for institutions*. (ETS RR-06-13). Princeton, NJ: Educational Testing Service.

Haberman, S. J., & von Davier, M. (2006). Some notes on models for cognitively based skills diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics 26* (pp. 1031–1038). Amsterdam: Elsevier North-Holland.

Haladyna, T. M., & 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). *Methods of examining the usefulness of subscores*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Kane, M. T. (2006). Content-related validity evidence. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*, pp. 131–154. Mahwah, NJ: Lawrence Erlbaum Associates.

Kuder, G., & Richardson, M. (1937). The theory of estimation of test reliability. *Psychometrika*, *2*, 151–160.

Longford, N. T. (1990). Multivariate variance component analysis: An application in test development. *Journal of Educational Statistics*, *15*(2), 91–112.

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

Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education*, *17*(2), 89–112.

Wainer, H., Sheehan, K., & Wang, X. (2000). Some paths toward making praxis scores more useful. *Journal of Educational Measurement*, *37*, 113–140.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., Swygert, K. A., & Thissen, D. (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.

Yao, L. H., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, *31*(2), 83–105.

Yen, W. M. (1987). *A Bayesian/IRT index of objective performance*. Paper presented at the annual meeting of the Psychometric Society, Montreal, Quebec, Canada, June 1–19.