International Journal of Learning, Teaching and Educational Research
Vol. 4, No.1, pp. 61-74, April 2014

Comparison and Properties of Correlational and Agreement Methods for Determining Whether or Not to Report Subtest Scores

Oksana Babenko, PhD
W. Todd Rogers, PhD
University of Alberta
Edmonton, Canada

Abstract. Large-scale testing agencies often report subtest scores in addition to reporting the total test score. But is there evidence that subtests reveal differences in student performances? Three methods for determining whether subscore reporting is warranted were examined and evaluated using large-scale data as well as samples of various sizes for Reading and Mathematics assessments. Results revealed that subtests did not differ among themselves and added no value over the total test. The method statistics were determined to be accurate and precise estimators of the population parameters. Implications for subscore reporting are discussed.

Keywords: subscore reporting; accuracy; precision; large-scale assessment

Introduction

Results from a large-scale achievement test can be reported in the form of the total test score and, if justified, as a series of subtest scores together with the total test score. The common practice is to report the total score as a summary of achievement of the total domain tested. However, large-scale testing agencies have increasingly adopted the practice of reporting subtest scores in addition to reporting the total test score because of the potential diagnostic value of subtest scores (Wainer, Sheehan, & Wang, 2000; Tate, 2004; Sinharay, Haberman, & Puhan, 2007; Yao & Boughton, 2007; Sinharay, Puhan, & Haberman, 2009; Sinharay, 2010). Part of the argument put forward to support subscore reporting is based on the fact that the items comprising a test are referenced to a curriculum that is multidimensional in nature, with each dimension characterized by specific content and/or cognitive skills. For example, items on a Mathematics achievement test can be referenced to (a) content areas, such as number sense and numeration, measurement, geometry and spatial sense, patterning and algebra, and data management and probability, and/or (b) cognitive skills, such as knowledge and understanding, application, and problem solving. In test development, the table of specifications serves to ensure that the test reflects the multidimensionality of the curriculum. However, what needs to be recognized is that there must be evidence that the variables or skills measured by the subtests are indeed sufficiently distinct to warrant reporting scores from the subtests. Additionally, while the number of test items typically reflects the proportional weighting given to each cell within the table of specifications, the total number of items included in the test is limited by the amount of available test administration time. Consequently, more often than not the number of items for each dimension in the table of specifications is not sufficient to achieve a high degree of reliability or a low error of measurement.

Despite these cautions, officials at large-scale assessment agencies still want to report subtest scores even though no deliberate attempt was made to ensure that (a) the variables (e.g., number sense and numeration) assessed by subtests are distinct and not highly related, and (b) there is a sufficient number of items for each subtest to ensure high reliability. The evidence that is usually used to determine if the variables are sufficiently distinct is the correlation among the subtest scores whereas the internal consistency of the items in each subtest provides evidence of subscore reliability. What is desired are low subtest correlations and high subtest internal consistencies (American Educational Research Association, American Psychological Association, & National Council on Measurement and Evaluation, 1999; Wainer, Vevea, Camacho, Reeve, et al., 2001; Tate, 2004).

Various methods for determining whether subtest scores are distinct and/or add value over and above the total test score have been developed. These methods include the agreement method (Kelley, 1923; see also Gulliksen, 1951; Lord & Novick, 1968; Ryan, 2003; Haladyna & Kramer, 2004), correlations corrected for attenuation due to unreliability of the measures (McPeek, Altman, Wallmark, & Wingersky, 1976; Harris & Hanson, 1991; Haladyna & Kramer, 2004), factor analytic method (McPeek et al., 1976; Grandy, 1992), statistical model fit (Harris & Hanson, 1991), and, in the case of determining only whether a subtest has value over the total test, the proportional reduction of the mean squared error (Haberman, 2005, 2008; Sinharay, Haberman, & Puhan, 2007). Three of these methods were considered in the present study: Kelley's agreement method (*KR*; Kelley, 1923), correlations corrected for attenuation (${}_{c}\hat{\rho}_{jk}$; McPeek, et al., 1976), and the proportional reduction of the mean squared error (*PRMSE*; Haberman, 2005; Sinharay, Haberman, & Puhan, 2007).

The agreement method takes into account the actual differences between observed scores on subtests *j* and *k* expressed in the same score metric (Kelley, 1923; Gulliksen, 1951; Lord & Novick, 1968; Ryan, 2003; Haladyna & Kramer, 2004). Working with *z*-scores ($\mu = 0; \sigma = 1$) or scores in some other standardized metric to remove the effects of different means and standard deviations of subtests *j* and *k*, the difference, d_i, between two standard scores for student *i* is given by:

$$d_i = z_{ij} - z_{ik},$$

where z_{ij} is the observed standard score of student *i* on subtest *j*, and z_{ik} is the observed standard score of student *i* on subtest *k*. If the estimated standard error of the difference for a given student, $\sqrt{2-\alpha_{jj}-\alpha_{kk}}$, where α_{jj} and α_{kk} are the reliabilities of subtests *j* and *k*, and the estimated standard deviation of the obtained differences for the group of students, $\sqrt{2-2\rho_{jk}}$, where ρ_{jk} is the correlation between the scores on subtests *j* and *k*, are close in value, then the obtained differences are no greater than what would be expected by chance (Kelley, 1923, p. 329). In order to determine directly the percentages of students with differences beyond what would be expected by chance, Kelley computed the proportion of cases in excess of the chance as a function of the ratio:

$$KR = \frac{\sqrt{2-\alpha_{jj}-\alpha_{kk}}}{\sqrt{2-2\rho_{jk}}} \text{ (Kelley, 1923, p. 330).}$$

Kelley (1923) illustrated his agreement method with the eight subtests of the Stanford Achievement Test Battery and found 10% to 44% of the students had differences beyond chance for 36 pairs of subtests (p. 331). Values of *KR* closer to one led to small proportions of students with differences beyond chance; values of *KR* further from one (i.e., closer to zero) led to larger proportions of students with differences beyond chance.

The correlation corrected for attenuation due to unreliability, ${}_c\rho_{jk}$, is given by:

$${}_c\rho_{jk} = \frac{\rho_{jk}}{\sqrt{\alpha_{jj}\alpha_{kk}}},$$

where ρ_{jk} is the uncorrected correlation between the scores on subtests *j* and *k*, and α_{jj} and α_{kk} are the internal consistency estimates (Cronbach, 1951) for subtests *j* and *k*, respectively. If ${}_c\rho_{jk}$ is less than 0.90, then it is concluded that student performances on subtests *j* and *k* differ and that reporting of subtest scores is warranted (McPeek et al., 1976; Haladyna & Kramer, 2004). For example, Haladyna and Kramer (2004) used the ${}_c\rho_{jk}$ method to determine whether subtest scores on a basic biomedical science test revealed any differences in examinees' performances. They found that the corrected correlations were higher than 0.90, suggesting a high degree of similarity in examinees' performances on the subtests of the test.

The proportional reduction of the mean squared error method involves predicting the true scores on subtest *j* from the observed scores on subtest *j* and from the total test score:

$$\tau_{ij} = \mu_j + \alpha_{jj}(s_{ij} - \mu_j) \qquad (1)$$

and

$$\tau_{iX} = \mu_j + \rho_{s_t x}\frac{\sigma_{\tau_j}}{\sigma_X}(x_i - \mu_X), \qquad (2)$$

where τ_{ij} and τ_{iX} are, respectively, the true score for student *i* on subtest *j* when predicted from the observed subtest score and the true score of student *i* on subtest *j* when predicted from the total test *X* score for student *i*; μ_j and μ_X are the means of subtest *j* and the total test *X*;

s_{ij} and x_i are the observed scores on subtest *j* and the total test *X*, respectively, for student *i*;

α_{jj} and α_{XX} are the internal consistencies of subtest *j* and the total test *X*, respectively;

σ_X is the standard deviation of the total test scores, and $\sigma_{\tau_j} = \sigma_j \sqrt{\alpha_{jj}}$, where σ_j is the standard deviation of the scores on subtest *j;* and $\rho_{s_t x} = \sqrt{\rho^2_{s_\tau x_\tau} \alpha_{XX}}$, where $\rho^2_{s_\tau x_\tau}$ is computed as outlined in Haberman (2005). The corresponding mean squared errors (*MSE*) are given by:

$$MSE_{\tau_j / s_j} = \sigma^2_{\tau_j} (1 - \alpha_{jj})$$

and

$$MSE_{\tau_j / X} = \sigma^2_{\tau_j} (1 - \rho^2_{s_j x}) ,$$

where $\sigma^2_{\tau_j}$ is the true score variance for subtest *j*. The *MSE* when τ_j is simply predicted from s_j is the subtest score error variance, $\sigma^2_{e_j} = \sigma^2_j (1 - \alpha_{jj})$.

The proportional reduction of the mean squared errors when the true score is predicted from a subtest score using equation (1) is given for each subtest by:

$$PRMSE_{\tau / s} = \frac{\sigma^2_e - MSE_{\tau / s}}{\sigma^2_e}$$

The *PRMSE* when the true score is predicted from the total test score is computed in the same way but using the $MSE_{\tau / x}$ as the base. If $PRMSE_{\tau / s} > PRMSE_{\tau/x}$, then reporting the scores for subtest *j* adds value over reporting only the total test scores (Haberman, 2005, 2008; Sinharay et al., 2007, 2009; Lyren, 2009; Sinharay, 2010). Haberman (2008) used the *PRMSE* method to determine whether or not the subtest scores on SAT I "had added value over and above the value of the total score" and found that "none of the section scores of SAT I math or SAT I verbal provide any appreciable information concerning an examinee that is not already provided by the math or verbal total score" (p. 221). Using the *PRMSE* method, Sinhary (2010) examined 25 operational tests to see if the subtests within each test had added value over the full test. He found that 16 of the 25 tests had no subtest scores with added value even though subtest scores were reported in many cases. Of the remaining nine tests, some but not all of the subtests had added value. However, it should be noted that in contrast to correlations corrected for attenuation, the *PRMSE* does not compare subtest scores to determine if they are distinct from one another.

In contrast to correlations corrected for attenuation and proportional reduction of the mean squared error methods, which do not specifically look at the agreement between two observed scores obtained from two subtests, the agreement method takes into account the actual differences between observed scores on subtests *j* and *k* expressed in the same score metric. In the case of the ${}_c\rho_{jk}$ method, if differences among the subtests are revealed, then the agreement method will need to be used to determine which students have pairs of scores that differ. In the case of the *PRMSE* method, if a subtest is found to have value over the total test, then the agreement method will need to be used to determine which students have subtest scores that differ from the total test. Thus, it seems reasonable to use the Kelley's agreement method alone.

Hence, one purpose of the present study was to determine whether Kelley's agreement and correlations corrected for attenuation methods would lead to the same or different decision regarding the identification of pairs of distinct subtests and whether Kelley's agreement and proportional reduction of the mean squared error methods would lead to the same or different decision about subtests having added value over the total test. If the decisions were the same in both, then the agreement method could simply be used.

A second purpose of the study was to examine the accuracy and precision of the statistics used in the *KR*, ${}_{c}\rho_{jk}$, and *PRMSE* methods. No studies were found in the published literature that comparatively examined accuracy and precision of the statistics used in these methods. If one method produced biased or imprecise estimates, then different decisions could be made when using samples rather than the population. However, if the method produced unbiased and precise estimates, then the decisions made would not be due to bias or impreciseness.

Method

The two data sets used in the study were population data sets for the Junior (Grade 6) English-language Reading and Mathematics assessments conducted by the Education Quality and Accountability Office (EQAO) in Ontario, Canada (www.eqao.com). EQAO conducts annual province-wide assessments in both of Canada's official languages (English and French) at the Primary (Grade 3) and Junior (Grade 6) levels in the areas of Reading, Writing, and Mathematics, at the Grade 9 level in Academic Mathematics and Applied Mathematics, and at the Grade 10 level in Literacy (Reading and Writing). Results are reported at the provincial, district, school, and student levels and are publically available on the EQAO website, with emphasis on progress from the previous year. EQAO requested that the present study be conducted to determine if reporting subtest scores was justified given no explicit attempt was made to develop subtests with psychometric characteristics that allowed subtest score reporting.

Description of the Reading Test

The English-language Reading test items are referenced to three knowledge and skills categories as specified in the curriculum for the province: *Explicit Information*, *Implicit Information*, and *Connections*. The items in the Explicit Information subtest require students to detect and understand information and ideas stated explicitly in a variety of text types identified in the provincial curriculum. The items in the Implicit Information subtest probe students' understanding of implicitly stated information and ideas. The items in the Connections subtest require students to demonstrate their understanding of text passages by connecting, comparing, and contrasting the ideas presented in the passages and drawing upon their own knowledge, experience and insights, other texts, and the world around them. Thus, the three subtests can be ordered in terms of complexity, with the Explicit Information and Connections subtests at the lowest and the highest levels of complexity, respectively. The Explicit Information subtest contains six multiple-choice items, the Implicit Information subtest contains 14 multiple-choice items and four open-response items, and the Connections subtest contains six multiple-choice items and six open-response items. The 10 open-response items are scored using four-point scoring rubrics.

Description of the Mathematics Test

In contrast to the Reading assessment, the items on the Mathematics assessment are referenced by content areas (i.e., strands) and by cognitive skills as specified in the mathematics curriculum. The five content areas include: *Number Sense and Numeration* (8 items involving estimation, rate, ratio, and use of fractions), *Measurement* (8 items involving the use of area relationships, understanding of the dimensions of the shapes needed to calculate their areas, and the conversion of metric area units), *Geometry and Spatial Sense* (6 items dealing with the identification, performance and description of transformations, the identification of angles, and accurate use of rulers and protractors), *Patterning and Algebra* (7 items dealing with growing patterns, use of diagrams, tables and number sequences to represent the stages of patterns), and *Data Management and Probability* (7 items involving concepts of probability, predicting and representing the probability of an outcome, comparing probabilities using common representations (e.g., common denominators, percents or decimals), and interpreting graphs). The five content areas are not ordered in terms of complexity.

Cognitive skills are divided into three categories: *Knowledge and Understanding* (8 items), *Application* (15 items), and *Problem Solving* (13 items). The items referenced to the Knowledge and Understanding category require students to demonstrate subject specific content (knowledge) and the comprehension of its meaning and significance (understanding). The Application items require students to select and fit an appropriate mathematical tool or get the necessary information. The Problem Solving items require students to select and sequence a variety of tools to solve a problem and demonstrate a critical-thinking process. That is, to answer Problem Solving items, students need to make a plan. In contrast to the content subtests, the cognitive subtests can be ordered in terms of complexity, with the Knowledge and Understanding subtest and the Problem Solving subtest being at the lowest and the highest levels of complexity, respectively. The total number of items on the Mathematics assessment is 36, including 8 open-response items scored using a four-point scoring rubric and distributed such that each content subtest has at least one open-response item.

Analyses

The analyses were conducted in two main stages corresponding to the two purposes of the study. First, the responses of the population of students were analysed to obtain the population value of each test statistic for each of the three detection methods. Following this, the analyses were repeated for 1,000 replicated independent samples of five different sizes – 250, 500, 1,000, 2,000, and 5,000 – randomly drawn from the population with replacement to (1) determine the effect of sample size on the accuracy and precision of the estimators, and then to (2) assess the consistency of the decisions made using the three detection methods in light of the findings about accuracy and precision. At the first stage, means and standard deviations of the distributions of sample statistics were used to evaluate the three detection methods with respect to their accuracy and precision. At the second stage, the *KR* and ${}_{c}\rho_{jk}$ methods were applied, first, at the population level to determine if the subtests were distinct, and then applied to each of 1,000 replicated samples for each of the five sample sizes to see if the same decision was made. The *PRMSE* method was applied at the population level and then for each replicated sample to see if the subtests added value over the total test. The consistency of the decisions made was assessed using the percentage of samples that led to the same decision that was made at the population level.

Results and Discussion

Psychometric Properties of the Reading and Mathematics Tests

The psychometric properties of the Junior Reading and Mathematics subtests and the total tests are provided in Table 1 for the population of students. The means and standard deviations are reported in the observed score units and as percentages (in parentheses).

Table 1. Means, Standard Deviations, and Internal Consistencies for Reading and Mathematics Tests

Subtest/Total Test	*k/ms*[a]	$\bar{X}$	s_X	Correlations					
Reading, *N* = 128,089				EI	II	C	TT		
Explicit Information (EI)	6/6	4.59 (76.5)	1.28 (21.3)	*0.47*[b]	0.59	0.52	0.69		
Implicit Information (II)	18/30	20.92 (69.7)	4.42 (14.7)		*0.76*	0.74	0.93		
Connections (C)	12/30	17.16 (57.2)	4.33 (14.4)			*0.74*	0.92		
Total Test (TT)	36/60	42.68 (64.7)	8.97 (13.6)				*0.87*		
Mathematics, *N* = 127,596									
Content Area				N	M	A	P	G	TT
Numeration (N)	8/14	8.44 (60.3)	3.02 (21.6)	0.63	0.66	0.62	0.67	0.63	0.87
Measurement (M)	8/11	6.34 (57.6)	2.67 (24.3)		*0.63*	0.59	0.63	0.63	0.84
Algebra (A)	7/10	6.63 (66.3)	2.14 (21.4)			*0.58*	0.62	0.59	0.80
Probability (P)	7/13	7.20 (55.4)	2.71 (20.8)				*0.61*	0.62	0.85
Geometry (G)	6/12	7.18 (59.8)	2.80 (23.3)					*0.60*	0.83
Total Test (TT)	36/60	35.79 (59.7)	11.20 (18.7)						*0.89*
Cognitive Skill				K/U	A	PS	TT		
Know/Understand (K/U)	8/8	5.45 (68.0)	1.87 (23.4)	*0.60*	0.70	0.67	0.80		
Application (A)	15/24	14.98 (62.4)	4.91 (20.5)		*0.75*	0.79	0.94		
Problem Solving (PS)	13/28	15.36 (54.8)	5.42 (19.4)			*0.78*	0.94		
Total Test (TT)	36/60	35.79 (59.7)	11.20 (18.7)				*0.89*		

[a] *k* is number of items in a subscale or the total test and *ms* is the maximum score given the use of dichotomously scored multiple-choice items and polytomously scored open-response items.

[b] Internal consistencies of the subtests and the total test are shown in italics along the principal diagonal of each correlation panel.

Reading. The mean percentages revealed that students' performance declined on the original three subtests as the complexity of the constructs increased. The standard deviations (percentages) were essentially the same for the Implicit Information and Connections subtests, which are at the two higher levels of complexity, but smaller than for the Explicit Information subtest, likely because of the smaller number of items and, therefore, total points for this subtest. As shown along the main diagonal of the correlation matrix on the right side of Table 1, the internal consistency (alpha; Cronbach, 1951) of the Explicit Information subtest was much lower than the internal consistencies of the Implicit Information and Connections subtests, which were essentially the same. The low reliability of the Explicit Information is due to the relatively small number of items (6) in this subtest in comparison to the other subtests (18 and 12, respectively). The estimate of the internal consistency of the total test was 0.87, reflecting the typical practice mentioned above of ensuring that the total test reliability is at an acceptable high level. The values of the correlations were either greater than the corresponding reliabilities or close in value, which suggests that the three procedures examined in this study will show that the subtests are not distinct and the subtests do not add value over and above the total test.

Mathematics content area. The mean percentages revealed that the mean for the Algebra subtest was the highest, the mean on the Probability subtest was the lowest, and the means of the other three subtests were between and essentially the same. The standard deviations were somewhat larger for the Measurement and Geometry subtests than the standard deviations for the Numeration, Algebra, and Probability subtests, which were essentially the same. Given the numbers of items in each subtest did not differ much as they did in the case of Reading and Mathematics cognitive skills, the internal consistencies of the five content subtests were essentially the same, ranging from 0.58 to 0.63. However, as with Reading, the values of the correlations were close to the values of the reliabilities, suggesting again that the three procedures examined in this study will show that the subtests are not distinct and the subtests do not add value over and above the total test. Further, some of the values of *KR* and ${}_{c}\rho_{jk}$ will exceed 1.00, which theoretically should not happen.

Mathematics cognitive skills. Similar to Reading, the students' performance on the three mathematics cognitive subtests declined as the level of required thinking increased from knowledge and understanding to application to problem solving. The standard deviations were essentially the same for the Application and Problem Solving subtests, which are of higher complexity, but smaller than the standard deviation for the knowledge and understanding subtest, again likely because of the smaller number of items in the latter subtest. The internal consistency of the knowledge and understanding subtest, 0.60, was lower than the internal consistencies of the Application and Problem Solving subtests, which were more alike, 0.75 and 0.78, respectively. The somewhat low value of reliability for the knowledge and understanding subtest was likely due to the relatively smaller number of items (8) in this subtest as compared to the numbers of items in the other two subtests (15 and 13, respectively). The estimate of the internal consistency of the total test was 0.89, again reflecting the typical practice mentioned above of ensuring that the total test reliability is at an acceptable high level. Again, we see as for the Reading and Mathematics content areas that the values of the correlations are close to the values of the reliabilities, suggesting that the three procedures examined in this study will show that the subtests are not distinct and the subtests do not add value over and above the total test, with some of the values of KR and ${}_{c}\rho_{jk}$ will exceed 1.00.

Accuracy and Precision of the Estimators

The means and standard deviations of the 1,000 replications for each sample size are reported in Table 2 for the *KR* method, Table 3 for the ${}_c\rho_{jk}$ method, and Table 4 for the *PRMSE* method. As shown in Table 2, all but two of the means of the sampling distributions of 1,000 replications across the pairs and sample sizes were within 0.01 of the corresponding population value of *KR* for each subtest pair and sample size (the difference is 0.02 for the Numeration and Probability subtest pair, with n = 250 and n = 500). The standard error of *KR* decreased as the sample size increased. For example, for n = 250 the standard errors were between 0.043 and 0.049, whereas for n = 5,000 the standard errors were between 0.009 and 0.011.

Table 2. Accuracy and Precision: Kelley's Ratio (KR)

	Sample Size					
Subtest Pairs	250	500	1,000	2,000	5,000	Population
Reading						
Exp Info–Imp Info	0.97 (0.049)[a]	0.97 (0.035)	0.97 (0.024)	0.97 (0.017)	0.97 (0.011)	0.97
Exp Info–Con	0.92 (0.046)	0.91 (0.032)	0.91 (0.022)	0.91 (0.015)	0.91 (0.010)	0.91
Imp Info–Con	0.96 (0.043)	0.95 (0.032)	0.95 (0.021)	0.95 (0.015)	0.95 (0.009)	0.96
Mathematics Content Area						
Num–Mea	1.04 (0.051)	1.04 (0.035)	1.04 (0.025)	1.04 (0.017)	1.04 (0.011)	1.04
Num–Alg	1.03 (0.050)	1.03 (0.034)	1.03 (0.025)	1.03 (0.016)	1.03 (0.011)	1.02
Num–Prob	1.09 (0.050)	1.09 (0.037)	1.08 (0.026)	1.08 (0.026)	1.08 (0.011)	1.07
Num–Geo	1.03 (0.049)	1.03 (0.037)	1.03 (0.025)	1.02 (0.017)	1.02 (0.011)	1.02
Mea–Alg	0.98 (0.045)	0.98 (0.033)	0.98 (0.023)	0.98 (0.016)	0.98 (0.010)	0.98
Mea–Prob	1.02 (0.046)	1.02 (0.032)	1.02 (0.024)	1.02 (0.017)	1.02 (0.010)	1.01
Mea–Geo	1.02 (0.052)	1.03 (0.035)	1.02 (0.026)	1.02 (0.017)	1.02 (0.011)	1.02
Alg–Prob	1.03 (0.049)	1.03 (0.033)	1.03 (0.024)	1.03 (0.017)	1.03 (0.010)	1.03
Alg–Geo	1.00 (0.047)	1.00 (0.033)	1.00 (0.023)	1.00 (0.017)	1.00 (0.010)	1.00
Prob–Geo	1.03 (0.048)	1.03 (0.034)	1.03 (0.023)	1.03 (0.017)	1.03 (0.011)	1.02
Mathematics Cognitive Skill						
Kno/Und–App	1.04 (0.052)	1.04 (0.035)	1.04 (0.026)	1.03 (0.017)	1.04 (0.011)	1.04
Kno/Und–Prob Sol	0.98 (0.049)	0.98 (0.031)	0.98 (0.024)	0.97 (0.018)	0.97 (0.011)	0.97
App–Prob Sol	1.05 (0.048)	1.05 (0.035)	1.05 (0.024)	1.05 (0.018)	1.05 (0.011)	1.05

[a] The first value is the mean and the value in parentheses is the standard deviation of the sampling distribution (i.e., standard error) of 1,000 replications.

The means of the sampling distributions of ${}_c\hat{\rho}_{jk}$ were within 0.01 of the corresponding population values of ${}_c\rho_{jk}$ for all the pairs of subtests and sample sizes (see Table 3). The standard errors of sample estimators decreased as the sample size increased. For n = 250, the standard errors ranged between 0.029 and 0.081, whereas for n = 5,000, the standard errors were as low as 0.007 and as high as 0.017. Given the low reliability of the Explicit Information subtest in the Reading assessment, the standard errors for the pairs involving this subtest were consistently higher than the standard errors for the remaining pairs of subtests.

Similar to *KR* and ${}_c\hat{\rho}_{jk}$, the means of the distributions of sample estimators of $PRMSE_{\tau/s}$ and $PRMSE_{\tau/x}$ were within 0.01 of the corresponding population values for all four subtests (Table 4). The standard errors of sample estimators were the largest when the Explicit Information subtest was considered (e.g., the standard error $PRMSE_E = 0.100$ for n = 250) but decreased as the sample size increased, ranging between 0.003 and 0.020 for n = 5,000. Taken together, the results provided in Tables 2, 3, and 4 reveal that sample estimates of KR, ${}_c\rho_{jk}$, and $PRMSE_{\tau/s}$ and $PRMSE_{\tau/x}$ are accurate and precise. Therefore, any differences among the three detection methods used for the detection of subtest differences or subtest-total test differences are not confounded by presence of biased or imprecise estimators.

Table 3. Accuracy and Precision: Correlation Corrected for Attenuation (${}_c\rho_{jk}$)

	Sample size					Popu-
Subtest Pairs	250	500	1,000	2,000	5,000	lation
Reading						
Exp–Imp Info	0.99 (0.078)[a]	0.99 (0.051)	0.99 (0.037)	0.98 (0.026)	0.98(0.016)	0.98
Exp Info–Con	0.90 (0.081)	0.89 (0.056)	0.89 (0.039)	0.89 (0.017)	0.89 (0.017)	0.89
Imp Info–Con	0.97 (0.029)	0.97 (0.021)	0.97 (0.015)	0.97 (0.011)	0.97 (0.007)	0.97
Mathematics Content Area						
Num–Mea	1.05 (0.054)	1.05 (0.037)	1.05 (0.026)	1.05 (0.018)	1.04 (0.012)	1.04
Num–Alg	1.04 (0.062)	1.04 (0.042)	1.04 (0.031)	1.03 (0.020)	1.03 (0.014)	1.03
Num–Prob	1.09 (0.051)	1.09 (0.038)	1.09 (0.027)	1.09 (0.018)	1.09 (0.012)	1.09
Num–Geo	1.03 (0.058)	1.03 (0.040)	1.03 (0.029)	1.03 (0.021)	1.03 (0.013)	1.03
Mea–Alg	0.98 (0.062)	0.97 (0.045)	0.97 (0.032)	0.97 (0.022)	0.97 (0.014)	0.97
Mea–Prob	1.02 (0.055)	1.02 (0.038)	1.02 (0.028)	1.02 (0.020)	1.02 (0.013)	1.02
Mea–Geo	1.02 (0.062)	1.03 (0.041)	1.03 (0.031)	1.03 (0.020)	1.03 (0.013)	1.03
Alg–Prob	1.04 (0.064)	1.04 (0.043)	1.04 (0.030)	1.04 (0.022)	1.04 (0.013)	1.04
Alg–Geo	0.99 (0.066)	1.00 (0.046)	0.99 (0.033)	0.99 (0.023)	0.99 (0.014)	0.99
Prob–Geo	1.04 (0.059)	1.03 (0.042)	1.03 (0.029)	1.03 (0.021)	1.03 (0.014)	1.03
Mathematics Cognitive Skill						
Kno/Und–App	1.04 (0.047)	1.04 (0.031)	1.04 (0.023)	1.04 (0.016)	1.04 (0.010)	1.04
Kno/Und–Prob Sol	0.99 (0.048)	0.99 (0.031)	0.98 (0.023)	0.98 (0.016)	0.98 (0.010)	0.98
App–Prob Sol	1.03 (0.026)	1.03 (0.018)	1.03 (0.013)	1.03 (0.009)	1.03 (0.006)	1.03

[a] The first value is the mean and the value in parentheses is the standard deviation of the sampling distribution (i.e., standard error) of 1,000 replications.

Detection of Performance Differences and Consistency of Decisions

As foreshadowed in the presentation of the psychometric properties of the subtests and total test and as revealed by the results in Tables 2, 3, and 4, the subtests were determined to be not distinct nor did the subtests add value over the total test. The values of *KR* were close to one with one exception (Reading, Explicit Information and Connections subtests; Table 2). Further, 11 of the 16 *KR* values exceeded one, which theoretically should not happen. For the agreement procedure to work, the sum of the reliabilities of the two subtests has to be greater than two times the correlation between the two subtests being compared. This was not the case with the subtests considered in the present study, with the sum of the reliabilities in the 11 cases being less than two times the corresponding correlations.

The decision rule for the method of correlations corrected for attenuation is a value less than 0.90 indicates that the two subtests being correlated are sufficiently different to warrant reporting the scores on each (McPeek et al., 1976). With one possible exception (Reading, Explicit Information and Connections subtests; Table 3), the values of ${}_c\hat{\rho}_{jk}$ exceeded 0.95, with 10 of the 16 values being greater than 1.00, which theoretically should not happen. The decision rule for the *PRMSE* method is: if $PRMSE_{\tau/s} > PRMSE_{\tau/x}$, then the subtest has added value over and above the total test and, therefore, the score on the subtest should be reported. As shown in Table 4, for all subtests, $PRMSE_{\tau/s} < PRMSE_{\tau/x}$. For both the ${}_c\rho_{jk}$ and *PRMSE* methods, the reliabilities of the subtests must be high, which was not the case in the present study.

In the case of Reading, with perhaps one exception, the decisions made using population values of *KR* and ${}_c\rho_{jk}$ were that the subtests did not differ, and the population values of $PRMSE_{\tau/s}$ and $PRMSE_{\tau/x}$ indicated that the three subtests did not add value over the total test. For the Explicit Information and Connections pair of subtests, *KR* suggested that there was a difference beyond chance for 5% of the students, and that the value of ${}_c\rho_{jk}$, 0.89, was just less than 0.90. The sample data revealed that with exception of two subtest pairs, Explicit Information and Connections and Explicit Information and Implicit Information with *n* = 250 and *n* = 500, the same decision was made using sample data for at least 91% of the replications using the *KR*, ${}_c\rho_{jk}$, and *PRMSE* methods across the different sample sizes. In the case of the Explicit Information and Connections pair, the decision consistency for ${}_c\hat{\rho}_{jk}$ varied from 51.4% to 78.8% across the five sample sizes (i.e., 514 of the 1,000 replications led to the same decision made at the population level). This finding is attributable to the low reliability of the Explicit Information subtest, 0.47, and the observation that the value of ${}_c\rho_{jk}$ was only 0.01 below the decision value of 0.90. In the case of the Explicit and Implicit Information pairs, the decision consistency for *n* = 250 was 90.5% and for *n* = 500, 95.1%, while for $n \geq 1,000$ the decision consistencies were 99.1%, 100%, and 100%.

In the case of Mathematics, *KR* and ${}_c\rho_{jk}$ indicated that there were no distinct subtests and $PRMSE_{\tau/s}$ and $PRMSE_{\tau/x}$ indicated that no subtest added value over the total test. Further, the majority of values for *KR* were greater than 1.00 due to the fact that the sum of the reliabilities was greater than two times the uncorrected correlation. Similarly, the majority of the values for ${}_c\hat{\rho}_{jk}$ were greater than 1.00 due to the fact that the square root of the product of the reliabilities was less than the uncorrected correlation between the pairs of subtests. The sample data revealed that, with three exceptions, Measurement and Algebra with *n* = 250 and *n* = 500 and Algebra and Geometry with *n* = 250, the same decision was made using sample data for at least 97% of replications using the *KR* and ${}_c\rho_{jk}$ methods and, in the case of the subtest-total test pairs, the *PRMSE* method. The exceptions included the Algebra subtest, which had the lowest reliability out of the all Mathematics subtests. Again, as for Reading, the sample values of the correlations were close to the sample values of the reliabilities, and in the majority of cases the two times the correlation exceeded the sum of the reliabilities, leading to sample estimates greater than 1.00.

Table 4. Accuracy and Precision: Proportional Reduction of the Mean Squared Error (PRMSE)

		Sample Size					Popu-
Subtest/*PRMSE*		250	500	1,000	2,000	5,000	lation
Reading							
Exp Info	$PRMSE_{\tau/s}$	0.47 (0.054)[a]	0.47 (0.038)	0.47 (0.027)	0.47 (0.020)	0.47 (0.012)	0.47
	$PRMSE_{\tau/x}$	0.80 (0.100)	0.79 (0.065)	0.79 (0.047)	0.79 (0.033)	0.79 (0.020)	0.79
Imp Info	$PRMSE_{\tau/s}$	0.76 (0.023)	0.76 (0.016)	0.77 (0.011)	0.77 (0.008)	0.77 (0.005)	0.77
	$PRMSE_{\tau/x}$	0.87 (0.017)	0.87 (0.013)	0.87 (0.008)	0.87 (0.006)	0.87 (0.004)	0.87
Con	$PRMSE_{\tau/s}$	0.73 (0.023)	0.73 (0.016)	0.73 (0.011)	0.74 (0.008)	0.74 (0.005)	0.74
	$PRMSE_{\tau/x}$	0.85 (0.020)	0.85 (0.014)	0.85 (0.010)	0.85 (0.007)	0.85 (0.004)	0.85
Mathematics Content Area							
Num	$PRMSE_{\tau/s}$	0.63 (0.028)	0.63 (0.019)	0.63 (0.015)	0.63 (0.010)	0.63 (0.006)	0.63
	$PRMSE_{\tau/x}$	0.94 (0.037)	0.94 (0.025)	0.94 (0.019)	0.94 (0.012)	0.94 (0.008)	0.94
Mea	$PRMSE_{\tau/s}$	0.63 (0.029)	0.63 (0.024)	0.63 (0.015)	0.63 (0.010)	0.63 (0.006)	0.63
	$PRMSE_{\tau/x}$	0.90 (0.042)	0.90 (0.031)	0.90 (0.021)	0.89 (0.014)	0.89 (0.009)	0.89
Alg	$PRMSE_{\tau/s}$	0.58 (0.038)	0.58 (0.025)	0.58 (0.019)	0.58 (0.013)	0.58 (0.008)	0.58
	$PRMSE_{\tau/x}$	0.89 (0.055)	0.88 (0.039)	0.88 (0.027)	0.88 (0.019)	0.88 (0.012)	0.88
Prob	$PRMSE_{\tau/s}$	0.61 (0.032)	0.61 (0.022)	0.61 (0.019)	0.61 (0.011)	0.61 (0.007)	0.61
	$PRMSE_{\tau/x}$	0.94 (0.042)	0.93 (0.029)	0.93 (0.020)	0.93 (0.015)	0.93 (0.009)	0.93
Geo	$PRMSE_{\tau/s}$	0.60 (0.033)	0.60 (0.024)	0.60 (0.017)	0.60 (0.011)	0.60 (0.007)	0.60
	$PRMSE_{\tau/x}$	0.90 (0.045)	0.90 (0.031)	0.90 (0.023)	0.90 (0.016)	0.90 (0.010)	0.90
Mathematics Cognitive Skill							
Kno/ Und	$PRMSE_{\tau/s}$	0.60 (0.038)	0.60 (0.026)	0.60 (0.019)	0.60 (0.013)	0.60 (0.008)	0.60
	$PRMSE_{\tau/x}$	0.90 (0.055)	0.89 (0.036)	0.89 (0.027)	0.89 (0.018)	0.89 (0.012)	0.89
App	$PRMSE_{\tau/s}$	0.75 (0.019)	0.75 (0.013)	0.75 (0.010)	0.75 (0.007)	0.75 (0.004)	0.75
	$PRMSE_{\tau/x}$	0.91 (0.016)	0.91 (0.011)	0.91 (0.008)	0.91 (0.005)	0.91 (0.004)	0.91
Prob Sol	$PRMSE_{\tau/s}$	0.78 (0.017)	0.78 (0.012)	0.78 (0.009)	0.78 (0.006)	0.78 (0.004)	0.78
	$PRMSE_{\tau/x}$	0.89 (0.014)	0.90 (0.010)	0.90 (0.007)	0.90 (0.005)	0.90 (0.003)	0.90

[a] The first value is the mean and the value in parentheses is the standard deviation of the sampling distribution (i.e., standard error) of 1,000 replications.

Taken together, the results provided in Tables 2, 3, and 4 revealed that there were no differences among the abilities of the three detection methods to detect subtest differences or subtest-total test differences. Kelley's agreement and correlations corrected for attenuation methods led to the same decisions regarding the identification of pairs of distinct subtests. Likewise, Kelley's agreement and proportional reduction of the mean squared error methods led to the same decisions about subtests having added value over the total test. Specifically, the decisions were that the subtests did not differ among themselves and the subtests did not add value over the total test.

Conclusion and Implications for Practice

Whether or not to report subtest results is an important topic that has immediate practical implications. Given a profile of subtest scores, teachers and school counsellors can identify areas of strength and areas that need to be addressed for individual students. Similarly, changes in curriculum and instruction designed to maintain strength and address issues at the school and class levels can be made to improve student learning and achievement.

Subscore reporting will most likely be enhanced if subtests are specifically developed to measure a multidimensional construct or domain. The subdomains to be assessed must be clearly defined and, if supportable, weakly to moderately correlated. The number of items used to assess each dimension or subdomain must be large enough to ensure an adequate level of reliability. The correlations between the subtests examined in the present study were moderate to moderately strong and the reliabilities of the subtests were not high, resulting in reliabilities and correlations being similar in value.

But it seems reasonable to assume that the values of the correlations for the pairs of subtests in the present study are likely to be found in other large-scale assessments of achievement in the school system. Consequently, given this assumption, it is necessary to increase the reliabilities of the subtests. For example, assuming the median observed correlation among Mathematics content subtests in the present study, 0.63, the percentage of students who would be identified with subtest differences beyond chance using the agreement method would be approximately 5% if the reliability of the two subtests was 0.70, 9% if the reliability of the two subtests was 0.75, 15% if the reliability of the two subtests was 0.80, and 20% if the reliability of the two subtests was 0.85. Likewise, for the correlations corrected for attenuation and the proportional reduction of the mean squared error methods, pairs of subtests are most likely to be found distinct and subtests are most likely to have value over and above the total test if the subtests have relatively high reliabilities and the true subtest scores and the true total scores have only moderate correlations. The results for replicated random samples (*n* = 250, 500, 1,000, 2,000, and 5,000) revealed that the statistics for the three detection methods were accurate and precise estimators of the corresponding population parameters.

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.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297–334.

Grandy, J. (1992). Construct validity study of the NTE core battery using confirmatory factor analysis. (ETS Research Report No. RR-92-03). Princeton, NJ: Educational Testing Service.

Gulliksen, H. (1950, 1967). Theory of mental tests. New York: John Wiley & Sons, Inc.

Haberman, S. J. (2005). When can subscores have value? (ETS Research Report No. RR-05-08). Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2008). Subscores and validity. (ETS Research Report No. RR-08-64). Princeton, NJ: Educational Testing Service.

Haladyna, T. M. & Kramer, G. A. (2004). The validity of subscores for a credentialing test. Evaluation and the Health Professions, 27, 349–368.

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

Kelley, T. L. (1923). A new method for determining the significance of differences in intelligence and achievement scores. Journal of Educational Psychology, 14, 300–303.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. New York: Addison–Wesley.

Lyrén, P. E. (2009). Reporting subscores from college admission tests. Practical Assessment, Research and Evaluation, 14(4), 1–10.

McPeek, M., Altman, R., Wallmark, M., & Wingersky, B. C. (1976). An investigation of the feasibility of obtaining additional subscores on the GRE Advanced Psychology Test (GRE Board Professional Report No. 74 - 4P). Princeton, NJ: Educational Testing Service. (ERIC Document No. ED163090)

Ryan, J. (2003). An analysis of item mapping and test reporting strategies. Greensboro, NC: South Carolina Department of Education.

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

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

Sinharay, S., Puhan, G., & Haberman, S. (2009). Reporting diagnostic scores: Temptations, pitfalls, and some solutions. Paper presented at the National Council on Measurement in Education, San Diego, CA, USA.

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

Wainer, H., Sheehan, K. M., & 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 Test Scoring (pp. 343–387). Mahwah, NJ: Lawrence Erlbaum Associates.

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