Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 20, Number 21, November 2015 ISSN 1531-7714

Methods for Examining the Psychometric Quality of Subscores: A Review and Application

Jonathan Wedman, Per-Erik Lyrén, *Umeå University, Sweden*

When subscores on a test are reported to the test taker, the appropriateness of reporting them depends on whether they provide useful information above what is provided by the total score. Subscores that fail to do so lack adequate psychometric quality and should not be reported. There are several methods for examining the quality of subscores, and in this study seven such methods, four of which are based on classical test theory and three of which are based on item response theory, were reviewed and applied to empirical data. The data consisted of test takers' scores on four test forms – two administrations of a first version of a college admission test and two administrations of a second version – and the analyses were carried out on the subtest and section levels. The two section scores were found to have adequate psychometric quality with all methods used, whereas the results for subtest scores ranged from almost all scores having adequate psychometric quality to none having adequate psychometric quality. The authors recommend using Haberman's method and the related utility index because of their solid theoretical foundation and because of various issues with the other subscore quality methods.

A test score is intended to reflect the test takers' knowledge in the domain purportedly measured by the test. *The Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014) states that in a situation where a test score is reported back to the test taker, those responsible for the testing programs should provide appropriate interpretations of the score. The same is true for subscores, which are scores derived from any subset of a test. In recent years, there has been an increasing interest in examining subscores, in terms of their psychometric quality and potential usefulness when they are reported to test takers (e.g. Haberman, 2008; Ling, 2012; Sinharay, 2010; Stone, Ye, Zhu & Lane, 2010).

When subscores are reported, the appropriate interpretations of them should be supported by relevant evidence, but the question is what can be considered relevant evidence. The *Standards* (2014) has a few passages that are relevant for all types of subscores. First, it says that "When interpretation of subscores … is suggested, the rationale and relevant evidence in support of such interpretation should be provided." (from Standard 1.14, p. 27), and that "when a test provides more than one score, the distinctiveness and reliability of the separate scores should be demonstrated" (comment to Standard 1.14, p. 27). Second, Standard 2.3 states that "For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported." (p. 43). That is, for subscores to be reported, evidence of their distinctiveness (referred to as "subscore orthogonality" by Feinberg & Wainer, 2014a) and reliability must have been demonstrated. Subscores that are both sufficiently distinct and reliable are said to have *adequate psychometric quality* (e.g. Sinharay, Puhan, & Haberman, 2011). The issue of whether subscores have adequate psychometric quality or not is, on a test taker level, primarily a

concern when the information is to be used as a basis for remedial study decisions (Monaghan, 2006) or is to be used for high-stakes decisions such as certification, placement, or college admission. Consequently, it is important to find empirical evidence for the adequate psychometric quality of subscores. In this study, we review different methods to find such evidence.

Purpose

The purpose of this study is to review different methods that are used for examining the psychometric quality of subscores, and to provide an empirical application for each of these methods. We also want to examine to what extent the conclusions that can be drawn when applying the different methods to operational tests are in agreement with each other.

The remainder of the paper is structured as follows. We first present a review with information about, and previous applications of, methods for examining subscore quality. This is followed by the empirical application, starting with the method section, which provides information about the tests, participants, data, and software. The results are then presented separately for each method. Finally, in the discussion section, the consequences of the results, the limitations of the study, a recommendation on which method to use when examining subscore quality, and some thoughts on further research are presented.

Descriptions and Previous Applications of Methods for Examining Subscore Quality

Sinharay, Puhan, and Haberman (2011) provided an introduction to, and review of, several methods for examining the psychometric quality of subscores, and given the comprehensiveness of their review we have based our own review on these methods and applied them on empirical data. The methods they discussed were factor analysis, the beta-binomial model, multidimensional item response theory (MIRT), DIMTEST, DETECT, and Haberman's (2008) method based on classical test theory (CTT). Unfortunately, the beta-binomial model, proposed by Lord (1965) and used by Hanson (1989) and Harris and Hanson (1991), is applicable only to tests with two subscores and we therefore chose to not include it in this study. However, in addition to these methods we also applied the utility index and subscore augmentation. Subscore augmentation is not a method for examining the psychometric quality of subscores per se. Instead, it serves as a means of increasing the psychometric quality of the subscores in situations where observed subscores are deemed insufficient.

The methods have not been explicitly referred to as methods for examining the psychometric quality of subscores. Instead, different terms such as added value and usefulness have been used. In this study, we use all these terms interchangeably.

Haberman's Method

Haberman (2008) proposed a CTT-based method for examining whether subscores have what he called *added value* over total scores. This method is based on the concept that there is value in reporting a certain subscore if the observed subscore is a more reliable predictor of the true subscore than the observed total score is. Assuming that S_s and S_x are the relevant predictors based on the subscore and the total score, respectively, Haberman suggested using the proportional reduction in mean squared error (PRMSE) of the predictors compared to the mean squared error of the trivial predictor as a criterion for comparing predictors of true subscores. The PRMSE for the predictor S_s, PRMSE_s, is simply the subscore reliability and the PRMSE for the predictor S_x, PRMSE_x, is a quantity that can be thought of as the reliability of the observed total score as an estimate of the true subscore. For a subscore to have added value, therefore, PRMSE_s must be larger than PRMSE_x. If, for example, a subscore has a PRMSE_s of 0.63 and a PRMSE_x of 0.57, the subscore is considered to have added value. If a subscore has a PRMSE_s of 0.82 and a PRMSE_x of 0.84, the subscore is considered to lack added value. For more details about computation of the PRMSE's, see Haberman (2008) or Haberman, Sinharay, and Puhan (2009).

Haberman (2008) applied his method to SAT I data from 2002 and found that none of the subtest scores had added value but that both section scores did. Haberman (2008) concluded that the relative value of subscores increases when their reliability increases, when the reliability of the total score decreases, and when the correlation between the true subscores and the true total score decreases. A simulation study by Sinharay (2010) provided more details concerning the

extent to which reliabilities and correlations affect subscore value.

Sinharay, Haberman and Puhan (2007) examined a test for prospective and practicing teacher's aides, and all subscores were found to lack added value. Puhan, Sinharay, Haberman and Larkin (2010) applied the method to six tests used for educational certification and also found a lack of added value for all subscores in those six tests. Lyrén (2009) applied the method to a college admissions test and found, in contrast to the other studies, that most subscores had added value.

Feinberg and Wainer (2014b) proposed using a simple equation to predict the outcome of using Haberman's method in terms of whether a subscore has added value or not, and they provide examples of the accuracy of this equation. However, Sinharay, Haberman and Boughton (2015) claim that the equation is often inaccurate and therefore of limited value. This claim is challenged by Feinberg and Wainer (2015), who point out that with a more reasonable level of precision in the reported statistics than that used by Sinharay et al. (2015) the simple equation will still be a useful tool.

Utility Index

The utility index was suggested by Brennan (2012), using a method based on CTT that is algebraically equivalent to Haberman's method. The difference from Haberman's method is in the underlying assumptions and in the presentation of the results. The utility, U, is the same as Haberman's $PRMSE_x$, the relative utility, $\tilde{U}$, is the proportional change in subtest length needed for the subscore reliability to equal U ($\tilde{U} > 1$ implies an increase and $\tilde{U} < 1$ implies a decrease in subtest length), and k is the number of items to add to or delete from a subtest to bring that subscore's reliability equal to U. Both $\tilde{U}$ and k provide for a more detailed analysis of the subscores than just U. Extending the example from Haberman's method, assume that a subscore is based on 20 items and has a $PRMSE_s$ of 0.82, a $PRMSE_x$ of 0.84, a $\tilde{U}$ of 1.10 and a k statistic of 2. The interpretation of this would be that the subscore lacks added value and needs to be lengthened by 10%, which means that two comparable items (in terms of item statistics) need to be added to the subtest for the subscore to have added value.

Because of the algebraic equivalence to Haberman's method, the results will always be identical as far as determining whether the subscores have added value or not. Brennan (2012) applied the utility index to the SAT I data analyzed by Haberman (2008). Because of the algebraic equivalence of the methods, Brennan's results matched those of Haberman.

Subscore Augmentation

Wainer, Sheehan and Wang (2000) developed a method intended to stabilize subscores "by augmenting data from any particular subscale with information obtained from other portions of the test" (p. 119). It deserves to be repeated that subscore augmentation is not a method for examining the psychometric quality of subscores per se. Instead, it is used to examine if the augmented subscore is a better estimate of the true subscore compared to the observed subscore and thereby if it provides more information and is more useful for the test taker.

Wainer and colleagues (2001) applied augmentation in two tests; the first test was composed of six subtests and the second test was composed of four subtests. All six augmented subtest scores in the first test had substantially smaller mean squared errors than the corresponding observed subtest scores and were, therefore, considered to provide more information than the observed subtest scores. In the second test, using the same criterion, three of the four augmented subtest scores were found to provide more information than the corresponding observed subtest scores.

Haberman (2008) suggested a special case of augmented subscores, which he called weighted averages. For augmented subscores to provide more information than the observed subtest scores, the PRMSE of the augmented subscores ($PRMSE_a$) should be substantially larger than $PRMSE_s$ and $PRMSE_x$, defined as reducing the distance of both $PRMSE_s$ and $PRMSE_x$ from 1.0 by at least 10% (Haberman & Sinharay, 2013). For example, if a subscore has a $PRMSE_s$ of 0.82 and a $PRMSE_x$ of 0.84, then the $PRMSE_a$ has to be at least 0.016 larger than 0.84 to provide more information than the observed subscore. Therefore, if the $PRMSE_a$ is at least 0.86 the augmented subscore is more useful than the observed score.

Sinharay (2010) conducted studies on operational and simulated data and found that weighted averages had added value more often than observed subscores and that Wainer and colleagues' (2001) augmented subscores and Haberman's (2008) weighted averages performed very similarly as predictors of the true subscores.

Multidimensional Item Response Theory (MIRT)

Haberman and Sinharay (2010) proposed an approach that can be viewed as a MIRT version of Haberman's CTT-based method. The general idea behind this approach is to compare the PRMSE's of predictors based on different types of scores. For comparison with CTT-based scores, they used the previously described $PRMSE_s$ and $PRMSE_x$. They then proposed an MIRT-based PRMSE ($PRMSE_m$) and the unidimensional IRT (UIRT) equivalent of this ($PRMSE_u$). The difference between the UIRT- and the MIRT-based PRMSE's is that in the MIRT case the model is fitted to all subtests at once, but in the UIRT case the model is fitted to each subtest individually.

Again, extending the example from Haberman's method, assume that a subscore has a $PRMSE_s$ of 0.82, a $PRMSE_x$ of 0.84, a $PRMSE_u$ of 0.87 and a $PRMSE_m$ of 0.90. Here, the multidimensional IRT estimate contains the most information and therefore is more useful than the observed score.

Haberman and Sinharay (2010) applied a method of using MIRT, analogous to Haberman's method, to determine subscore quality in data obtained from five tests used for teacher certification. Their findings showed that the use of MIRT provided overall better estimates of the true subscores than those estimates obtained from observed subscores. This does not imply that the observed subscores lacked adequate psychometric quality, but rather that the MIRT-based subscores had added value over the observed subscores. In all cases, the subscores obtained using MIRT were considered more useful than those obtained using UIRT (i.e., the former had added value over the latter).

DIMTEST

Stout (1987) proposed a nonparametric IRT-based approach, DIMTEST, to investigate the assumption of unidimensionality in a test. DIMTEST conducts hypothesis tests of two sets of items. The subtest in focus is called the assessment subtest. The other subtest is called the partitioning subtest and is made up of all or some of the remaining items in the test. This method tests the null hypothesis that there is a dimensional similarity between the assessment subtest and the partitioning subtest, and rejection of the null hypothesis indicates a lack of unidimensionality.

For example, if a subscore has a DIMTEST *p*-value of 0.02 when tested against the total score, the null hypothesis is rejected and the subscore is considered to be dimensionally different from the total score and therefore to have adequate psychometric quality. Or, if a subscore has a DIMTEST *p*-value of 0.06 the null hypothesis is accepted and the subscore is considered to be dimensionally similar to the total score and therefore to lack adequate psychometric quality.

DIMTEST (and DETECT) were used by Ackerman and Shu (as cited in Sinharay et al., 2011) to examine the usefulness of subscores on a 5th-grade assessment. They found none of the subscores to be useful.

DETECT

The DETECT procedure (Zhang & Stout, 1999) is nonparametric in the same sense as DIMTEST. The procedure, either exploratory or confirmatory, searches for dimensionally homogenous clusters and produces an index value that indicates the amount of multidimensionality present in the test. Roussos and Ozbek (2006) found that values below 0.2 indicate approximate unidimensionality, values between 0.2 and 0.4 indicate weak to moderate multidimensionality, values between 0.4 and 1.0 indicate moderate to strong multidimensionality, and values above 1.0 indicate strong multidimensionality. For example, if a test has a DETECT index of 0.24 this indicates weak multidimensionality which supports the existence of subscores. If three clusters are found, then three subscores are empirically supported. If a cluster is composed solely of items from a single theorized subtest, for example a vocabulary subtest, (with reservation for random noise, see Zhang & Stout, 1999, pp. 241–242) the accompanying subscore is considered to have adequate psychometric quality. If items from a subtest are spread over multiple clusters, that subscore lacks adequate psychometric quality. For details about computations used in the DETECT procedure, see Stout (1990) and Zhang and Stout (1999).

Zhang and Stout (1999) applied DETECT to four analytical reasoning passages in an administration of the Graduate Record Examinations and found four clusters that corresponded perfectly to the four passages. They also applied DETECT to four reading comprehension passages in an administration of the Law School Admission Test and found three clusters. The DETECT index values obtained in the two studies were 0.799 and 0.709 indicating moderate to strong multidimensionality within the sections of both tests that were analyzed.

Factor Analysis

A factor analysis seeks to examine if the variables in a test or other instrument can be grouped together into a smaller number of variables, called factors. The number of factors indicated by a factor analysis represents the number of scores that can be considered to provide useful information to the test taker. Sinharay and colleagues (2011) discuss factor analysis in general terms, that is, both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). However, EFA seems to be used the most when examining subscores (e.g. Sinharay et al., 2007; Stone et al., 2010). In EFA, pattern coefficients of 0.40 or above indicate an association between a variable and a factor, whereas pattern coefficients below 0.40 indicate a lack of association (as used by Thompson, 2005). For example, in a factor structure with two factors, a subscore that has a pattern coefficient of 0.16 on Factor 1 and of 0.82 on Factor 2 is considered to be associated with Factor 2 but not with Factor 1. The same rule of association is also used on the item level.

Sinharay and colleagues (2007) conducted an exploratory factor analysis on a test with six subtest scores given to prospective and practicing teacher's aides. The test was made up of two subtests in the areas of reading, writing, and mathematics, and a lack of adequate psychometric quality was found in all subscores because the results of their analysis suggested only one dominant factor. Stone and colleagues (2010) used an exploratory factor analysis to determine the extent to which a one-factor model described a mathematics tests with four proposed subscores given to 8th-graders in the United States. They found one dominant factor and thus a lack of adequate psychometric quality in the subscores.

Method

Participants and Data

The two tests used for the empirical application in this study are two versions of a test used for selection to higher education. The first version (Test A) was designed to give only a total score that was a composite of five subtest scores. The second version (Test B) was designed to give a quantitative score and a verbal score (both of which are composites of four subtest scores) that are separately scaled and equated in a manner similar to that of the SAT. It is important to note that under both test versions and in addition to the total scores and section scores, the subtest scores have been reported to test takers. This is an action that has been supported by little empirical evidence and, therefore, should be examined more thoroughly.

Test A consisted of the subtests DS (data sufficiency; 22 items), DTM (diagrams, tables, and maps; 20 items), WORD (vocabulary; 40 items), READ (reading comprehension; 20 items), and ERC (English reading comprehension; 20 items), which gave a total of 122 items. Test B had a quantitative section that consisted of the subtests DS (12 items), DTM (24 items), XYZ (mathematical problem solving; 24 items), and QC (quantitative comparisons; 20 items). The verbal section consisted of WORD (20 items), READ (20 items), ERC (20 items), and SEC (sentence completion; 20 items). Each section thus had 80 items for a total of 160 items.

The participants in this study were test takers of Test A in the spring of 2010 (test form A1, n_{A1} = 57,050) and in the fall of 2010 (test form A2, n_{A2} = 40,662), and test takers of Test B in the fall of 2011 (test form B1, n_{B1} = 40,431) and in the spring of 2012 (test form B2[1], n_{B2} = 56,358). Of those taking test A1, 51% were female and the age range was 14–76 years with 80% being between the ages of 18 and 24. The other test forms had gender and age distributions

[1] The verbal section in B2 was scored using only 76 of the 80 items due to four WORD items being posted on an Internet forum before the items were to be administered. The items omitted from scoring were still used in this study because the item parameters appeared to be unaffected by the exposure of these items.

similar to A1. The data used were the test takers' scores on each item.

Intercorrelations

Intercorrelations between subtests as well as their correlation to the total score are displayed in Tables 1 and 2. All subtests had high correlations to the total score and READ had the highest.

In Test B, the subtest DTM differed from the other subtests in that its correlations to the subtests within the quantitative section were only slightly higher than its correlations to the subtests in the verbal section. A likely explanation for this is that DTM, to a higher degree than the other quantitative subtests, measured reading comprehension, which is a verbal skill. Because of the requirement to interpret numerical data DTM was, however, theorized to be a quantitative subtest, which was empirically supported by Table 2.

Table 1. Mean reliabilities (diagonal, in bold), intercorrelations (below the diagonal), and disattenuated intercorrelations (above the diagonal) between subtests in tests A1 and A2.

	WORD	DS	READ	DTM	ERC	Total test
WORD	**.86**	.51	.82	.56	.77	.95
DS	.42	**.79**	.66	.88	.65	.87
READ	.65	.50	**.72**	.70	.87	.99
DTM	.45	.67	.51	**.73**	.68	.91
ERC	.64	.52	.66	.52	**.79**	.96
Total test	.85	.75	.81	.75	.83	**.93**

Table 2. Mean reliabilities (diagonal, in bold), intercorrelations (below the diagonal), and disattenuated intercorrelations (above the diagonal) between subtests in tests B1 and B2.

	XYZ	QC	DS	DTM	WORD	READ	SEC	ERC	Q	V	Total test
XYZ	**.80**	.97	.83	.65	.29	.32	.44	.48	1.04	.45	.84
QC	.73	**.72**	.90	.70	.30	.52	.46	.52	1.07	.47	.87
DS	.60	.62	**.66**	.77	.43	.63	.58	.60	1.03	.59	.91
DTM	.48	.49	.52	**.69**	.50	.64	.61	.64	.96	.63	.90
WORD	.23	.23	.31	.37	**.80**	.81	.94	.75	.40	1.01	.80
READ	.36	.35	.41	.43	.58	**.64**	.91	.87	.61	1.07	.95
SEC	.34	.34	.41	.44	.73	.63	**.75**	.85	.55	1.08	.93
ERC	.37	.38	.42	.46	.57	.60	.63	**.74**	.60	1.01	.91
Q	.88	.86	.79	.76	.34	.46	.45	.49	**.90**	.57	.95
V	.38	.38	.45	.50	.86	.81	.89	.83	.51	**.91**	.95
Total	.72	.71	.71	.72	.69	.74	.77	.76	.87	.87	**.93**

In Test B, the highest intercorrelations for XYZ, QC, DS and DTM were with each other. The highest intercorrelations for WORD, READ, SEC and ERC were also with each other. This supported the respective composition of the two theorized sections of Test B, quantitative and verbal. The correlation between the sections was moderate, which indicated that the sections measured different aspects of the construct measured by the total score, which was desirable from a validity standpoint. The same pattern of subtest intercorrelations was found in Test A. Although Test A was not formally divided into sections, this supported the theorized "orientation" (quantitative/verbal) of each subtest.

The disattenuated correlations above 1 between subtests and sections in Test B were due to the subtests being part of the sections and therefore being correlated, in part, to themselves. This violates the assumption of error independence between correlated variables, which leads to inaccurate results when correcting for attenuation (Zimmerman, 2007).

Software and Estimation Notes

Four different software programs were used for the analyses. *SPSS Statistics 22* was used to estimate the

statistics necessary for the application of Haberman's method, utility index, and subscore augmentation; *FACTOR* (Lorenzo-Seva & Ferrando, 2011) was used for the factor analyses; *DIMPACK* (William Stout Institute for Measurement, 2006) was used for DIMTEST and DETECT; and *MIRT* (Haberman, 2013) was used to perform the MIRT-based analyses.

Exploratory factor analyses were carried out in this study in accordance with the method described by Sinharay and colleagues (2007) and Stone and colleagues (2010). *DIMPACK* had a restriction of 7,000 cases and 150 variables that led to randomly selected subsamples being used for the DIMTEST and DETECT analyses. We also used these samples for some of the analyses of Test A using the *MIRT* software to apply a five-factor MIRT to the data. The reason for this was that the analyses could not be completed for full datasets due to hardware limitations. Furthermore, we used the default settings in MIRT that included a two-parameter logistic model in both the MIRT and UIRT cases and a between-item model in the MIRT case.

When applying the DIMTEST and DETECT methods, 14 items from Test B were omitted from the analysis due to software limitations. The omitted items were selected at random within each subtest. In the quantitative section two items were omitted from XYZ, DTM, and QC, and one item was omitted from DS. This was determined based on DS having considerably less items than the rest of the subtests in the quantitative section making the omission of items proportionally similar among the four subtests. To ensure comparable results for the two section scores, seven items were removed from the verbal section as well. Because the subtests in the verbal section had an equal number of items, one item was omitted from one subtest selected at random (WORD) and two items were omitted from the remaining subtests.

When applying the DIMTEST method, the mean values from the analyses of three random samples of only 500 cases were used instead of the maximum of 7,000 supported by the DIMPACK software. This was due to problems with type I error rates with larger sample sizes and because a sample size of 500 was used as a lower bound in a previous study (Seo & Roussos, 2010). The DETECT analyses in this study were exploratory.

Data requirements

Haberman's method and therefore also utility index require subscores consisting of more than 10 items in order to detect added value (Sinharay, 2010). There is no previous literature on the recommended sample size of Haberman's method and utility index, so using a repeated sampling procedure on our own data we found that a sample size of 1200 was required in order to be at least 95% confident that the decision on subscore value in the sample was in agreement with the decision in the population.

Subscore augmentation has successfully been applied to subscores consisting of five items (Wainer et al., 2000). There is no previous literature on the recommended sample size of subscore augmentation, so using a repeated sampling procedure on our own data we found that a sample size of 1600 was required in order to be at least 95% confident that the decision in the sample, on whether augmented subscores provide more information that observed subscores or not, was in agreement with the decision in the population.

Reckase (1997) states that sample sizes of 1,000 and "fairly long tests" (p. 33) have been found to provide stable parameters when using MIRT. Data requirements for the DIMTEST statistic to work reasonably well, suggested in Stout (1987), are at least 300 examinees, at least 80 items in the test and at least five items in a subtest. For DETECT, Zhang and Stout (1999) used 400 examinees, 20 items in the total test and five items in a subtest as their minimum condition in two simulation studies, although no specific requirements were suggested.

MacCallum, Widaman, Zhang and Hong (1999) found that the required sample size when using factor analysis is between 100 and "well over 500" (p. 96), depending on the communalities and the strength with which factors are determined. A sample size of 100 produces good recovery of population factors when communalities are high, factors are well determined, and computations converge into a proper solution. A sample size of 500 or more is required when communalities are low and there are a large number of weakly determined factors.

Results

The results from the empirical application of the reviewed methods are presented below. The relevant

statistics for Haberman's method, utility index, subscore augmentation and MIRT are presented first because these results are provided in the same table (Table 3). The results for DIMTEST, DETECT and factor analysis come at the end of this section.

Haberman's Method and Utility Index

The PRMSE's based on Haberman's method for each subtest score and section score are shown in Table 3. It can be seen in both test forms of Test A that all subtest scores except READ had added value. In B1 the subtest scores for XYZ and DTM had added value and in B2 the subtest scores for DTM and WORD had added value. The quantitative and the verbal section scores had added value in both B1 and B2.

The results obtained with the utility index are also shown in Table 3. As stated previously, these are identical to those obtained with Haberman's method in terms of whether the subscores have added value or not. In A1 and A2 it would take 7 and 13 additional READ items, respectively, as indicated by k, for READ to have added value. In Test B, the length of several subtests would need to be at least doubled for their

Table 3. Estimated PRMSE, $\tilde{U}$ (in percentages), and k, for the subtest scores in all test forms and for the section scores in B1 and B2.

		Subtest scores								Section Scores	
Test form	Statistic	XYZ	QC	DS	DTM	WORD	READ	SEC	ERC	Q	V
A1	$PRMSE_s$			79	74	87	74		79		
	$PRMSE_x$			61	67	72	80		77		
	$PRMSE_a$			82	81	88[a]	85		85		
	$PRMSE_u$			80	76	88	76		79		
	$PRMSE_m$			86	85	89	88		87		
	$\tilde{U}$			43	62	44	135		86		
	k			-13	-8	-23	7		-3		
A2	$PRMSE_s$			79	72	86	69		79		
	$PRMSE_x$			61	64	74	79		76		
	$PRMSE_a$			82	79	88	83		85		
	$PRMSE_u$			79	73	87	70		81		
	$PRMSE_m$			85	83	89	85		88		
	$\tilde{U}$			43	69	48	165		83		
	k			-13	-7	-21	13		-4		
B1	$PRMSE_s$	83	76	65	68	80	63	76	75	91	91
	$PRMSE_x$	81	87	80	65	81	81	88	77	73	73
	$PRMSE_a$	87	88	83	76	86	82	89	83	91	91
	$PRMSE_u$	84	77	66	70	81	67	78	78	92	92
	$PRMSE_m$	89	89	83	78	88	85	90	85	92	92
	$\tilde{U}$	86	203	210	85	111	255	243	107	27	26
	k	-4	21	14	-4	3	31	29	2	-59	-59
B2	$PRMSE_s$	76	68	67	70	80	65	74	72	89	91
	$PRMSE_x$	81	82	77	61	77	81	87	76	72	75
	$PRMSE_a$	84	83	81	75	85	82	88	81	89	91
	$PRMSE_u$	78	72	68	72	77	70	74	68	90	90
	$PRMSE_m$	87	88	81	78	87	87	85	76	90	92
	$\tilde{U}$	131	217	165	70	84	221	233	121	34	31
	k	8	24	8	-8	-4	25	27	5	-54	-56

Note: Strictly speaking, k is always a positive integer but because k indicates a decrease in subtest length when $\tilde{U}$ <100 it is more descriptive to present k as a negative integer in these cases. [a]The $PRMSE_a$ is substantially larger than the $PRMSE_s$ but appears not to be, due to rounding.

corresponding subtest scores to have added value.

Subscore Augmentation

The $PRMSE_a$ values for each augmented subtest score and for each augmented section score are shown in Table 3. All of the augmented subtest scores in Test A had a substantially larger $PRMSE_a$ than both the $PRMSE_s$ and $PRMSE_x$ of the corresponding observed subscores. In Test B, all of the augmented subtest scores, except for QC, READ, and SEC, had a substantially larger $PRMSE_a$ than both the $PRMSE_s$ and $PRMSE_x$ of the corresponding observed subtest scores. The augmented section scores did not have a substantially larger $PRMSE_a$ than both $PRMSE_s$ and $PRMSE_x$ of the corresponding observed section scores. This implies that the augmented subtest scores in general can be viewed as containing more information than the observed subtest scores with the exception of QC, READ, and SEC in Test B. This also implies that the augmented section scores in general cannot be viewed as containing more information than the observed section scores.

Multidimensional Item Response Theory

$PRMSE_m$ and $PRMSE_u$ for each subscore are shown in Table 3. For all subtest scores in both Test A and Test B, $PRMSE_m$ was larger than $PRMSE_u$ and this means that the MIRT-based subtest scores had added value over the UIRT-based subtest scores. Also, for all subtest scores in Test A, $PRMSE_m$ was the largest of the five PRMSE's and this means that the MIRT-based subtest score in this test had added value over any other type of subtest score. For the subtest scores in Test B, $PRMSE_a$ was occasionally equal to or larger than $PRMSE_m$ indicating that the advantage of MIRT-based subtest scores over CTT-based subtest scores was not as evident when looking at this test. In regards to the section scores in Test B, there was no major difference between $PRMSE_m$ and $PRMSE_u$ and these were only somewhat, if at all, higher than $PRMSE_s$.

In only a few instances was $PRMSE_s$ larger than or equal to $PRMSE_u$. These were ERC in A1, DS in A2, and WORD, SEC, and ERC in B2. This implies that IRT-based subtest scores in general can be viewed as containing more information than the CTT-based subtest scores.

DIMTEST

The *p*-values for all subtest scores with the DIMTEST method for Test A were significant (α = .05), which means that each subtest score was dimensionally distinct from the total score. In Test B the subtests XYZ, DS, DTM, WORD, READ, and ERC were distinct from the section scores whereas QC and SEC were not. Both section scores were distinct from the total score.

DETECT

On a section level, the exploratory DETECT analyses for A1 and A2 showed moderate multidimensionality (the DETECT indices for the total score on Test A were $Total_{A1}$ = 0.40 and $Total_{A2}$ = 0.39). For test B, the analysis showed weak multidimensionality within the test sections (the DETECT indices for the quantitative and the verbal section of Test B were $Quantitative_{B1}$ = 0.28, $Quantitative_{B2}$ = 0.29, $Verbal_{B1}$ = 0.26, and $Verbal_{B2}$ = 0.21), and the quantitative section seemed to be somewhat more multidimensional than the verbal section. However, the found clusters (dimensions) in both Test A and Test B did not correspond to the theorized clusters, which are defined by each item's belonging to a specific subtest. Therefore, according to the DETECT analysis the subtest scores in both tests were considered to lack adequate psychometric quality.

On the total test level, the exploratory DETECT analyses showed moderate multidimensionality (the DETECT indices for the total score on Test B were $Total_{B1}$ = 0.59 and $Total_{B2}$ = 0.46). In all of the analyses, either three or four clusters maximized the DETECT statistic with one cluster containing mainly quantitative items, one cluster containing mainly verbal items, and one or two random noise clusters containing between one and four items. Therefore, according to the DETECT analysis the section scores of Test B were considered to have adequate psychometric quality.

Factor Analysis

On the section level, the exploratory factor analysis yielded one factor in both forms of Test A and two factors in both forms of Test B. These factors were determined using the Minimum Average Partial procedure (Velicer, 1976) and confirmed via parallel

analysis (Horn, 1965). In Test A, the one factor explained 64.7% (A1) and 63.4% (A2) of the variance, respectively. In Test B, the first factor explained 53.8% (B1) and 52.7% (B2) of the variance and the second factor explained an additional 18.2% and 17.3%, respectively. The rotated factor matrices showed that WORD, READ, SEC and ERC were associated with the first factor and XYZ, QC, DS, and DTM were associated with the second factor. The results were in accordance with the theoretical model of Test B where WORD, READ, SEC, and ERC constituted one section (verbal) and XYZ, QC, DS, and DTM constituted another section (quantitative). From a factor analysis perspective, therefore, both section scores in both test forms were considered to have adequate psychometric quality.

On the subtest level, the exploratory factor analysis yielded three factors in both forms of both tests. The factors in Test A corresponded to one quantitative factor (DS and DTM), and two verbal factors – one made up of WORD and part of READ, and one made up of ERC and the other part of READ. The factors in Test B corresponded to a quantitative factor (XYZ, QC and DS), a verbal factor (WORD, READ, and SEC), and a third factor consisting of DTM and ERC. The three factors together accounted for 16.4% (A1) and 15.7% (A2) of the explained variance in Test A and 22.0% (B1) and 20.1% (B2) in Test B. Because none of the factors coincided with the theorized subtests, all subtests in all test forms were considered to lack adequate psychometric quality.

Summary of Results

An overview of the results from all methods is shown in tables 4 (Test A) and 5 (Test B). The results are presented in terms of whether the subscores were

Table 4. The results from all methods showing whether the observed subtest scores in Test A had adequate psychometric quality or not, and whether they contained more or less information than the augmented subscores or MIRT estimates. The answers are presented as Yes (Y), No (N), More (M), or Less (L).

	Subtest scores for Test A				
Method	**DS**	**DTM**	**WORD**	**READ**	**ERC**
Haberman's Method and Utility Index	Y	Y	Y	N	Y
Augmentation	L	L	L	L	L
MIRT	L	L	L	L	L
DIMTEST	Y	Y	Y	Y	Y
DETECT	N	N	N	N	N
Factor Analysis	N	N	N	N	N

Note: For all methods, the same results were found for both test forms of both tests so only one answer is given.

Table 5. The results from all methods showing whether the observed subtest scores and section scores in Test B had adequate psychometric quality or not, and whether they contained more or less information than the augmented subscores or MIRT estimates. The answers are presented as Yes (Y), No (N), More (M), or Less (L).

	Subtest scores for Test B								Section Scores	
Method	XYZ	QC	DS	DTM	WORD	READ	SEC	ERC	Q	V
Haberman's Method and Utility Index	Y/N	N	N	Y	N/Y	N	N	N	Y	Y
Augmentation	L	M	L	L	L	M	M	L	M	M
MIRT	L	L	L	L	L	L	L	L	L	L
DIMTEST	Y	N	Y	Y	Y	Y	N	Y	Y	Y
DETECT	N	N	N	N	N	N	N	N	Y	Y
Factor Analysis	N	N	N	N	N	N	N	N	Y	Y

Note: Whenever the same result is obtained using the same method on both test forms only one answer is given. When different results were obtained, the first answer refers to test form B1 and the second to B2.

found to have adequate psychometric quality or not. It should be noted again that augmented subscores and MIRT-based subscores do not test whether or not observed subscores have adequate psychometric quality, but rather if this can be improved. Therefore, the results using those methods are classified as either 'More' (the observed subscores contain more information than the augmented subscores or MIRT estimates) or 'Less' (the observed subscores contain less information).

Discussion

In this paper we have provided a review and application of methods for examining the psychometric quality of subscores. From the empirical application we note that all of the methods used in this study suggest that the two section scores in Test B have adequate psychometric quality. Consequently, it could be useful to report these to the test takers. In contrast, on the subtest level the results varied greatly for both tests. Exploratory factor analysis and DETECT indicated a lack of adequate psychometric quality for the subtest scores, while DIMTEST indicated adequate psychometric quality for almost all subtest scores. A possible explanation for this is what was mentioned earlier, that although the methods all involve assessment of dimensionality, they still have some unique characteristics. Exploratory factor analysis and DETECT are methods developed specifically for assessing "optimal" dimensionality, while DIMTEST involves significance testing and Haberman's method, utility index, subscore augmentation, and MIRT explicitly consider score reliability information.

Several methods showed a lack of quality for most subtest scores, but DIMTEST indicated quality for all subtest scores in A1 and A2 as well as for all subtest scores except QC and SEC in B1 and B2. All methods indicated added value for the section scores. This is an important result for Test B because it provides empirical evidence for the validity of the decision to scale, equate and report the two section scores separately. The MIRT methodology indicated that IRT-based scores are preferable to CTT-based scores.

When applying Haberman's method, the subscore value changed for some subtest scores between test forms in Test B. In B1, subtests DTM and XYZ provided added value but in B2 subtests DTM and WORD provided added value. This change was not surprising because subscore reliabilities vary between test forms due to sampling variability. This means that subscores with small differences between $PRMSE_s$ and $PRMSE_x$ will sometimes be shown to provide added value and sometimes not. Therefore, no definitive conclusions can be drawn in regards to these subtest scores using Haberman's method. Still, as pointed out by Haberman (2008), as long as the subtest scores are rather reliable they "can be employed by themselves to provide relatively accurate approximations to the true subscores" (p. 224). This applies to IRT-based PRMSE's as well.

Exploratory factor analysis and DETECT are both used to determine the number of dimensions present in a test and this is how they were used in this study. They are also the only two methods that suggest a lack of adequate psychometric quality for all subtest scores. This might be due to all subtest scores truly lacking adequate psychometric quality, but, because the other methods found adequate psychometric quality, it might also be a sign that these methods are inappropriate to use to examine the psychometric quality of subscores on an item level in large scale assessments. This is because, similar to the alternative scores when using subscore augmentation and MIRT, they only provide information on the "optimal" dimensionality instead of assessing the appropriateness of the theorized dimensionality, thus failing to evaluate the adequacy of the subscores. The findings in this study concerning the empirical applications of the methods are of use to test users who want to examine subscore quality before deciding which scores to report to the test takers. These findings are likely to be generalizable to other large scale tests with similar structure to Test B, such as the SAT in the U.S. or the PET in Israel. A possible limitation, and a basis for further research, is that the analyzed tests were not developed from a strict factor analytic perspective where highly separated factors are the primary goal. While internal structure is an important consideration for these tests, as well as for many other educational tests, content considerations are even more important. It is possible, or even likely, that the differences between the methods would have been different if we had analyzed tests that were developed in such a way that the different subtest scores were clearly distinct, which is the case with many psychological tests.

The findings concerning the psychometric value of the subtest scores in Test B primarily affect those who

retake Test B and who base their remediation strategy on their subtest scores from previous administrations of that test. The percentage of repeaters has increased over the past few years and repeaters constituted 46% of the test takers in A1 and A2, 47% in B1, and 49% in B2. This means that about 20,000 test takers or more per test form or at least 40,000 test takers per year are potentially affected if their remediation strategies are based on subtest scores from previous administrations of the test. Those affected by the findings concerning the psychometric value of the section scores in Test B are the policy makers and researchers who took part in the decision to scale, equate and report the two sections scores separately, and those who continually support this.

The results concerning the empirical applications of the methods can be generalized to future administrations of Test B but might not be generalized to other tests. This limitation should also be evident from the difference in results between the old and new version of the test. It is, however, important to remember that the results for ERC, when using PRMSE methods, might very well change at some point to indicate that ERC lacks adequate psychometric quality as was explained previously.

As described earlier there were some hardware and software issues in applying some of the methods. *DIMPACK* has restrictions on sample size and test length while *MIRT* requires rather powerful computers when there are four or more factors. These software restrictions will probably make the related methods less attractive to people in the testing industry. Exploratory factor analysis will probably also be less attractive, as explained above, unless the examined test is specifically developed using factor analysis.

If we were to recommend a method that will fit the need of most potential users involved in testing it would be Haberman's method complemented by the test length statistics ($\tilde{U}$ and k) from the utility index. Haberman's method is based on simple statistical measures and takes into account the two key properties of subscores when it comes to score reporting: reliability and distinctiveness. We see the utility index as an extension of Haberman's method, introducing the concept of the relative utility and providing an estimate of the statistic k, the number of items to add to a subtest to make it reliable enough to report. Both these statistics should be highly relevant for all practitioners involved in the reporting of test scores.

Future research in this area might be to investigate how many of the repeated test takers actually pay attention to their subscores when deciding on remedial strategies. It might also be of interest to find out how many repeated test takers are aware of the implications of several subtest scores potentially lacking adequate psychometric quality and why these should not be used for remedial studies. A study of the effects of remedial studying using subscores that lack adequate psychometric quality should also be considered. A closely related area is external analysis of subscore quality, which focuses on the subscores' predictability for a criterion (e.g., Davidson, Davenport, Chang, Vue & Su, 2015). External analysis of subscore value has great potential to complement internal analysis and is a topic that deserves more attention.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Brennan, R. L. (2012). Utility indexes for decisions about subscores (CASMA Research Report No. 33). Iowa City, IA: University of Iowa, Center for Advanced Studies in Measurement and Assessment. Retrieved from: http://www.education.uiowa.edu/docs/default-source/casma---research/33utility-revised.pdf?sfvrsn=2

Davison, M. L., Davenport, E. C., Jr., Chang, Y.-F., Vue, K., Su, S. (2015). Criterion-related validity: Assessing the value of subscores. *Journal of Educational Measurement*, 52(3), 263–279.

Feinberg, R. A., & Wainer, H. (2014a). When can we improve subscores by making them shorter?: The case against subscores with overlapping items. *Educational Measurement: Issues and Practice,* 33(3), 47–54.

Feinberg, R. A., & Wainer, H. (2014b). A simple equation to predict a subscore's value. *Educational Measurement: Issues and Practice*, 33(3), 55–56.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, 33(2), 204–229. doi: 10.3102/1076998607302636

Haberman, S. J. (2013). A general program for item-response analysis that employs the stabilized Newton-Raphson algorithm (Report No. RR-13-32). Retrieved from Educational Testing Service website: http://www.ets.org/Media/Research/pdf/RR-13-32.pdf

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, 75(2), 209–227. doi: 10.1007/S11336-010-9158-4

Haberman, S. J., & Sinharay, S. (2013). Does subgroup membership information lead to better estimation of true subscores? *British Journal of Mathematical and Statistical Psychology,* 66, 452–469. doi: 10.1111/j.2044-8317.2012.02061.x

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, 62, 79–95. doi: 10.1348/000711007X248875

Hanson, B. A. (1989). Scaling the P-ACT+. In R. L. Brennan (Ed.), *Methodology used in scaling the ACT Assessment and P-ACT+* (pp. 57–73). Iowa City, IA: American College Testing.

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

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185.

Ling, G. (2012). Why the major field test in business does not report subscores: Reliability and construct validity evidence (Report No. RR-12-11). Princeton, NJ: Retrieved from Educational Testing Service website: https://www.ets.org/Media/Research/pdf/RR-12-11.pdf

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika*, 30(3), 239–270.

Lorenzo-Seva, U., & Ferrando, P. J. (2011). FACTOR (Version 8.02) [Computer software]. Tarragona, Spain: Rovira i Virgili University.

Lyrén, P.-E. (2009). Reporting subscores from college admission tests. *Practical Assessment, Research, & Evaluation*, 14(4). Available online: http://pareonline.net/pdf/v14n4.pdf

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4(1), 84–99. doi: 1082-989X/99/S3.00

Monaghan, W. (2006). The facts about subscores (Report No. RDC-04). Princeton, NJ: Educational Testing Service.

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: an evaluation of methods using empirical data. *Applied Measurement in Education*, 23(3), 266–285. doi: 10.1080/08957347.2010.486287

Reckase, M. D. (1997) The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21(1), 25–36.

Roussos, L. A., & Ozbek, O. (2006). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. *Journal of Educational Measurement*, 43(3), 215–243.

Seo, M., & Roussos, L. A. (2010). Formulation of a DIMTEST effect size measure (DESM) and evaluation of the DESM estimator bias. *Journal of Educational Measurement*, 47(4), 413–431.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, 47(2), 150–174.

Sinharay, S., Haberman, S. J., & Boughton, K. (2015). Too simple to be useful: A Comment on Feinberg and Wainer (2014). *Educational Measurement: Issues and Practice*, 34(3), 6–8.

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.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, 30(3), 29–40.

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education*, 23(1), 63–86.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52(4). 589–617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, 55(2), 293–325.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3), 321–327.

Wainer, H., Sheehan, K. M., & Wang, X. (2000). Some paths toward making Praxis scores more useful. *Journal of Educational Measurement*, 37(2), 113–140.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., III, Rosa, K., Nelson, L., Swygert K. A., & Thissen, D. (2001). Augmented scores—"borrowing strength" to compute scores based on small number of items. In D.

Thissen & H. Wainer (Eds.), *Test Scoring* (pp. 343–388). Mahwah, NJ: Lawrence Erlbaum Associates.

William Stout Institute for Measurement. (2006). Nonparametric dimensionality assessment package DIMPACK (Version 1.0) [Computer software]. St. Paul, MN: Assessment Systems Corporation.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, 64(2), 213–249.

Zimmerman, D. W. (2007). Correction for attenuation with biased reliability estimates and correlated errors in populations and samples. *Educational and Psychological Measurement*, 67(6), 920–939.

Citation:

Wedman, Jonathan, and Lyrén, Per-Erik (2015). Methods for examining the psychometric quality of subscores: A review and application. *Practical Assessment, Research & Evaluation*, 20(21). Available online: http://pareonline.net/getvn.asp?v=20&n=21

Note:

The authors are grateful to Marie Wiberg for her valuable comments on the manuscript and to Shelby Haberman for his generous technical support regarding the MIRT software.

Corresponding Author:

Jonathan Wedman
Department of Applied Educational Science
Umeå University
SE-90187 Umeå
Sweden
jonathan.wedman@umu.se