Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

Estimating Optimal Weights for Compound Scores: A Multidimensional IRT Approach

Hendrika G. van Lier, Liseth Siemons, Mart A.F.J. van der Laar & Cees A.W. Glas

To cite this article: Hendrika G. van Lier, Liseth Siemons, Mart A.F.J. van der Laar & Cees A.W. Glas (2018) Estimating Optimal Weights for Compound Scores: A Multidimensional IRT Approach, Multivariate Behavioral Research, 53:6, 914-924, DOI: 10.1080/00273171.2018.1478712

To link to this article: https://doi.org/10.1080/00273171.2018.1478712

Published online: 21 Nov 2018.

MULTIVARIATE BEHAVIORAL RESEARCH
2018, VOL. 53, NO. 6, 914–924
https://doi.org/10.1080/00273171.2018.1478712

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Estimating Optimal Weights for Compound Scores: A Multidimensional IRT Approach

Hendrika G. van Lier[a], Liseth Siemons[b], Mart A.F.J. van der Laar[a], and Cees A.W. Glas[b]

[a]Department of Psychology, Health and Technology, Universiteit Twente; [b]Department of Research Methodology, Measurement and Data-analysis, Universiteit Twente

ABSTRACT

A method is proposed for constructing indices as linear functions of variables such that the reliability of the compound score is maximized. Reliability is defined in the framework of latent variable modeling [i.e., item response theory (IRT)] and optimal weights of the components of the index are found by maximizing the posterior variance relative to the total latent variable variance. Three methods for estimating the weights are proposed. The first is a likelihood-based approach, that is, marginal maximum likelihood (MML). The other two are Bayesian approaches based on Markov chain Monte Carlo (MCMC) computational methods. One is based on an augmented Gibbs sampler specifically targeted at IRT, and the other is based on a general purpose Gibbs sampler such as implemented in OpenBugs and Jags. Simulation studies are presented to demonstrate the procedure and to compare the three methods. Results are very similar, so practitioners may be suggested the use of the easily accessible latter method. A real-data set pertaining to the 28-joint Disease Activity Score is used to show how the methods can be applied in a complex measurement situation with multiple time points and mixed data formats.

KEYWORDS

Bayesian estimation; full-information factor analysis; item response theory; marginal maximum likelihood; multidimensional item response theory

Introduction

Researchers and practitioners in the fields of psychology, sociology, health, educational measurement and epidemiology often combine multiple measures into an index, i.e., a compound, or composite score. An example is the Economic Social and Cultural Status (ESCS) index used in the Programme for International Student Assessment (PISA) project, which is an index of economic, social and cultural status made up of subscales measuring Family Wealth, Cultural Resources, Home Educational Resources and the Educational and Income Level of the Parents (OECD, 2015). In this article, an index from health assessment is used as an example, that is, the 28-joint Disease Activity Score (DAS28), made up of four different measures that will be discussed below.

In classical test theory (CCT), expressions for the reliability of composite scores are well known (see, e.g., Feldt & Brennan, 1989; Rudner, 2005). Further, Mosier (1943) proposed a method to optimize the reliability of a linear composite by finding weights for each measure by minimizing the error variance in the index. So this method provides the linear combination of the multiple measures that maximizes the reliability.

Nowadays, IRT has emerged as an alternative statistical framework for addressing measurement problems, or rather, as an extension of CCT (see, e.g., Bechger, Maris, Verstralen, & Béguin, 2003). IRT provides a well-founded framework for the construction of measurement instruments, linking and equating measurements, and evaluation of test bias and differential item functioning. Further, IRT has provided the underpinnings for item banking, optimal test construction and various flexible test administration designs, such as multiple matrix sampling, and computerized adaptive testing. An important advantage of IRT over CCT is that missing data and complex data collection designs (such as adaptive tests, multistage tests, and booklet-rotation designs) can be easily accommodated. Since global reliability as defined in

CONTACT: Cees Glas c.a.w.glas@utwente.nl Department of Research Methodology, Measurement and Data-analysis, Universiteit Twente, Enschede, Netherlands.

IRT can be seen as a generalization of the reliability concept of CCT (Bechger et al., 2003), the current approach solves many problems associated with the estimation of reliability in more traditional fashions, especially in the case of missing data and complex data collection designs.

In IRT, the problem of composite scores can be tackled in the framework of multidimensional IRT (MIRT, see Ackerman, 1992, 1994, 1996; Bock, Gibbons, & Muraki, 1988; Reckase, 1985, 1997, 2009). In this approach, every measurement instrument loads on one specific latent dimension and the ensemble of latent dimensions has a multivariate normal distribution. This framework is, for instance, used to compute subscores (see, e.g., Haberman, 2008; Haberman, Davier & Lee, 2008; Haberman, Sinharay, & Puhan, 2009; Haberman & Sinharay, 2010; Reise, Bonifay, & Haviland, 2013; Sinharay, 2010; Sinharay, Puhan & Haberman, 2010, 2011) and to augment the test reliability by collateral information (see, e.g., Wainer, Vevea, Camacho, Reeve, Rosa, & Nelson, 2001). This framework is further generalized by Rijmen, Jeon, von Davier, and Rabe-Hesketh (2014) who present a model where the latent dimensions associated with the subscores load on a second-order latent principal component. Also relevant here is a study by Culpepper (2013) into the precision of the IRT-based approach relative to an approach based on CCT.

The problem addressed in the present article is how to estimate optimal weights to optimize the global reliability of an index based on multiple measures. Nowadays, the prominent frameworks for estimating MIRT models are marginal maximum likelihood (MML) and a fully Bayesian framework. The procedure for the estimation of optimal weights will be implemented in both. Estimation of multidimensional IRT models in the likelihood-based MML framework was outlined by Bock et al. (1988; also see, Bock & Schilling, 1997; Schilling & Bock, 2005). In the Bayesian approach, the posterior distribution of the optimal weights together with their posterior expectation (EAP) as point estimates and their credibility regions are obtained in an MCMC procedure. The procedure to compute the estimates is either implemented in a Gibbs Sampler (Gelfand & Smith, 1990) with data-augmentation developed for MIRT models (Albert, 1992; Béguin & Glas, 2001; Johnson & Albert, 1999) or in the general purpose sampler implemented in OpenBugs (Lunn, Spiegelhalter, Thomas, & Best, 2009; also refer to the package Jags, by Plummer, 2003). The second Bayesian approach needs an additional software application, because the optimization step needs to be performed in every iteration of the MCMC procedure. Therefore, the complete sample of all parameter draws must be saved first (the Coda option in OpenBugs) and the optimal weights are then afterwards computed using this saved information. Still, this last method is simple and easy-to-implement because apart from the tool needed to process the Coda file (available on the author's website) all computations can be made in the public domain software packages OpenBugs or Jags. The main motive to consider the other two procedures is to validate the third one.

The three frameworks discussed here will be labeled MML, MCMC-daug and OpenBugs.

The article is organized as follows. First, the IRT models are explained, the optimization problem is outlined, and the estimation procedures are discussed. Next, a number of simulation studies is presented to demonstrate the feasibility of the method and to compare the output of the three estimation procedures. Then, the model is applied to a complex measurement situation with multiple time points and mixed data formats (both discrete and continuous). The example pertains to the DAS28, a multiple-measures index for disease activity in Rheumatoid Arthritis (RA) patients. Finally, some suggestion for further research will be given.

IRT model and optimization problem

MIRT models can be seen as factor analysis models for discrete observations that use all available information in individual item response patterns; hence the alternative name full-information factor analysis (Bock et al., 1988). In fact, Takane and de Leeuw (1987) show MIRT models in a representation as used below are equivalent to factor analysis models for categorical data. The model will be presented here for polytomously scored items, with dichotomously scored items as a special case. So let the response variable y_{nqij} be equal to one if a person indexed n ($n = 1, \ldots, N$) gives a response in category j ($j = 0, \ldots, M_i$) of item i ($i = 1, \ldots, K_q$) of subscale q ($q = 1, \ldots, Q$). For dichotomously scored items, $M_i = 1$ leads to a special case. It is assumed that the responses on the items of every subscale are given according to the normal ogive representation of the Graded Response Model (GRM) by Samejima (1969). Alternative representations such as a logistic version of the GRM and the generalized partial credit model (Muraki, 1992) are possible, but Verhelst, Glas, and de Vries (1997) show that the results under the various representations are quite similar. In the GRM, conditional on a latent variable θ_{nq}, the probability of a response in category j ($j=0, \dots, M_i$) is given by

$$P_{ij}\left(\theta_{nq}\right) = \Pr\left(Y_{nqij} = 1|\theta_{nq}, a_{qi}, b_i\right) = \phi\left(a_{qi}\theta_{nq} - b_{qij}\right) - \phi\left(a_{qi}\theta_{nq} - b_{qi(j+1)}\right), \tag{1}$$

where $\phi(.)$ is the cumulative normal distribution. For the item location parameters, it holds that $b_{qij} < b_{qi(j+1)}$ and it is assumed that $b_{qi0} = -\infty$ and $b_{qi(M_i+1)} = \infty$. Further, referring to the analogy between MIRT and factor analysis pointed out by Takane and de Leeuw (1987), a_{qi} can be seen as a factor loading. Note that in the present application, it is assumed that every item loads on one latent dimension only. In IRT, this is referred to as between-items multidimensionality, as opposed to within-items multidimensionality, the case where an item loads on more than one of the dimensions (see, e.g., Reckase, 1997, 2009).

The second assumption of the model is that the latent variables $\theta_n = (\theta_{n1}, \dots, \theta_{nq}, \dots, \theta_{nQ})$ have a Q-variate normal distribution with a density given by

$$g(\theta_n|\mu, \Sigma). \tag{2}$$

Often, $\boldsymbol{\mu} = \boldsymbol{0}$ and Σ equals a correlation matrix to identify the model. These restrictions are also used in the simulation study. However, in multiple group applications the restriction on the means can be relaxed, for instance, by assuming that only the mean of one group is fixed and other groups have free means; this is also done in the real-data application presented below.

In the unidimensional case of latent variable modeling, reliability can be based on the variance decomposition

$$var(\theta) = var[E(\theta|\mathbf{y})] + E[var(\theta|\mathbf{y})],$$

where $\mathbf{y}$ is the person's observed response pattern, where $var(\theta)$ is the population variance of the latent variable, $var[E(\theta|\mathbf{y})]$ is the posterior variance of the expected person parameters (say, the EAP estimates of θ, and $E[var(\theta|\mathbf{y})]$ is the expected posterior variance of the EAP estimate. Reliability is given by the ratio

$$\rho = \frac{var[E(\theta|\mathbf{y})]}{var(\theta)} = 1 - \frac{E[var(\theta|\mathbf{y})]}{var(\theta)} \tag{3}$$

(see, Bechger et al., 2003). The middle expression is the variance of the estimates of the person parameters relative to the "true" variance, and the right-hand expression is one minus the average variance of the estimates of the person parameters relative to the "true" variance.

In a multidimensional case, reliability can be defined as follows. An estimate of the latent person parameter θ_n can be obtained by its posterior expectation given of a response pattern $\boldsymbol{y}_n$, that is, by $\frac{E(\boldsymbol{\theta}_n|\boldsymbol{y}_n)=\int\dots\int\boldsymbol{\theta}_n P(\boldsymbol{y}_n|\boldsymbol{\theta}_n)g(\boldsymbol{\theta}_n|\Sigma)d\theta_{n1},\dots,d\theta_{nQ}}{P(\boldsymbol{y}_n),}$ where

$$P\left(\boldsymbol{y}_n|\boldsymbol{\theta}_n\right) = \prod_{q,i,j} P_{ij}\left(\theta_{nq}\right)^{y_{nqij}} \tag{5}$$

is the probability of the response pattern, and

$$P(\boldsymbol{y}_n) = \int\dots\int P\left(\boldsymbol{y}_n|\boldsymbol{\theta}_n\right)g(\boldsymbol{\theta}_n|\Sigma)d\theta_{n1}, \dots, d\theta_{nQ} \tag{6}$$

is the marginal probability of response pattern $\boldsymbol{y}_n$.

Then

$$Cov\left(E(\boldsymbol{\theta}|\boldsymbol{y}), E(\boldsymbol{\theta}|\boldsymbol{y})^t\right)$$

is the covariance matrix of this estimate. If an index is defined as the linear combination $\boldsymbol{w}^t\boldsymbol{\theta} = \sum_q w_q\theta_q$, then for a compound index the definition of reliability in formula (3) generalizes to

$$\rho = \frac{\boldsymbol{w}^t Cov\left(E(\boldsymbol{\theta}|\boldsymbol{y}), E(\boldsymbol{\theta}|\boldsymbol{y})^t\right)\boldsymbol{w}}{\boldsymbol{w}^t\Sigma\boldsymbol{w}}. \tag{7}$$

This reliability can be optimized by choosing appropriate weights. This leads to the constraint maximization problem $\text{Max}[\boldsymbol{w}^t Cov(E(\boldsymbol{\theta}|\boldsymbol{y}), E(\boldsymbol{\theta}|\boldsymbol{y})^t)\boldsymbol{w}]$ with respect to $\boldsymbol{w}$, subject to the constraint $\boldsymbol{w}^t\Sigma\boldsymbol{w} = 1$. The constraint is chosen such that the variance of the compound scores is equal to one. Other choices for the variance are possible, but they will not change the weights because they figure in a variance ratio.

Solving the optimization problem

The solution is an adaptation of the solution of a more general problem solved by Albers, Critchley, and Gower (2011). As a first step, rewrite Equation (7) as

$$\rho = 1 - \frac{\boldsymbol{w}^t E\left(Cov\left(\boldsymbol{\theta}, \boldsymbol{\theta}^t|\boldsymbol{y}\right)\right)\boldsymbol{w}}{\boldsymbol{w}^t\Sigma\boldsymbol{w}}.$$

Introducing the notation

$$\boldsymbol{E} = E\left(Cov\left(\boldsymbol{\theta}, \boldsymbol{\theta}^t|\boldsymbol{y}\right)\right)$$

problem becomes

$$\begin{array}{l}\text{Minimize } \boldsymbol{w}^t\boldsymbol{E}\text{w} \\ \text{with respect to } \boldsymbol{w}^t\Sigma\boldsymbol{w} = 1.\end{array} \tag{8}$$

Define $\boldsymbol{E} = \boldsymbol{U}\Delta\boldsymbol{U}^t$ and $\boldsymbol{C} = \boldsymbol{E}^{-1/2}\boldsymbol{U}'\Sigma\boldsymbol{U}\ \boldsymbol{E}^{-1/2} = \boldsymbol{V}'\Gamma\boldsymbol{V}$. Next, we introduce the change of variables $\boldsymbol{z} = \boldsymbol{T}^{-1}\boldsymbol{w}$ with $\boldsymbol{T} = \boldsymbol{U}\boldsymbol{E}^{-1/2}\boldsymbol{V}$ so $\boldsymbol{T}^{-1} = \boldsymbol{V}^t\boldsymbol{E}^{1/2}\boldsymbol{U}^t$. As a result $\boldsymbol{T}^t\boldsymbol{E}\mathrm{T} = \boldsymbol{I}$, where $\boldsymbol{I}$ is an identity matrix, because $\boldsymbol{V}^t\Delta^{-1/2}\boldsymbol{U}\boldsymbol{E}\boldsymbol{U}\Delta^{-1/2}\boldsymbol{V} = \boldsymbol{V}^t\Delta^{-1/2}\boldsymbol{U}^t\boldsymbol{U}\Delta\quad\boldsymbol{U}^t\boldsymbol{U}\Delta^{-1/2}\boldsymbol{V} = \boldsymbol{I}$, $\boldsymbol{U}^t\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{V}^t\boldsymbol{V} = \boldsymbol{I}$ (orthonormal basis of eigenvectors). This changes to problem to

$$\text{Minimize } \boldsymbol{z}^t\boldsymbol{z}$$
$$\text{with respect to } \boldsymbol{z}^t\Gamma\boldsymbol{z} \;= 1.$$

However, this is a simple problem, since $\boldsymbol{z}^t\boldsymbol{z} = z_1^2 + z_1^2 + ... + z_Q^2$ with respect to $z_1^2\Gamma_1 + z_1^2\Gamma_2 + ... + z_Q^2\Gamma_Q = 1$ is minimized upon ordering $\Gamma_1, \Gamma_2, ..., \Gamma_Q$ from large to small and setting $z_1 = \Gamma_1^{-1/2}, z_2 = 0, ..., z_Q = 0$.

Marginal likelihood inference

Marginal inference in IRT models (see, Bock & Aitkin, 1981) proceeds by considering a likelihood function that is marginalized with respect to the latent person parameters, that is, the likelihood function of all item and population parameters is given by

$$L(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\mu}, \Sigma) = \prod_n P(\boldsymbol{y}_n)$$
$$= \prod_n \int \cdots \int P(\boldsymbol{y}_n|\boldsymbol{\theta}_n) g(\boldsymbol{\theta}_n|\Sigma) d\theta_{n1}, ..., d\theta_{nQ} \qquad (9)$$

where $P(\boldsymbol{y}_n|\boldsymbol{\theta}_n)$ is the probability of a response pattern $\boldsymbol{y}_n$ given the latent person parameter $\boldsymbol{\theta_n}$ (Equation (5)) and $g(\boldsymbol{\theta}_n|\Sigma)$ is the density of $\boldsymbol{\theta_n}$ (Equation (2)).

For solving the estimation equations, Bock et al. (1988) employ the EM algorithm (expectation-maximization algorithm, Dempster, Laird, and Rubin, 1977), where the values of $\boldsymbol{\theta_n}$ are seen as missing data. It handles missing data, firstly, by replacing missing values by a distribution of estimated values, secondly, by estimating new parameters, thirdly, by re-estimating the distribution of missing values assuming the new parameter estimates are correct, and fourth, re-estimate parameters, and so forth, iterating until convergence. The multiple integrals that appear above (e.g.,, Equations (4), (6), and (9)) can be evaluated using Gauss-Hermite quadrature. A critical point related to using Gauss-Hermite quadrature is the dimensionality of the latent space, that is, the number of latent variables that can be analyzed simultaneously. Bock et al. (2003) indicate that the maximum number of factors is 10 with adaptive quadrature, 5 with nonadaptive quadrature, and 15 with Monte Carlo integration.

Bayesian inference

There are several motives for choosing a Bayesian framework for estimation of the parameters in an IRT model. One of them is that all uncertainty regarding the parameters can be incorporated in the analysis. However, in IRT, another very important motive is that likelihood-based inference of the more complex versions of IRT models requires the evaluation of highly dimensional integrals which, at some point, becomes infeasible. Bayesian interference using MCMC computational methods does not suffer from this problem. In the MCMC approach, samples are drawn from the posterior distribution of all parameters and so the problem of complex multiple integrals does not arise here. Another important point is that along with the parameters, also complex functions of the parameters can be sampled, and, thus, also their credibility regions become available. This will be exploited below to obtain optimal weights and their credibility regions.

The MCMC-daug algorithm (-daug stands for data-augmentation) used in the present article is a combination of the algorithm for MIRT for dichotomously scored items by Béguin and Glas (2001) and the Metropolis-Hastings-algorithm for a unidimensional GRM by Johnson and Albert (1999). The procedure is a Gibbs Sampler (Gelfand & Smith, 1990) with data-augmentation (Tanner & Wong, 1987). The algorithm iterates over the following steps:

1. Draw augmented data Z_{nqi} given a draw of all other parameters. These variables are independent and normally distributed as
$$N\left(a_{qi}\theta_{nq}, 1\right) I\left(b_{qij} < Z_{nqi} < b_{qi(j+1)}\right) \text{ if } Y_{nqij} = j.$$
This step maps the discrete item responses into continuous responses. As a result, all remaining steps are based on regression models with normally distributed error terms.
2. Draw item parameters $\boldsymbol{b}$ given all other parameters using a Metropolis-Hastings step.

   Refer to Johnson and Albert (1999) for details.

   For the remaining steps, refer to Béguin and Glas (2001). These steps are:
3. Draw the item discrimination parameters $\boldsymbol{a}$ given the draw of all other parameters;
4. Draw values for the latent variables θ given the draw of all other parameters;
5. Draw the mean μ (if needed) and covariance matrix Σ from a normal-inverse Wishart distribution, given the draw of all other parameters.

After a number of burn-in iterations, the draws are draws from the full posterior distribution. The priors are independent standard normal for b_{qij}, independent normal with expectation 1 and standard deviation 10, confined to the positive values for a_{qi}, and an uninformative inverse-Wishart prior for μ and Σ.

Obtaining estimates of $Cov(E(\boldsymbol{\theta}|\boldsymbol{y}), E(\boldsymbol{\theta}|\boldsymbol{y})^t)$, of $E(Cov(\boldsymbol{\theta}, \boldsymbol{\theta}^t|\boldsymbol{y}))$, of the weights $\boldsymbol{w}$ and of ρ can be done in two ways. The first one is to evaluate of Equation (1) in every iteration of the MCMC-daug algorithm using Gaussian quadrature, and solve the optimization problem defined by Equation (8) to obtain a value for $\boldsymbol{w}$ and ρ. The second approach is to divide the MCMC-daug chain in a number of batches. Within every iteration, we compute $Cov(\boldsymbol{\theta}, \boldsymbol{\theta}^t|\boldsymbol{y})$, and over every batch we compute $E(Cov(\boldsymbol{\theta}, \boldsymbol{\theta}^t|\boldsymbol{y}))$ by taking the average over the iterations within the batch. So essentially, we are performing a Monte Carlo integration. Further, for every batch we solve the optimization problem to obtain a value for $\boldsymbol{w}$ and ρ. Then, the EAP estimate of $E(Cov(\boldsymbol{\theta}, \boldsymbol{\theta}^t|\boldsymbol{y}))$, $\boldsymbol{w}$, and ρ is computed as the mean over batches. Note that the computation of the optimal weights does not interfere with the MCMC estimation procedure. That is, the procedure behaves as normal and produces results that are analogous to the results obtained with a normal run without the optimization added.

Both approaches are examined in the simulation study using the MCMC-daug procedure. Implementing the first procedure in OpenBugs or Jags is problematic, so there only the second procedure will be used.

Simulation studies

Two sets of simulation studies are presented below, one for dichotomously scored items and one for polytomously scored items. The purpose of the simulation studies is to show what type of results can be expected in a number of typical situations, and to assess the differences between the three procedures.

Dichotomously scored items

Three simulation studies were made with dichotomously scores items. The first study was made to assess the effects of varying correlation, sample size and test length, the second study was made to assess the effects of unequal numbers of items in subscales, and the third study was made to assess interaction effects of differences in correlation and test length in subscales.

The setup of the first set of simulations can be inferred from Table 1. For all simulations, three subscales were used. The covariance matrix was a correlation matrix with equal correlations between subscales. In the first column, it is shown that this correlation was varied as 0.20 and 0.80. In the next columns, it can be seen that the sample size was varied as 400 and 1000 and that the test length of every subscale was varied as 5 and 9. The discrimination parameters were redrawn for every replication from a log-normal distribution with a mean equal to one, and a standard deviation equal 0.20. Both for the condition with 5 items and 9 items, the location parameters were equally spaced between −1.00 and 1.00.

Table 1. Effects of correlation between subscales (*R*), sample size (*N*), and subscale length (*K*) on weights (*W*), and reliability (ρ).

			MML Weights				MCMC-daug Weights				Open Bugs Weights			
R	*N*	*K*	*W1*	*W2*	*W3*	ρ	*W1*	*W2*	*W3*	ρ	*W1*	*W2*	*W3*	ρ
.20	400	5	.36	.27	.28	.74	.33	.32	.33	.75	.33	.32	.33	.75
		9	.32	.33	.30	.83	.34	.34	.31	.83	.34	.34	.31	.83
	1000	5	.32	.32	.32	.74	.34	.33	.33	.75	.34	.33	.33	.75
		9	.30	.35	.32	.82	.32	.34	.34	.82	.32	.34	.34	.82
.80	400	5	.59	.12	.08	.75	.33	.34	.35	.84	.33	.34	.35	.84
		9	.32	.32	.36	.90	.34	.33	.34	.90	.34	.33	.34	.90
	1000	5	.33	.30	.36	.82	.33	.34	.34	.84	.33	.34	.34	.84
		9	.35	.31	.34	.89	.34	.33	.34	.90	.34	.33	.34	.90

W1, *W2*, and *W3* are the weights for subscales 1, 2, and 3, respectively.

The columns labeled MML, MCMC-daug and OpenBugs give the estimates of the weights and reliabilities for the three procedures. For the MML and MCMC-daug procedures, 100 replications were made for each condition. For the MML procedure, the number of quadrature points was equal to 8 for each dimension, resulting in a grid of 512 points. Quadrature was adaptive (see, Schilling & Bock, 2005). For both MCMC procedures, 3000 burn-in iterations were followed by 8100 operational iterations. Checks on the convergence of the MCMC procedure showed this to be sufficient. To compute the posterior covariance matrices the 8100 operational iterations were divided into 90 batches with 90 iterations to compute each replicated posterior covariance matrix. Comparisons with the analogous computations using the explicit expression for the matrix and Gaussian quadrature showed that the results were very close. The number of within and between batches iterations will be returned to when discussing the results of the real-data example. Finally, the estimation procedure using OpenBugs is less suited for simulation studies, because it is difficult to run OpenBugs in batch, so here only 20 replications were used.

In Table 1, it can be seen that the agreement between estimated weights and reliabilities for the three estimations was very high. Also, the correlation between the estimates for the three methods was very high (not shown, but always above .90). Since all test lengths on all dimensions and the correlations

Table 2. Effects of varying correlation between subscales (*R*), test length of subscales (*K*) on weights (*W*) and reliability (ρ) MCMC-daug estimation method, 400 persons per replication.

	Subscale length			Average over 100 replications Weights				Standard error over 100 replications Weights			
R	*K1*	*K2*	*K3*	*W1*	*W2*	*W3*	ρ	*W1*	*W2*	*W3*	ρ
.20	9	21	21	.09	.43	.42	.91	.019	.035	.035	.004
	9	9	21	.10	.10	.63	.89	.028	.029	.040	.005
	5	9	21	.05	.10	.65	.90	.015	.030	.044	.006
.80	9	21	21	.17	.42	.42	.95	.010	.019	.018	.003
	9	9	21	.22	.22	.56	.93	.013	.013	.020	.004
	5	9	21	.13	.12	.62	.93	.010	.015	.027	.004

K1, K2, and *K3* are the numbers of items in subscales 1, 2 and 3, respectively.
W1, W2, and *W3* are the weights for subscales 1, 2 and 3, respectively.

between dimensions were all equal within each condition, average weights were all approximately equal. The variance across replications will be returned to when discussing the next study. Further, as expected, there were two main effects on the average reliability, that is, a main effect of the size of the correlation between dimensions and a main effect of the test length.

In the second set of simulations, the numbers of items in the subscales were varied. Overall, the setup was the same as in the first simulation study. Only sample sizes of 400 persons were considered here. Again, the agreement between the three methods was very high, so only the results of the MCMC-daug procedure are reported in Table 2, this time with the standard error over replications. In this set of simulations, there was no main effect of the test length on the attained reliability and its standard error. The attained reliability was very high and there seemed to be no room for further improvement by augmenting the test length. The weights were always largest for the longest subscale. However, the weights for the shorter subscales increased with the correlation. So for a correlation of .80, the contribution of the shorter tests to the overall reliability was weighted higher. Further, also the average reliability over the three conditions was slightly larger for a correlation of .80. Standard errors of the weights were smaller in this case.

In the third set of simulations, the correlation between the subscales was varied, along with the distribution of items across the subscales. Sample size was again 400, and 100 replications were made for each condition. This was done according to the following logic. Suppose a cluster is defined as a combination of two subscales. Two combinations are made: a cluster of two highly correlated subscales combined with a low correlation with the remaining cluster, or two subscales with a low correlation combined with a high correlation with the remaining cluster. These two configurators are crossed with test lengths as follows. Within the cluster we can have two short subscales, a short and a long subscale, or two long subscales. These three combinations are combined with a short or a long test for the subscale outside the cluster. Together this leads to $2 \times 3 \times 2 = 12$ combinations listed in Table 3.

In Table 3, it can be seen that, again, the weights were always largest for the longest subscale. The weights were higher when the correlation between the two subscales in a cluster was higher. Further, a high correlation of a cluster with the remaining subscale had an increasing effect on weight of the other subscale. Finally, again the attained reliability increased with the overall test length, but the pattern of the correlations did not have a discernible effect.

Polytomously scored items

The study with polytomously scored items generally had the same setup as the three studies regarding dichotomously scored items reported above. Sample size was again 400, and 100 replications were made for each condition. Results are reported in Table 4. Note that the correlations were varied as 0.20 and 0.80, and the subscale length was varied as 3 and 5.

In the fifth column, under the label M, it can be seen that the number of response categories was varied as 3 and 5, that is, as $M_i = 2$ and $M_i = 4$, respectively. The item parameters for the condition with $M_i = 2$ and subscale length 3 were fixed to $\beta_{i1} = \delta_i{-}0.5$ and $\beta_{i2} = \delta_i + 0.5$, where the mean item locations were fixed at $\delta_i = -0.25, \quad 0.00$ and 0.25. For the subscale length of 5, two items were added with mean item locations of $\delta_i = -0.50$ and 0.50. The item parameters for the condition with $M_i = 4$ and subscale length 3 were fixed to $\beta_{i1} = \delta_i{-}0.75$, $\beta_{i2} = \delta_i{-}0.25$,$\beta_{i3} = \delta_i + 0.25$ and $\beta_{i4} = \delta_i + 0.75$,

Table 3. Effects of varying correlations and subscale lengths on weights (*W*) and reliability (*ρ*) MCMC-daug estimation method, 400 persons per replication.

Correlation			Subscale length			Average over 100 replications Weights				Standard error over 100 replications Weights			
1,2	1,3	2,3	*K1*	*K2*	*K3*	*W1*	*W2*	*W3*	*ρ*	*W1*	*W2*	*W3*	*ρ*
.8	.8	.4	9	9	9	.38	.23	.30	.90	.017	.022	.021	.006
			9	9	21	.24	.17	.59	.93	.014	.017	.027	.004
			9	21	9	.24	.58	.18	.93	.013	.027	.016	.004
			9	21	21	.22	.41	.40	.94	.010	.023	.021	.004
			21	21	9	.43	.40	.14	.94	.016	.016	.017	.002
			21	21	21	.35	.31	.32	.95	.014	.014	.019	.002
.4	.4	.8	9	9	9	.27	.36	.36	.89	.024	.021	.017	.005
			9	9	21	.14	.22	.58	.92	.016	.015	.022	.005
			9	21	9	.14	.58	.22	.92	.017	.024	.012	.003
			9	21	21	.10	.42	.42	.94	.010	.016	.016	.003
			21	21	9	.53	.49	.02	.93	.017	.022	.017	.003
			21	21	21	.23	.23	.36	.95	.019	.012	.012	.004

K1, K2, and *K3* are the numbers of items in subscales 1, 2 and 3, respectively.
W1, W2, and *W3* are the weights for subscales 1, 2 and 3, respectively.

Table 4. Effects of correlation between subscales (*R*), varying length of subscales (*K*), and number of response categories of polytomously scored items (*M*) on weights (*W*) and reliability (*ρ*) MCMC-daug estimation method, 400 persons per replication.

	Subscale length				Average over 100 replications Weights				Standard error over 100 replications Weights			
R	*K1*	*K2*	*K3*	*M*	*W1*	*W2*	*W3*	*ρ*	*W1*	*W2*	*W3*	*ρ*
.20	3	3	3	2	.327	.316	.335	.727	.075	.068	.076	.009
				4	.324	.318	.325	.759	.075	.069	.065	.025
	3	3	5	2	.187	.174	.554	.776	.052	.049	.049	.017
				4	.193	.193	.542	.800	.048	.048	.042	.013
	3	5	5	2	.158	.396	.396	.797	.035	.069	.061	.001
				4	.157	.397	.404	.818	.038	.063	.057	.001
.80	3	3	3	2	.334	.340	.331	.829	.023	.030	.024	.000
				4	.334	.338	.333	.848	.027	.024	.027	.001
	3	3	5	2	.273	.264	.466	.853	.021	.025	.032	.018
				4	.268	.277	.455	.869	.027	.022	.032	.020
	3	5	5	2	.225	.397	.388	.873	.012	.024	.026	.001
				4	.231	.388	.387	.888	.019	.028	.014	.001

K1, *K2*, and *K3* are the numbers of items in subscales 1, 2, and 3, respectively.
W1, *W2*, and *W3* are the weights for subscales 1, 2 and 3, respectively.

where the mean item locations were fixed at $\delta_i = -0.50,\ 0.00$ and 0.50 . For the subscale length of 5, two items were added with mean item locations of $\delta_i = -1.00$ and 1.00. The item discrimination parameters were drawn from a lognormal distribution as indicated above.

In Table 4, it can be seen that the results were analogous to the results obtained with dichotomously scored items. The longer subscales obtained higher weights, and the shorter subscales obtained higher weights in the conditions with a correlation of .80 relative to the conditions with a correlation of .20. The reliability had a positive main effect of overall test length and of correlation, while standard errors over replications went down as the correlation went up. Further, for polytomously scored items, the reliability had a positive relation with the number of response categories.

A real-data example

The example pertains to the DAS-28, an index of disease activity for RA patients. The data are a set of patients drawn from the remission induction cohort of the Dutch Rheumatoid Arthritis Monitoring (DREAM) registry (Siemons et al., 2011). The remission induction cohort is a longitudinal observational, multicenter cohort of clinically diagnosed patients with early RA that aims to evaluate the effects of a protocolized treat-to-target strategy in daily clinical practice. The project started with recruiting patients in 2006 and data collection is still going on. However, the used data set was closed in 2012 with 546 patients included in the sample. These patients were measured at 0, 3, 6, 9 and 12 months. The DAS-28 consists of four different measures; that is, the 28-tender joint count (TJC28, a count of joints that are painful in the hands, shoulders, elbows, wrists and knees, 14 at each side of the body), the 28-swollen joint count (SJC28), a patient-reported measure of general health (GH), and a measure of inflammation consisting of either the erythrocyte sedimentation rate (ESR) (i.e., the DAS28-ESR) or the C-reactive protein (CRP) (i.e., the DAS28-CRP). Both indices, the DAS28-ESR and DAS28-CRP, are being used in practice. GH is measured on a 0 to 100 visual analog scale on general health, where 0 is very good and 100 is very bad. Tender joints and swollen joints are scored dichotomously and Siemons et al. (2011) showed that these counts fitted an IRT model (the Rasch model, Rasch, 1960) very well, although a log-likelihood ratio test showed that the 2-PL model had a significantly better fit to the TJC28 than the Rasch model (log-likelihood ratio test =163.81, df =27, $p < .01$).

The DAS28-ESR and DAS28-CRP were originally developed using linear regression. Since this article has no clinical interest, the ESR and CRP are included simultaneously. To create the present example, the values of the ESR, the CRP and GH were transformed such that they had a standard normal distribution at the first time point.

For some patients, measures at some time points were missing. The percent of missing data on each time point for the different instruments can be found in Table 5. Note that the percentage increases with time, this seems to imply that certain people drop out after a few measurement moments. The missingness is assumed to be at random and to have no relation with the outcome measures. One of the strong points of the present approach is that missing data do not complicate the estimation procedure. The missing data are just ignored and all estimates are based on the observed data only.

The model is as follows. It is assumed that all five measures at all time points t $(t = 1, \ldots, T,\ T = 5)$ loaded on a latent person parameter θ_{nqt}. Further, $\theta_{nqt} = \theta_{nq} + \delta_t + \varepsilon_{nqt}$, where θ_{nq} is a random person effect that is constant over time, δ_t is a change parameter and $\varepsilon_{nqt}$are independently normally distributed errors. The Q-dimensional latent variable $\boldsymbol{\theta}_{nt} = (\theta_{n1t}, \ldots, \theta_{nqt}, \ldots, \theta_{nQt})$ has a Q-variate normal distribution given by

$$\boldsymbol{\theta}_{nt} \sim N(\boldsymbol{\mu}_t, \Sigma)$$

where $\boldsymbol{\mu}_t$ has elements δ_t for $t = 1, \ldots, T$. Note that the covariance matrix is assumed constant over time, so Σ has elements $Cov(\theta_{nq}, \theta_{nq'})$. So though GH, ESRC and CRP are scalar-valued measures, the fact that they are measured repeatedly provides a measure of their reliability. The observations of the tender joints and the swollen joints were modeled by the 2-parameter normal ogive model, that is, the model given by Equation (1) with $M_i = 1$ for all i.

Table 5. Percent missing per time point for each instrument.

	TJC	SJC	GH	ESR	CRP
0 months ($t = 1$)	4.3	0.2	0.9	2.2	5.3
3 months ($t = 2$)	11.6	7.8	8.4	9.5	12.1
6 months ($t = 3$)	20.7	17.0	18.3	19.4	21.8
9 months ($t = 4$)	28.8	24.0	26.7	26.2	28.4
12 months ($t = 5$)	32.4	28.2	30.8	31.9	33.9

Table 6. Mean time effects, posterior standard deviation between brackets.

t	TJC	SJC	GH	ESR	CRP
1	−1.53 (.16)	−2.80 (.19)	−3.44 (.20)	−3.55 (.21)	−4.16 (.23)
2	−1.12 (.12)	−2.29 (.13)	−2.81 (.14)	−3.19 (.15)	−3.54 (.17)
3	0.58 (.04)	−0.03 (.04)	−0.10 (.04)	−0.22 (.05)	−0.26 (.05)
4	0.51 (.04)	−0.07 (.04)	−0.16 (.04)	−0.23 (.05)	−0.23 (.05)
5	0.49 (.04)	−0.11 (.04)	−0.18 (.05)	−0.17 (.05)	−0.16 (.05)

Table 7. Correlation matrix between instruments, posterior standard deviation between brackets.

	TJC	SJC	GH	ESR
SJC	0.58 (.05)	–	–	–
GH	0.74 (.04)	0.15 (.06)	–	–
ESR	0.02 (.06)	0.20 (.05)	0.02 (.06)	–
CRP	0.07 (.07)	0.26 (.06)	0.12 (.07)	0.77 (.04)

The model was estimated in OpenBugs. Priors for the item location parameters and the change parameters were independent standard normal distributions, the item discrimination parameters had independent log-normal distributions, and the inverse of the covariance matrix had a Wishart distribution. The number of burn-in iterations was 3000, then 21,000 iterations were made to estimate all model parameters.

The obtained change parameters δ_t are given in Table 6. Note that all disease indicators decreased over time, which was as expected.

The EAP estimate of the correlation matrix is given in Table 7, together with the posterior standard deviations. Note that the CRP and ESR correlated highly. As do the TJC and GH. The correlation between tender and swollen joints was moderate, 0.58. All other correlations were quite low, the correlation of ESR with GH and TJC was even far outside the 99% credibility region.

To obtain estimates of the weights and reliability the iterations following the burn-in iterations were divided into batches. Within a batch, we computed $E(Cov(\theta, \theta^t|\boldsymbol{y}))$ by taking the average of $Cov(\theta, \theta^t|\boldsymbol{y})$over the iterations. Then, the EAP estimate of $E(Cov(\theta, \theta^t|\boldsymbol{y}))$ was computed as the mean over batches. Further, for every batch we solve the optimization problem to obtain a value for $\boldsymbol{w}$ and ρ. To

Table 8. Estimates of weights and reliability, posterior standard deviation between brackets.

Number of iterations							
Within	Between	TJC	SJC	GH	ESR	CRP	Reliability
90	90	0.205 (.009)	0.236 (.011)	0.186 (.016)	0.172 (.027)	0.078 (.010)	0.840 (.003)
90	90	0.204 (.009)	0.235 (.013)	0.187 (.017)	0.178 (.026)	0.081 (.009)	0.841 (.003)
140	140	0.204 (.007)	0.230 (.010)	0.192 (.014)	0.185 (.025)	0.082 (.008)	0.837 (.003)
400	50	0.202 (.005)	0.225 (.007)	0.198 (.009)	0.200 (.015)	0.088 (.006)	0.832 (.001)
50	400	0.208 (.011)	0.242 (.015)	0.175 (.022)	0.155 (.030)	0.071 (.011)	0.849 (.004)

get an impression of the influence of the choice of the number of within and between batches iterations, several independent runs were made. The results are given in Table 8. It can be seen that the results were quite stable.

Discussion

Reliability of indices is an important issue when they are used as covariates in regression models (as in large-scale educational surveys such as PISA) or when they are used to classify respondents (such as patients with RA). It must be noted that, though maximal reliability is important in the applications addressed here, reliability is not the only thing that assessments might aims for. In educational testing, for instance, a test battery may be made up of several components where the weights must reflect the understanding and consequently the buy-in of various stakeholders. Therefore, reliability may be moderated by other, mostly content driven constraints for weighting the components. For such applications, research on maximal reliability under practical constraints is a useful topic, but beyond the scope of the present study. Here, the focus is on indices with a different purpose. An example is the ESCS index in the PISA project. The index has evolved over the years, but for the 2018 cycle, it is derived from questions about general wealth (based on several proxy variables including home possessions), parental education, and parental occupation (OECD, 2016, for more information on the index refer to Willms, 2006). The index is a combination of IRT scales and directly observable variables and maximizing its reliability is important because it functions as a predictor for educational achievement.

In this article, it was shown how the reliability of a compound score or index can be maximized by weighting the components making up the compound score. The problem was tackled in the framework of IRT, both because IRT is by now a much used framework for solving measurement and testing problems, and because of its flexibility in handling missing data and multiple groups. The simulation studies showed that the method behaves as expected. Of course, the simulation studies do not have the pretention of being exhaustive and many more conditions could be envisioned. However, the results clearly show the feasibility of the procedure. Since test length, subscale length, the average height of the correlation of the subscales and the number of response categories of polytomously scored items all positively contribute to test information, they are all positively related to the attained reliability of the compound score. Also, the weights behave as expected: the longer subscale has the higher weight and subscales that highly correlate with the longest subscale also gain in weight.

The simulations further showed that the statistical framework is not essential for the results. The OpenBugs scripts and the programs for the MML estimation, the MCMC estimation and the computations on the CODA file are available on the website https://www.utwente.nl/nl/bms/omd/Medewerkers/medewerkers/glas/. The real-data example shows that the method is very flexible and can easily be adapted for application in more complex measurement situation with multiple time points and mixed data formats (both discrete and continuous).

Further research pertains to the fact that the proposed procedure maximizes global reliability, that is, the extent to which random respondents from some population can be distinguished. IRT distinguishes between global and local reliability, where the latter refers the measurement precision locally on the ability scale, say the person-specific or score-specific reliability or measurement error. Optimal weighting to maximize local reliability in a multidimensional situation is not straight-forward. Several possibilities are open to define cutoff scores or cutoff lines in the multidimensional latent space, and it has to be whether decision rules are conjunctive, disjunctive of compensatory (see, e.g., Glas, 2014; Glas and Vos, 2010). Therefore, optimal weighting to maximize local reliability is one of the more interesting lines of research needing further attention.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest.

No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported by a grant.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institution is not intended and should not be inferred.

ORCID

Hendrika G. van Lier http://orcid.org/0000-0001-9539-5508
Cees A.W. Glas http://orcid.org/0000-0001-6531-5503

References

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67–91. doi:10.1111/j.1745-3984.1992.tb00368.x

Ackerman, T. (1994). Creating a test information profile for a two-dimensional latent space. *Applied Psychological Measurement*, *18*(3), 257–276. doi:10.1177/014662169401800306

Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, *20*(4), 311–330. doi:10.1177/014662169602000402

Albers, C. J., Critchley, F., & Gower, J. C. (2011). Quadratic minimization problems in statistics. *Journal of Multivariate Analysis*, *102*(3), 698–722. doi:10.1016/j.jmva.2009.12.018

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *17*(3), 251–269. doi:10.3102/10769986017003251

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Béguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, *27*(5), 319–334. doi:10.1177/0146621603257518

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*(4), 541–562. doi:10.1007/BF02296195

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM-algorithm. *Psychometrika*, *46*, 443–459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological*, *12*(3), 261–280. doi:10.1177/014662168801200305

Bock, R. D., & Schilling, S. G. (1997). High dimensional full-information item factor analysis. In M. Berkane (Ed.), *Latent variable modeling and applications of causality* (pp. 163–176). New York: Springer. doi:10.1007/978-1-4612-1842-5_8

Bock, R., Gibbons, R., Schilling, S., Muraki, E., Wilson, D., & Wood, R. (2003). *TESTFACT 4.0 computer software and manual*. Lincolnwood, IL: Scientific Software International.

Culpepper, S. A. (2013). The reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distribution. *Applied Psychological Measurement*, *37*(3), 201–225. doi:10.1177/0146621612470210

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

Feldt, L., & Brennan, R. (1989). Reliability. In R. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 105–146). New York, NY: The American Council on Education, MacMillan.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approach to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409. doi:10.1080/01621459.1990.10476213

Glas, C. A. W. (2014). Adaptive mastery testing using a multidimensional IRT model. In D. Yan, A. A., von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 205–218). Boca Raton, FL: Chapman and Hall/CRC.

Glas, C. A. W., & Vos, H. J. (2010). Adaptive mastery testing using a multidimensional IRT model. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 409–431). New York, NJ: Springer. doi:10.1007/978-0-387-85461-8_21

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., Davier, M., & Lee, Y. H. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions. ETS Research Report Series*, *2008*(2), i. doi:10.1002/j.2333-8504.2008.tb02131.x

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, *62*(1), 79–95. doi:10.1348/000711007X248875

Haberman, S. J., & Sinharay, S. (2010). Reporting subscores using item response theory. *Psychometrika*, *75*(2), 209–227. doi:10.1007/s11336-010-9158-4

Johnson, V., & Albert, J. H. (1999). *Ordinal data modeling*. New York: NJ: Springer. doi:10.1007/b98832

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*(25), 3049–3067. doi:10.1002/sim.3680

Mosier, C. I. (1943). On the reliability of a weighted composite. *Psychometrika*, *8*(3), 161–168. doi:10.1007/BF02288700

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 1992* (1), i–176. doi:10.1002/j.2333-8504.1992.tb01436.x

OECD. (2015). *PISA 2015 technical report*. Chapter 16. Retrieved form http://www.oecd.org/pisa/data/2015-technical-report/

OECD. (2016). *PISA 2018: Draft analytical frameworks, May, 2016*. Retrieved form https://www.oecd.org/pisa/data/PISA-2018-draft-frameworks.pdf

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*. Vienna, Austria.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Denmark Paedagogiske Institute.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9(4), 401–412. doi:10.1177/014662168500900409

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer. doi:10.1007/978-1-4757-2691-6_16

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer. doi:10.1007/978-0-387-89976-3

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*(2), 129–140. doi:10.1080/00223891.2012.725437

Rudner, L. (2005). Informed test component weighting. *Educational Measurement: Issues and Practice*, *20*(1), 16–19. doi:10.1111/j.1745-3992.2001.tb00054.x

Rijmen, F., Jeon, M., von Davier, M., & Rabe-Hesketh, S. (2014). A third-order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys. *Journal of Educational and Behavioral Statistics*, *39*(4), 235–256.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. doi:10.1002/j.2333-8504.1968.tb00153.x

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*, 533–555. doi:10.1007/S11336-009-9136-X

Siemons, L., ten Klooster, P. M., Taal, E., Kuper, I. H., van Riel, P. L. C. M., van de Laar, M. A. F. J., & Glas, C. A. W. (2011). Validating the 28-tender joint count using item response theory. *Journal of Rheumatology*, *38*(12), 2557–2564. doi:10.3899/jrheum.110436

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*(2), 150–174. doi:10.1111/j.1745-3984.2010.00106.x

Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic scores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research*, *45*(3), 553–573. doi:10.1080/00273171.2010.483382

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, *30*(3), 29–40. doi:10.1111/j.1745-3992.2011.00208.x

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408. doi:10.1007/BF02294363

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528–540. doi:10.1080/01621459.1987.10478458

Verhelst, N. D., Glas, C. A. W., & de Vries, H. H. (1997). A steps model to analyze partial credit. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 123–138). New York: Springer. doi:10.1007/978-1-4757-2691-6_7

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

Willms, J. D. (2006). *Learning divides: Ten policy questions about the performance and equity of schools and schooling systems*. Montreal: UNESCO Institute for Statistics.