Measuring the Power of Learning.™

Research Report
ETS RR–18-17

A Review of Subscore Estimation Methods

Jianbin Fu

Yanxuan Qu

December 2018

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

A Review of Subscore Estimation Methods

Jianbin Fu & Yanxuan Qu

Educational Testing Service, Princeton, NJ

Various subscore estimation methods that use auxiliary information to improve subscore accuracy and stability have been developed. This report provides a review of various subscore estimation methods described in the literature. The methodology of each method is described, then research studies on these subscore estimation methods are summarized. Comments on the methods and suggestions for future areas of research are provided, and preliminary guidelines for using subscore estimation methods in practice are recommended.

Keywords Subscore estimation; review; regression; item response theory; classical test theory

doi:10.1002/ets2.12203

Subscores provide test users with diagnostic information on fine-grained content domains or skills and are useful, for example, in making decisions about admissions or educational remedies (e.g., help for students who have lower achievement). In practice, many testing programs report subscores to their clients. Because subscores are often based on a small number of items, the reliabilities of these subscores may not be sufficiently high for reporting, even though the reliability of the total test score may be adequate to support the intended decisions. Various methods have been proposed that use information from other items within a test to improve subscore accuracy and stability (Haberman, 2008; Haberman & Sinharay, 2010; Kahraman & Kamata, 2004; Wainer et al., 2001; Yao & Boughton, 2007; Yen, 1987; Yen, Sykes, Ito, & Julian, 1997). These methods differ in how the auxiliary information is used to estimate subscores. This report provides an integrated review of various subscore estimation methods described in the literature. First, the methodology of each method is described. Next, research studies on these subscore estimation methods are summarized. Finally, the methods are discussed, future areas of research are proposed, and preliminary guidelines on the use of subscore estimation methods in practice are recommended.

Description of Subscore Estimation Methods

Table 1 lists the summary information of eight subscore estimation methods seen in the literature, including the underlying test theory/statistical method, whether and how collateral information is used to estimate subscores, and pros and cons, if any, for each method. These key points for each method are discussed subsequently in detail.

These methods are classified into two categories: methods based on *classical test theory* (CTT) and methods based on *item response theory* (IRT). In CTT, subscores are reported as *raw subscores* or linear transformations of raw subscores (e.g., *percentage correct* and *standardized subscores*), whereas in IRT, subscores may be reported as *theta estimates*, linear transformations of theta estimates (e.g., *scale scores*), *IRT true score estimates*, or linear transformations of IRT true score estimates (e.g., *percentage correct IRT true score estimates*). Note that IRT true score is a function of theta and item parameters based on an IRT model. Theta estimates can be further classified into seven types based on estimation methods and observed data types (Thissen, Pommerich, Billeaud, & Williams, 1995; Thissen & Wainer, 2001; Yen, 1984; see Table 2). Raw or transformed subscores and IRT-based subscores based on a unidimensional IRT model applied to the items in a subtest are two straightforward estimation methods that do not use collateral information. The procedures of the other subscore estimation methods that use collateral information are described in the following pages. The emphasis is on the main idea of each method; for computational details, please refer to the references therein. After the introduction of subscore estimation methods, a sample of subscore estimates by the various methods is provided.

Corresponding author: J. Fu, E-mail: jfu@ets.org

Table 1 Summary of Subscore Estimation Methods

Subscore estimation methods	Type	Collateral information	Advantages	Disadvantages
Raw subscores or linear transformations of raw subscores (e.g., percentage correct and standardized subscores)	CTT	No	• Very easy computation	• Subscore estimates least accurate and reliable for a short subtest
Subscore augmentation (Wainer et al., 2001) using observed raw scores	CTT, regression approach	Use the observed raw scores of all subtests to predict the true scores of each subtest.	• Provide accurate and reliable subscore estimates if subscores are highly correlated • Relatively easy computation	• Hard to explain to test users why a subscore estimate depends not only on the observed subscore but also on other observed subscore(s)
Haberman's (2008) methods	CTT, regression approach	Use the observed raw score of the predicted subtest, total raw scores, or both to predict the true scores of each subtest.	• Provide a quick tool to judge whether subscores should be reported in addition to total scores • Relatively easy computation	• Hard to explain to test users why a subscore estimate depends not only on the observed subscore but also on other observed subscore(s)
Subscores (i.e., thetas, IRT true scores, or their linear transformations) based on a unidimensional IRT model on items within each subtest separately	IRT	No	• Relatively easy computation	• Subscore estimates not accurate and reliable for a short subtest
Subscores based on a unidimensional IRT model on each subtest, but with item parameter estimates based on all items in a test	IRT	The items in other subscores may impact the item parameter estimates in a target subscore.	• Relatively easy computation	• Subscore estimates not accurate and reliable for a short subtest
Kahraman and Kamata's (2004) method (Ackerman & Davey, 1991; Davey & Hirsh, 1991)	IRT	The items in other subscores may impact the theta estimates in a target subscore.		• Lack theoretical ground; using thetas estimated from all items to represent a subscale is questionable
Objective performance index (OPI; Yen, 1987; Yen et al., 1997)	IRT	The items in other subscores are used as prior information for a target subscore.		• Does not perform well when the correlations between subscores are low or intermediate • Make unrealistic assumptions: (a) all items in a subtest have the same response probability for the examinee and (b) a polytomous item with M_i score categories is treated as $M_i - 1$ dichotomous items
Subscores based on a multidimensional IRT model (e.g., de la Torre & Patz, 2005; Fu, 2009; Haberman, 2013; Haberman & Sinharay, 2010; von Davier, 2008; Yao & Boughton, 2007)	IRT	The items in other subscores may impact the item parameter and theta estimates in a target subscore.	• Subscore estimates are most accurate and reliable	• Heavy computational burden
Subscore augmentation (Wainer et al., 2001) using theta estimates from an IRT model	IRT, regression approach	Uses the theta estimates of all subtests to predict the augmented theta scores of each subtest.	• Subscore estimates are most accurate and reliable • Relatively easy computation	• Hard to explain to test users why a subscore estimate depends not only on the observed subscore but also on other observed subscore(s)

Note. CTT = classical test theory; IRT = item response theory.

Table 2 Types of Theta Estimates

Data type	Estimation methods
Item pattern	Maximum likelihood Maximum a posteriori Expected a posteriori
Summed raw score	Maximum likelihood Maximum a posteriori Expected a posteriori Raw to IRT scale score conversion based on the test characteristic curve

Note. IRT = item response theory.

Regression Approach

The *regression approach* uses a linear regression to predict classical true scores of subtests from observed raw scores. Subscore estimation methods under this approach differ in the observed score predictors used. *Subscore augmentation* (Wainer et al., 2001) uses the observed raw scores of all subtests to predict the true scores of each subtest (referred to as *augmented scores*), while in Haberman's (2008) methods, the three common models use the observed raw score of the predicted subtest, total raw scores, and both, respectively. The estimation of regression coefficients is based on the typical least squares method for regression and classical true score theory.

To measure the stability of augmented score estimates, the R^2 statistic of augmented scores (i.e., the squared correlation between true and estimated augmented scores) is used (Sinharay & Haberman, 2008, p. 25). Haberman (2008) interpreted the R^2 statistic as the proportional reduction of the mean squared errors (PRMSE) obtained by a regression model compared to the approximation of the true scores using an unbiased estimate (i.e., the mean of the raw scores). The PRMSE is computed as the mean squared errors from the approximation minus those from the regression model, divided by the mean squared errors from the approximation. Note that when the observed subscore in the target subtest is used as the only predictor, the R^2 of augmented scores is the same as the classical reliability of this subtest under CTT (e.g., Cronbach's alpha).

Haberman (2008) used PRMSE to judge if subscores had added value given that total scores had been reported. In particular, if the PRMSE for the model with the total score as the predictor is larger than for the model with the target subscore as the predictor, then it indicates that subscores have no added value. If the PRMSE is smaller, then reporting subscores is justified (Sinharay & Haberman, 2008). The third model in Haberman (2008) uses both target subscore and total scores as predictors, which has PRMSE at least as large as the other two models, because the other two models are nested within the third model. Sinharay (2010), Sinharay and Haberman (2008), and Sinharay, Puhan, and Haberman (2010) suggested using the third model only if its PRMSE reduction was substantial (a cut point of 0.01 was used in their papers). Note that the third model is a submodel of subscore augmentation (Wainer et al., 2001) with the regression coefficients of the subscores other than the target subscore constrained to be equal. However, the third model and subscore augmentation were found to produce similar predictions (Sinharay, 2010; Sinharay & Haberman, 2008).

Wainer et al. (2001, pp. 357–358) used the R^2 of augmented scores to infer test dimensionality so as to check whether subscores were worth reporting. In particular, R^2 of augmented scores for all subtests that are similar to the reliability of the whole test indicates that the test is essentially unidimensional. Wainer et al. (2001, pp. 352–353) also pointed out that the variance–covariance matrix of observed subscores could be used for the same purpose. If the matrix has a dominant eigenvalue, then the test is virtually unidimensional and subscores have little added value.

Haberman, Sinharay, and Puhan (2009) extended the three models in Haberman (2008) for subscore estimation at an aggregate level, for example, at the level of the institutions to which the examinees belong.

Subscore Augmentation With Theta Estimates From an Item Response Theory Model

Wainer et al. (2001; also see de la Torre & Patz 2005) also used IRT theta estimates of subtests as predictors in the subscore augmentation model to get the augmented theta scores of a subtest. IRT theta estimates on each subtest are based on a unidimensional IRT model. Because *maximum* a posteriori (MAP) or *expected* a posteriori (EAP) theta estimates shrink

toward the population mean, it should be first corrected to remove the shrinkage. Assuming that the population mean of thetas on each subtest is zero and the standard errors of theta estimates are constant, the correction is made on the theta estimates by dividing them by their reliabilities. These corrected theta estimates of subtests are then used as predictors in the regression model in the same way as raw subscores, as described earlier. Note that for *maximum likelihood* (ML) theta estimates of subtests, the correction is unnecessary, and they can be used directly as predictors in the regression model.

Unidimensional Item Response Theory Model on Subtest With Item Parameter Estimates Based on All Items in a Test

This subscore estimation method first applies a unidimensional IRT model using all the items in a test to get the item parameter estimate(s) for each item. And then IRT subscale scores for a subtest are estimated based on the items in the subtest, using their item parameters from the calibration based on all the items in the test. The commonly used IRT models are the one-, two-, or three-parameter logistic models (1 – 3PL) for dichotomous items and the partial credit model (PCM) or the generalized partial credit model (GPCM; also known as the two-parameter partial credit model) for polytomous items.

Kahraman and Kamata's Method

Kahraman and Kamata's (2004) method used the information from other subtests by estimating a target subscore based on all the items in a test. However, the item parameters in other subtests were aligned to the scale of the target subscore prior to using them to estimate the target subscore. In particular, this method was carried out for each subscale as follows:

1. Calibrate the item parameters of the items in the target subscale (called in-scale items) based on a unidimensional IRT model.
2. Fix the in-scale items' parameters from Step 1, and calibrate each of the remaining items in the test (called out-of-scale items), one by one, by the same IRT model.
3. Estimate the IRT subscale scores based on all items in the test with item parameters from Steps 1 and 2 by the same IRT model.

These steps were repeated for each subscale.

Objective Performance Index

The OPI (Yen, 1987; Yen et al., 1997) applies Bayesian and IRT theories to estimate subscores. In the OPI, the target subscore is the expected percentage correct subscore. The OPI assumes the distribution of an examinee's observed raw score conditional on the examinee's expected percentage correct subscore following a binomial distribution, and the prior distribution of the expected percentage correct subscore is a beta distribution. Because a beta distribution is a conjugate prior for a binomial distribution, the posterior distribution of the expected percentage correct subscore conditional on the observed raw score is also a beta distribution. The OPI subscore is just the mean of the posterior distribution.

The prior information comes from a calibration of all items in a test by a unidimensional IRT model. Specifically, a unidimensional IRT model such as one of those mentioned previously is applied to all test items to get item parameter estimates for each item and a theta score estimate for each examinee. Then the mean and variance of the prior distribution of an examinee's expected percentage correct subscore are assumed to be the examinee's percentage correct IRT true subscore and its variance, respectively. With the known mean and variance of the prior distribution of an examinee's expected percentage correct subscore, the mean of the posterior distribution of the examinee's expected percentage correct subscore (i.e., OPI subscore) can be computed (for details, see Yen, 1987; Yen et al., 1997).

Because theta estimates are based on all the items in the test, the prior information overlaps with the posterior information. An adjustment for reducing the overlapping information could be made on the variance of the prior distribution of an examinee's expected percentage correct subscore by multiplying the variance of the examinee's percentage correct IRT true subscore by $(U - U_f)/U$, where U and U_f are the maximum score possible on the whole test and the target subtest, respectively. This is called the adjusted OPI.

In the OPI, a chi-square test is used to check whether an examinee's percentage correct IRT true subscore is a good estimate of the examinee's observed percentage correct subscore. If not, the prior information is not used to calculate the OPI, and the OPI is equal to the observed percentage correct subscore.

Note that the OPI assumes that the distribution of an examinee's observed subscore, conditional on the examinee's expected percentage correct subscore, follows a binomial distribution. This is actually a simplified assumption that (a) all items in a subtest have the same response probability for the examinee and (b) a polytomous item with M_i score categories is treated as $M_i - 1$ dichotomous items. This simplified assumption can cause estimation errors, especially if a test is composed of a large number of polytomous items (Yen et al., 1997).

Multidimensional Item Response Theory Model

Many multidimensional IRT (MIRT) models, such as the multidimensional extensions of 1–3PL, PCM, and GPCM, have been developed in the past 30 years (e.g., Adams, Wilson, & Wang, 1997; Béguin & Glas, 2001; Fu, 2009; Haberman, von Davier, & Lee, 2008; McDonald, 1997; Muraki & Carlson, 1995; Reckase, 1997; von Davier, 2008). Using MIRT, IRT subscales are represented by theta estimates on different dimensions (commonly referred to as attribute or skill). For example, if a test has two subtests, a two-dimensional IRT model can be used to fit the test data. In this case, each dimension represents a subtest, and each item has loading(s) on the subtest(s) to which it belongs. The collateral information used in MIRT for a subscore estimation is the item responses in the other subscores; the key information is the correlation structure of subscores. The reliability of IRT subscores (e.g., Adams, 2006; Haberman & Sinharay, 2010) is defined as the squared correlation between true and estimated IRT subscores and is comparable to the R^2 of augmented scores from the regression models. If the reliability of IRT subscores is larger than the R^2 of augmented scores, then the subscale from the MIRT approach is more stable.

Note that other MIRT models can be used for subscore estimation, such as the conjunctive Rasch model (Maris, 1995) and the multicomponent latent trait model (e.g., Embretson, 1997). Some latent class models were proposed to report performance levels on fine-grained skills, for example, the reparameterized unified model (Roussos et al., 2007), the deterministic-input noisy and gate model; the noisy-inputs deterministic "and" gate model (Junker & Sijtsma, 2001), and the Bayesian inference networks (e.g., Almond, DiBello, Moulder, & Zapata-Rivera, 2007). In addition, some other statistical models are proposed in the literature for diagnostic score estimation, for example, the rule-space method (Tatsuoka, 1983), the attribute hierarchy method (Leighton, Gierl, & Hunka, 2004), and the tree-based approach (Sheehan, 1997). These models/methods are rarely used in current large-scale assessments and thus are not covered in this review. Interested readers may refer to the review papers or monographs for a complete accounting of multidimensional psychometric models that may be used for diagnostic score estimation, for example, Dibello, Roussos, and Stout (2007), Fu and Li (2007), Reckase (2009), Rupp and Templin (2008), Rupp, Templin, and Henson (2010), and von Davier, Dibello, and Yamamoto (2008).

Sample of Subscore Estimates

Table 3 shows various subscore estimates for three hypothetical students on two subtests by the methods listed in Table 1, except for Haberman's methods and Kahraman and Kamata's method. The three students were selected from a simulated data set with 3,000 examinees and two subtests to have low, medium, and high subscores, respectively. The two subtests had 10 and 20 items, respectively, and in each subtest, 80% of the items were dichotomous items and 20% of the items were three-point category items. The data were generated based on the two-dimensional GPCM with the correlation between subtest thetas of .5. Subscore estimates are reported as percentage correct subscores in Table 3 to facilitate comparisons across the subscore estimation methods. As one can see, the subscore estimates by the seven methods are not the same. The extent of the variation depends on the methods and tests (e.g., subtest length and correlation between subtests). For example, the subscore estimates across the different methods for 20 items are more similar than for 10 items. Because of this variation, it is important to compare these methods under various conditions, study the implications of the differences in the subscore estimates in practical situations, and possibly recommend the best methods. In the subsequent sections, the studies on subscores in the literature are summarized and discussed.

Table 3 Sample Estimates of Percentage Correct Subscores by Seven Methods

Subscore estimation method	Subscore 1[a] (%)			Subscore 2[b] (%)		
	Student 1	Student 2	Student 3	Student 1	Student 2	Student 3
Raw score	17	50	100	8	50	88
Subscore augmentation with raw subscores	24	50	84	15	50	83
Subscore augmentation with theta subscores from GPCM	27	54	81	17	48	83
GPCM on subtest	31	54	78	17	47	81
GPCM with item parameters based on all the items in a test	38	51	69	17	47	81
OPI	28	48	77	15	48	85
MGPCM	26	54	82	16	48	84

Note. GPCM = generalized partial credit model; MGPCM = multidimensional GPCM; OPI = objective performance index.
[a]Ten items. [b]Twenty items.

Research on Subscore Estimation Methods

Research on subscore estimation methods in the literature focuses on the following research questions.

Are Subscores Worth Reporting Given That Total Scores Have Been Reported?

This research question asks whether reporting a subscore adds any value to the total score. Haberman (2008) has shown that subscores are increasingly favored as the reliability of the subtest increases, the reliability of the total test decreases, and the disattenuated correlation between subtest scores and total test scores decreases. Sinharay (2010), Sinharay and Haberman (2008), and Sinharay et al. (2010) applied Haberman's three models to a bunch of real data sets and found that only a few subscores added value over and above reporting total scores, which, in general, had more items, high reliabilities, and low disattenuated correlations. Sinharay (2010) also examined the conditions necessary to increase the value of subscore reporting using a simulation study; the conclusion was that the subscores had to be long (at least 20 items) and sufficiently distinct from each other (disattenuated correlations less than .85).

The value of reporting a subscore is also closely related to test dimensionality. If a test is essentially unidimensional, then reporting subscores will not provide more information than what has already been included in total scores. For reviews of the methods to assess test dimensionality, see, for example, De Champlain and Gessaroli (1998), Hattie (1984, 1985), and Levy and Svetina (2010). Under the score augmentation method, Wainer, Sheehan, and Wang (2000) and Wainer et al. (2001) used the R^2 of augmented scores and the observed variance–covariance matrix of subscores to assess the dimensionality of a test. They found that subscores for the Education in the Elementary School Assessment (one of the *PRAXIS*® assessments for teacher licensure; Wainer et al., 2000) and the 1994 American Production and Inventory Control Society Certification Examination provided no added value because these two tests were virtually unidimensional (Wainer et al., 2001). However, Wainer et al. found that the four subscores in the North Carolina Test of Computer Skill had some added value.

A related question concerns the impact on theta estimation if a multidimensional test is estimated by a unidimensional IRT model. Many studies (e.g., Ackerman, 1994; Reckase, Ackerman, & Carlson, 1988; M. Wang, 1986) have shown that a unidimensional theta estimate can be thought of as a linear combination of multidimensional theta estimates. The quality of the unidimensional theta estimates increases as the subtests are more homogenous. Tate (2004) investigated the impact on the precision of unidimensional theta estimates in terms of standard errors of theta when the test data were actually multidimensional. Tate concluded that the precision of theta estimates based on a unidimensional IRT model with an underlying multidimensional structure increased as the number of subtests decreased and the correlation between thetas increased, that is, the test became more homogenous. However, the composite unidimensional theta estimates may lose information and lead to bias and differential item functioning (DIF), for example, if item difficulty and dimensionality are confounded in the data (Reckase, Carlson, Ackerman, & Spray, 1986) or if group differences are present in the distribution of the multidimensional thetas (Ackerman & Evans, 1994; Walker & Beretvas, 2001). Walker and Beretvas (2003) demonstrated that the use of a unidimensional IRT model on a two-dimensional test led to proficiency classification bias for some students.

Does the Subscore Estimation Method Improve the Quality of Subscore Estimates?

This question is a related but separate question from the first one. Subscores may not have added value, but their accuracy and stability can still be improved by different methods. Note that in this report, stability is used interchangeably with reliability or R^2 (i.e., the squared correlation between estimated subscores and true subscores),[1] and accuracy is used interchangeably with root mean squared error (RMSE), where error refers to the difference between estimated and true subscores. The subscore estimation methods using ancillary information (i.e., those methods described in the previous section) are presumed to provide more accurate estimates of subscores than raw subscores or theta estimates provided by a unidimensional IRT model calibrated on in-scale items only. However, the improvement in degree of accuracy depends on the relevance and amount of the ancillary information, which, in general, is guided by several factors. The first factor is the reliability of an in-scale subtest, which is closely related to the subtest length. If a subtest has high reliability, then the influence from the ancillary information will be relatively small. The second is the reliability and test length of out-of-scale subtests. Longer and more reliable out-of-scale subtests will provide more and consistent information. The third is the degree of correlation between subtests. Highly correlated out-of-scale subtests provide more relevant information.

Edwards and Vevea (2006) studied subscore augmentation using the EAP theta estimates based on summed scores (i.e., raw subscores) and individual item scores, respectively, as predictors, and the control factors were the number of in-scale items, the number of out-of-scale items, and the correlation of thetas. All items were dichotomous items, and the 3PL model was used. The results show that the augmented subscores in both cases were more accurate than the nonaugmented subscores (number correct). The magnitude of accuracy improvement through the use of an augmentation procedure increased when the correlations between subscales and the lengths of the ancillary subtests increased and the length (reliability) of the target subtest decreased.

Kahraman and Kamata (2004) studied their incorporation of out-of-scale information method using 3PL and EAP theta estimates with conditions varied on subtest length, correlation between subtests, and item discrimination power (low vs. high). The results show that the proposed method produced theta estimates having smaller standard errors in all conditions when compared to the traditional approach where the 3PL was applied only to in-scale items. The standard errors became smaller as the correlation between subtests increased and the number of out-of-scale items increased. However, the theta estimates from the incorporation of out-of-scale information method were more accurate than the traditional method only if the correlation between in-scale and out-of-scale items was high (at least >.5). When the out-of-scale items were highly discriminating, the increased accuracy was only observed if the correlation was larger than .9.

Yen (1987) studied the OPI in the tests composed of dichotomous items only. The ML theta estimates were based on the summed scores and the item parameter estimates from the 3PL model. The results indicated that the unadjusted OPI procedure underestimated both the length of the 67% credibility intervals of OPI scores and their standard errors, and that the adjusted OPI did too, but to a lesser degree. However, in a test with only dichotomous items, the unadjusted OPI was sufficiently accurate to report because the underestimation was small and was unlikely to have any practical importance. Yen et al. (1997) studied the OPI in the tests with mixed-item formats. The ML theta estimates were based on the item patterns and the item parameter estimates derived from the 3PL or GPCM. Their results show that OPIs resulted in subscores with substantially smaller standard errors than the percentage correct raw subscores. As the number of polytomous items increased, the degree of underestimation of both the length of the credibility intervals of OPI scores and their standard errors increased for both unadjusted and adjusted OPI. The adjusted OPI procedure performed better than the unadjusted one and was recommended for tests with mixed-item formats, especially if the number of polytomous items was large, because the credibility intervals and standard errors that it produced were sufficiently accurate for practical uses.

de la Torre and Patz (2005) studied the accuracy of theta estimates from the M3PL based on item patterns in a simple structure (i.e., an item belongs to only one subtest) estimated by the hierarchical Bayesian approach with the Markov chain Monte Carlo (MCMC) technique. The three factors considered were number of subtests, number of items, and correlation of thetas. Their study found that as the correlation between thetas, the number of items, and the number of subtests increased, the correlation between the true theta and the estimated EAP theta increased, and posterior variance of theta EAP estimates decreased. de la Torre (2008) did a similar study and obtained the similar findings using MGPCM on polytomous items which was also estimated by the hierarchical Bayesian method with MCMC. W. C. Wang, Chen, and Cheng (2004) compared the variances of EAP subscore estimates between the multidimensional 1PL model and

the 1PL model calibrated on in-scale items on two real test datasets using the maximum marginal likelihood estimation with the expectation–maximization algorithm. They found that the EAP subscore estimates from the multidimensional 1PL had the smaller variances on average than the 1PL; thus, the multidimensional 1PL was more efficient in subscore estimations than the 1PL. The similar results have been found in computer adaptive testing (e.g., Li & Schafer, 2005; W. C. Wang & Chen, 2004).

How Do the Subscore Estimation Methods Compare to Each Other in Terms of Accuracy and Stability of Subscore Estimates?

This is an important and interesting question, especially for those who want to choose among the various methods. Bock, Thissen, and Zimowski (1997) used resampling results from real data composed of dichotomous items and demonstrated that percentage correct IRT true scores calculated by 2PL and MGPCM EAP or ML theta estimates are more accurate than the percentage correct raw scores.

Shin (2007) compared five subscore estimation methods on mixed-format tests using simulated data sets: (a) percentage correct raw subscores, (b) 3PL/GPCM IRT true subscores based on item parameters estimated on in-scale items only, (c) unadjusted OPI with 3PL/GPCM theta estimates based on all items, (d) Wainer et al.'s (2001) score augmentation method using raw scores, and (e) a MCMC version of Method 4 (Shin, Ansley, Tsai, & Mao, 2005). The design factors were sample size, subtest length, correlation between subscale thetas, and ratio of number of constructed response items to multiple choice items. Theta estimates were based on item patterns. Shin found that Wainer et al.'s (2001) score augmentation method (Method 4) and its MCMC version (Method 5) yielded subscore estimates that had the highest R^2 and RMSE.

Yao and Boughton (2007) studied six subscore estimation methods on mixed-format tests using simulated and real data sets: (a) percentage correct raw subscores, (b) 3PL/GPCM EAP subscale thetas based on item parameters estimated on in-scale items only, (c) 3PL/GPCM ML subscale thetas based on item parameters estimated on all items, (d) M3PL/MGPCM EAP subscale thetas, (e) M3PL/MGPCM percentage correct IRT true subscore based on M3PL/MGPCM EAP subscale thetas, and (f) the unadjusted OPI with 3PL/GPCM EAP theta estimates based on all items. All the IRT theta estimates were based on item patterns, and all the IRT estimations used the Bayesian hierarchical approach with MCMC technique, except for the ML subscale theta estimation in Method 3. The design factors were sample size and correlation between thetas. Yao and Boughton compared the accuracy of subscale theta estimates among Methods 2, 3, and 4 and found that the accuracy of subscale theta estimates for Methods 2 and 4 was higher than that of Method 3 across all conditions and that the accuracy of Method 4 was higher than that of Method 2 when the correlations were at least medium high. They also compared the accuracy of subscore estimates between M3PL/MGPCM percentage correct IRT true subscore (Method 5) and OPI (Method 6) because their subscore estimates were on the same scale, and they found that the accuracy of Method 5 was higher than that of the OPI across all conditions and that the accuracy of the OPI increased with the correlation between subscale thetas and a correlation of .8–.9 was needed for unbiased subscale estimation. Regarding classification accuracy of performance levels, their result suggested that percentage correct raw subscores (Method 1) had the least accuracy.

de la Torre, Song, and Hong (2011) compared four subscore estimation methods on dichotomous items using simulated and real data sets: (a) M3PL EAP subscale thetas; (b) a multilevel 3PL model in which the first level is a 3PL model where EAP subscale thetas are estimated based on in-scale items and the second level is a regression model where each subscale theta is predicted by an overall EAP theta; (c) Wainer et al.'s (2001) subscore augmentation using the EAP theta estimates as predictors from the 3PL model on in-scale items; and (d) the unadjusted OPI with 3PL EAP theta estimates based on all items. As in Yao and Boughton (2007), all the EAP thetas were estimated using MCMC under the Bayesian framework based on item patterns. The design factors were number of subtests, subtest length, and correlation between thetas. de la Torre et al. (2011) concluded that the first three methods provided highly comparable subscore estimates, except for extreme thetas, where Methods 1 and 2 might perform better, while the OPI method produced the least accurate subscore estimates.

Fu and Qu (2010) conducted a more comprehensive comparison study on subscore estimation methods on dichotomous, polytomous, and mixed-format tests using simulated data. Each simulated test data set had two subtests and 2,000 examinees based on the two-dimensional PCM. The design factors were in-scale subtest length, out-of-scale subtest length, and correlations between two subscale thetas. The following six methods were compared: (a) percentage correct raw subscores, (b) percentage correct of Wainer et al.'s augmented scores with subtest raw scores as predictors, (c) percentage

correct IRT true subscores based on PCM and item parameters estimated on in-scale items, (d) percentage correct IRT true subscores based on PCM and item parameters estimated on all items, (e) the adjusted OPI with theta estimates based on PCM, and (f) percentage correct IRT true subscores based on multidimensional PCM (MPCM). The conditional likelihood estimation method was used for item estimation in PCM/MPCM. All theta estimates are EAP estimates based on item patterns. The main findings were as follows. First, in terms of accuracy and estimated R^2, the augmented score method (Method 2) and MPCM (Method 6) were the best methods when the subtest correlation was at least intermediate, and the subscore estimates from both methods were perfectly or near perfectly correlated. The two unidimensional IRT methods (Methods 3 and 4) basically had the same accuracy, and their subscore estimates were perfectly correlated across all conditions. The accuracy of the OPI (Method 5) was strongly related to subtest correlation. When the correlation was high, the OPI's accuracy was better than Methods 3 and 4 in most cases; however, when the correlation was low to intermediate, its accuracy was lower than that of the two unidimensional IRT methods and was even lower than the accuracy of percentage correct raw subscores (Model 1) in most cases. In general, subscore estimates from the OPI had the lowest correlations with those from other methods. This finding is consistent with the finding on the OPI in de la Torre et al. (2011) and Shin (2007). Percentage correct raw subscores had the lowest accuracy, except for the OPI cases mentioned earlier, which is consistent with the finding in Yao and Boughton (2007). However, Fu and Qu (2010) pointed out that the accuracy of the six methods was notably different only when the number of in-scale items was small (5 or 10). Second, in terms of R^2, the MPCM (Method 6) and PCM on in-scale items (Method 3) had the highest estimated R^2 when the correlation between subtest thetas was low, and the augmented score method and MPCM had the highest R^2 estimates when the correlation was at least intermediate. However, the differences among subscore R^2 estimates were notable only when subtest correlation was high and the number of in-scale items was smaller than 30.

Fu and Qu (2012) conducted a study similar to Fu and Qu's (2010). However, in Fu and Qu (2012), multidimensional GPCM (MGPCM) was used to simulate data, only mixed-format test data were generated, the GPCM or MGPCM was used for IRT calibrations, and item parameters were estimated by the marginal ML method. In addition, one more method was included for comparisons: the score augmentation method, with IRT theta subscores as predictors. The findings were similar to those in Fu and Qu (2010) except for the following result. The MGPCM and the score augmentation method with IRT theta subscores were the best methods in terms of accuracy and stability when the subtest correlation was at least intermediate and the correlations of their subscore estimates were 1 or almost 1. This result is consistent with the finding in de la Torre et al. (2011) as well as in de la Torre and Patz (2005) and Gessaroli (2004). The score augmentation method with raw subscores seems to perform well only when there were more out-of-scale items than in-scale items. This difference is understandable, considering the Rasch models were used in Fu and Qu (2010), whereas the two-parameter models were employed in Fu and Qu (2012). Note that in a Rasch model, raw score is a sufficient statistic for theta, whereas in a two-parameter model, it is not.

In summary, the consistent findings in the studies reviewed earlier are as follows:

- MIRT models and the score augmentation method with IRT theta subscores provided comparable subscore estimates and performed best among subscore estimation methods.
- In general, the subtest raw score was the least accurate and reliable.
- The OPI did not perform well when the correlations between subscores were low or intermediate.

As mentioned in Table 2, IRT thetas for subtests can be estimated by different methods. Some related research compared different approaches to IRT theta estimation on a whole test. Studies (e.g., Capar, 2001; Kim & Nicewander, 1993; Thissen & Orlando, 2001) showed that EAP theta estimation was preferred to ML estimation for its small bias and standard error. Yen (1984) proposed an ML estimation of thetas based on summed scores for dichotomous items by approximating a compound binomial distribution. The results showed that the ML theta estimates based on summed score and item pattern were very similar and were tau-equivalent; that is, a student's expected thetas from the two methods were the same, or in other words, for all groups of examinees, their mean thetas estimated from the two methods were the same. However, the theta estimates based on item patterns had lower standard errors, particularly for examinees with low scores. Yen concluded from her results that the estimated IRT true score scale was overall a more accurate and stable scale than the theta scale or linear transformations of theta. Thissen et al. (1995) provided a recursive algorithm to compute the ML, MAP, and EAP theta estimates based on summed scores for polytomous items. Their study indicated that the loss in accuracy for theta estimation based on summed scores was relatively small compared to that based on item patterns.

Discussion and Future Research

Comments on the Proposed Models

All of the proposed subscore estimation methods using collateral information improve the reporting of subscores in terms of accuracy and stability to some extent. However, each method is different in terms of its underlying assumptions, what collateral information is used, and how the information is incorporated into subscore estimation.

In CTT, the linear regression method is used to predict true subscores. Haberman's models and the PRMSE criterion provide quick tools for testing programs reporting raw scores to judge whether subscores should be reported in addition to total scores. Note that, as mentioned earlier, this issue is closely related to test dimensionality. Thus test dimensionality assessment methods, for example, factor analysis and conditional covariance-based dimensionality assessment methods (Stout et al., 1996), IRT-based methods, as well as those discussed in Wainer et al. (2001), can serve this purpose also. However, if a testing program wants to report true subscore estimates, then it is safe to report the estimates from subscore augmentation (Wainer et al., 2001), because it uses the most predictors (all subtest scores) and thus provides the most accurate estimates.

For the two methods that use a unidimensional IRT model to estimate subscores (i.e., Rows 5 and 6 in Table 1), both estimate subscale thetas (as well as the subscores derived from thetas) based on in-scale items; however, they differ on how item parameters are estimated. One method uses in-scale items, and the other uses all items in a test. It may be reasonable to argue that the method using item parameter estimates based on all items performs better than the one using item parameter estimates based on in-scale items if the correlations between subscores are high, and worse if the correlations are low (Yao & Boughton, 2007). However, Fu and Qu (2010, 2012) showed that these two methods provided very similar subscore estimates under the simple subscore structure across the design conditions. They argued that under simple structure, the impact of test heterogeneity was minor on item parameter estimates because item parameter estimates largely depended on the population distribution of thetas.

For Kahraman and Kamata's (2004) method, subscore estimates are based on all items in a test. However, this method aligns the parameters of the out-of-scale items to those of in-scale items one by one before using all items in a test to estimate subscale thetas. Because all items are used in theta estimation, test homogeneity does become relevant here, as Kahraman and Kamata indicated that high correlations between subtests were the critical requirement for this method to overperform the method using theta estimation based on in-scale items only. However, the scale alignment procedure seems to lack theoretical ground, and using thetas estimated from all items to represent a subscale is questionable.

Similarly, the OPI uses theta estimates based on all items as prior information and combines them with observed subscores to provide expected posterior subscore estimates. Hence test homogeneity is also relevant for the OPI. In addition, this method makes some assumptions that may not hold in certain cases, as discussed previously. Recent studies have shown that when the correlations between subtests were not high, the OPI performed less well than the score augmentation method and MIRT method (de la Torre et al., 2011; Fu & Qu, 2010, 2012; Shin, 2007; Yao & Boughton, 2007).

MIRT produces accurate theta estimates given an adequate item and examinee sample for stable calibration. Some studies (de la Torre et al., 2011; de la Torre & Patz, 2005; Fu & Qu, 2012; Gessaroli, 2004) have found that the score augmentation method using IRT theta scores as predictors produced highly comparable subscore estimates with the MIRT method, and both methods provided the most accurate subscore estimates. When test data fit a Rasch model, the score augmentation method using raw subscores was found to perform as well as MIRT (Fu & Qu, 2010). Actually, the MIRT and the score augmentation method can be seen as a general approach for subscore estimation in the sense that they take into account the magnitude of the correlations between subscores, while the other methods assume either that subscores are perfectly correlated or that they are not related to each other at all. A practical issue of applying MIRT to large-scale assessments is computational efficiency. It is well known that MIRT estimations may be unstable for models with too many parameters and impossible to carry out or very slow for large data sets (especially for the Bayesian estimation using MCMC) owing to the limitations of software and computer resources. Although in recent years, more efficient computer programs for estimating MIRT models have been developed to handle large data sets, for example, the MIRT package (Haberman, 2013), the mdltm program (von Davier, 2008), and the flexMIRT program (Cai, 2013), these issues still remain and prevent MIRT from being used in large-scale assessments. Instead, the common methods used in large-scale assessments are the unidimensional IRT method with item parameters estimated from all items and the OPI. The two methods are efficient on computation and assumed to provide subscore estimates with adequate accuracy. The assumption

is probably correct in most cases, as subscore estimates from different methods are highly correlated, and the differences may not have any practical impact on the inferences made from the subscore estimates, especially if a subtest is sufficiently long (e.g., 20 score points or more; Fu & Qu, 2010, 2012). On the other hand, the score augmentation method is an attractive method for large-scale tests in that it provides both efficient and accurate subscore estimation. Recent studies have recommended this method for subscore estimation.

Future Areas of Research

The following issues are of interest and could be areas of future research. First, the auxiliary information used in the subscore estimation methods described previously is examinees' responses to other subtests. Another type of auxiliary information that can be used to improve IRT estimations is the information outside the test, for example, examinees' demographic information. IRT models incorporating exterior auxiliary information are often referred to as multilevel IRT models or explanatory IRT models (Adams et al., 1997; de Boeck & Wilson, 2004; van den Noortgate, de Boeck, & Meulders, 2003). Mislevy (1987) and Mislevy and Sheehan (1989), for example, showed that incorporating examinees' demographic variables into a unidimensional IRT model could lead to more accurate ability estimates and consistent parameter estimates. Examinees' demographic variables can also be incorporated into MIRT models as covariates to subscore estimates, so that both types of auxiliary information are used for subscore estimations. de la Torre (2009) reported more stable and accurate EAP subscore estimates from the M3PL with examinees' demographics as covariates than the corresponding models ignoring either or both types of auxiliary information. However, research on the applications of multilevel IRT models on subscore estimations is limited, partly because of the lack of an efficient estimation method and computer program. Recently, Haberman (2013) developed the MIRT package, which can be used to estimate a variety of MIRT models with covariates on subscales including M3PL and MGPCM. This program employs a log-linear modeling approach and implements the maximum marginal likelihood method with the stabilized Newton–Raphson algorithm. This program is fast and can handle large data sets. It is anticipated that the MIRT package will facilitate research on multilevel IRT models, including their applications to subscore estimations.

Second, in some cases, subscores need to be reported at an aggregate level, such as class, school, district, and state. Haberman et al. (2009) proposed methods to check whether subscores are worth reporting at aggregate levels. However, research on aggregate-level subscores is scarce, and further studies are needed.

Third, although some studies have compared different types of theta estimation on a whole test (as described earlier), the impact of the different types of theta estimation (Table 2) on the quality of subscore estimation is rarely studied. In particular, the raw to scale score conversion based on the test characteristic curve needs to be checked carefully, as it is commonly used in practice.

Fourth, note that all the empirical studies on subscore estimation mentioned in this review assumed a simple test structure, because most large-scale tests have simple test structures based on their test blueprints. However, some cognitive diagnostic tests do have subtests that do not have simple structures. Thus it would be interesting to study these methods on tests with overlapping subtest structures.

Fifth, to establish a common scale on a subscore across test administrations, subscore equating is needed. Subscore equating poses more challenges than total score equating, as the stability of subscores is typically lower than that of total scores. While Puhan and Liang (2011) and Sinharay and Haberman (2011) have done some preliminary work in this area, more research is needed.

Finally, more well-designed studies can be carried out to better understand the performance of the models discussed in the report and provide guidelines for practical uses of these models.

Suggestions on Practical Uses of Subscore Estimation Methods

On the basis of the current research, the following preliminary recommendations can be made for using the subscore estimation methods in practice:

1. We need to determine if reporting subscores is justifiable from both content and statistical perspectives. From the content perspective, subscore reporting should be supported by the test design. If a test is not originally designed to provide subscore information, then no matter how good the statistical methods are, valid and reliable subscores

cannot be extracted (Sinharay, Puhan, & Haberman, 2011). From a statistical perspective, test data should support the subscore structure. Haberman's method as well as other test dimensionality assessment methods can be used to check if reporting subscores is warranted.

2. In general, the score augmentation method with theta subscores as predictors and MIRT can be used for subscore estimation. If test data fit a Rasch model, then the score augmentation method with raw subscores as predictors can be used too. The score augmentation method is much more computationally efficient than MIRT. However, if subtests are sufficiently reliable (e.g., Cronbach's alpha larger than .70), which often implies more items/raw score points in a subtest (e.g., 20 raw score points or more), then all the subscore estimation methods provide quite comparable results in terms of accuracy and stability.

Acknowledgments

Thanks are due to Lora Monfils, Marna Golub-Smith, Sandip Sinharay, and Gautam Puhan for their helpful suggestions and edits on early versions of this report. We are grateful to Ayleen Gontz for her editorial assistance.

Notes

1 Reliability and R^2 of subscore estimates for a subtest can both be defined as the squared correlation between estimated and true subscores. However, according to the definition of reliability, only when the true subscores are directly based on the observed item scores in the subtest can the squared correlation be referred to as reliability. Therefore the R^2 of augmented subscore estimates should not be referred to as reliability because of the regression effects from other subtests.

References

Ackerman, T. A. (1994). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement, 18*, 257 – 275.

Ackerman, T. A., & Davey, T. C. (1991, April). *Concurrent adaptive measurement of multiple abilities*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Ackerman, T. A., & Evans, J. A. (1994). The influence of conditioning scores in performing DIF analysis. *Applied Psychological Measurement, 18*, 329 – 342.

Adams, R. J. (2006, April). *Reliability and item response modeling: Myths, observations, and applications*. Paper presented at the 13th International Objective Measurement Workshop, Berkeley, CA. Retrieved from http://www.slideserve.com/kerryn/reliability-and-item-response-modelling-myths-observations-and-applications

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1 – 23.

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J. (2007). Modeling diagnostic assessment with Bayesian networks. *Journal of Educational Measurement, 44*, 341 – 359.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66*, 541 – 561.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement, 34*, 197 – 211.

Cai, L. (2013). *flexMIRT™ version 2: Flexible multilevel multidimensional item analysis and test scoring [computer software]*. Chapel Hill, NC: Vector Psychometric Group.

Capar, N. K. (2001). *Analyzing multidimensional response data structure represented by unidimensional IRT models* (Unpublished doctoral dissertation). Florida State University.

Davey, T., & Hirsh, T. M. (1991, April). *Examinee discrimination as measurement properties of multidimensional tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

de Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

De Champlain, A., & Gessaroli, M. E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. *Applied Measurement in Education, 11*, 231 – 253.

de la Torre, J. (2008). Multidimensional scoring of abilities: The ordered polytomous response case. *Applied Psychological Measurement, 32*, 355 – 370.

de la Torre, J. (2009). Improving the quality of ability estimates through multidimensional scoring and incorporation of ancillary variables. *Applied Psychological Measurement, 33*, 465 – 485.

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

de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT scoring. *Applied Psychological Measurement, 35*, 296 – 316.

DiBello, L. V., Roussos, L., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. *26*, pp. 979 – 1030). Amsterdam, Netherlands: Elsevier.

Edwards, C. M., & Vevea, J. L. (2006). An empirical Bayes approach to subscore augmentation: How much strength can we borrow? *Journal of Educational and Behavioral Statistics, 31*, 241 – 259.

Embretson, S. E. (1997). Multicomponent latent trait models. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305 – 322). New York, NY: Springer.

Fu, J. (2009, April). *Marginal likelihood estimation with EM algorithm for general IRT models and its implementation in R*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Fu, J., & Li, Y. (2007, April). *Cognitively diagnostic psychometric models: An integrative review*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Fu, J., & Qu, Y. (2010, April). *A comparison of subscore estimation methods on simulated data*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CA.

Fu, J., & Qu, Y. (2012). An evaluation of subscore estimation methods on mixed-format tests. Unpublished manuscript.

Gessaroli, M. E. (2004, April). *Using hierarchical multidimensional item response theory to estimate augmented subscores*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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

Haberman, S. J. (2013). *A general program for item-response analysis that employs the stabilized Newton-Raphson algorithm* (Research Report No. RR-13-32). Princeton, NJ: Educational Testing Service.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75*, 209 – 227.

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

Haberman, S., von Davier, M., & Lee, Y.-H. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous distributions* (Research Report No. RR-08-45). Princeton, NJ: Educational Testing Service.

Hattie, J. A. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*, 49 – 78.

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139 – 164.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258 – 272.

Kahraman, N., & Kamata, A. (2004). Increasing the precisions of subscale scores by using out-of-scale information. *Applied Psychological Measurement, 28*, 407 – 426.

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika*, *58*, 587 – 599.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy model for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*, 205 – 237.

Levy, R., & Svetina, D. (2010, May). *A framework for dimensionality assessment for multidimensional item response models*. Paper presented at the meeting of the National Council on Measurement in Education, Denver, CO.

Li, Y. H., & Schafer, W. D. (2005). Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. *Applied Psychological Measurement, 29*, 3 – 25.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60*, 523 – 547.

McDonald, R. P. (1997). Normal-Ogive multidimensional model. In W. J. van der Lindern & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 258 – 270). New York, NY: Springer.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11*, 81 – 91.

Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54*, 661 – 679.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19*, 73 – 90.

Puhan, G., & Liang, L. (2011). Equating subscores under the non equivalent anchor test (NEAT) design. *Educational Measurement: Issues and Practice, 30*, 23 – 35.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271 – 286). New York, NY: Springer.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25*, 193 – 203.

Reckase, M. D., Carlson, J. E., Ackerman, T. A., & Spray, J. A. (1986, June). *The interpretation of unidimensional IRT parameters estimated from multidimensional data*. Paper presented at the annual meeting of the Psychometric Society, Toronto, ON.

Roussos, L. A., DiBello, L. V., Stout, W. F., Hartz, S. M., Henson, R. A., & Templin, J. H. (2007). The fusion model skills diagnostic system. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275 – 318). New York, NY: Cambridge University Press.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*, 219 – 262.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*, 333 – 352.

Shin, C. D. (2007). *A comparison of methods of estimating subscale scores for mixed-format tests*. Retrieved from https://images.pearsonassessments.com/images/tmrs/tmrs_rg/EstimatingSubscaleScoresforMixedFormatItemsforPEMreportfinal.pdf?WT.mc_id=TMRS_A_Comparison_of_Methods_of_Estimating

Shin, C. D., Ansley, T., Tsai, T., & Mao X. (2005, April). *A comparison of methods of estimating objective scores*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, QC.

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. (2008). *Reporting subscores: A survey* (Research Memorandum No. RM-08-18). Princeton, NJ: Educational Testing Service.

Sinharay, S., & Haberman, S. J. (2011). *Equating of subscores and weighted averages under the NEAT design* (Research Report No. RR-11-01). Princeton, NJ: Educational Testing Service.

Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic subscores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research, 45*, 553 – 573.

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

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*, 331 – 354.

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

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345 – 354.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73 – 140). Mahwah, NJ: Lawrence Erlbaum.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*, 39 – 49.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring* (pp. 343 – 387). Mahwah, NJ: Lawrence Erlbaum.

van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28*, 369 – 386.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287 – 307.

von Davier, M., DiBello, L. V., & Yamamoto, K. (2008). Reporting test outcomes using models for cognitive diagnosis. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 151 – 176). Cambridge, MA: Hogrefe & Huber.

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

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L.,. .. Thissen, D. (2001). Augmented scores: "Borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343 – 387). Mahwah, NJ: Lawrence Erlbaum.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement, 38*, 147–163.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40*, 255–275.

Wang, M. (1986, April). *Fitting a unidimensional model to multidimensional response data*. Paper presented at the ONR Contractors Conference, Gatlinburg, TN.

Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement, 28*, 295–316.

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9*, 116–136.

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

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates for number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21*, 93–111.

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

Yen, W. M., Sykes, R. C., Ito, K., & Julian, M. (1997, April). *A Bayesian/IRT index of objective performance for a test with mixed-item types*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Suggested citation:

Fu, J., & Qu, Y. (2018). *A review of subscore estimation methods* (Research Report No. RR-18-17). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12203

Action Editor: Marna Golub-Smith

Reviewers: Gautam Puhan and Sandip Sinharay