Confirmatory Multidimensional IRT Unfolding Models for Graded-Response Items

Applied Psychological Measurement
2016, Vol. 40(1) 56–72
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621615602855
apm.sagepub.com

Wen-Chung Wang¹ and Shiu-Lien Wu²

Abstract

Most unfolding item response models for graded-response items are unidimensional. When there are multiple tests of graded-response items, unidimensional unfolding models become inefficient. To resolve this problem, the authors developed the confirmatory multidimensional generalized graded unfolding model, which is a multidimensional extension of the generalized graded unfolding model, and conducted a series of simulations to evaluate its parameter recovery. The simulation study on between-item multidimensionality demonstrated that the parameters of the new model can be recovered fairly well with the WinBUGS program. The Tattoo Attitude Questionnaire, with three subscales, was analyzed to demonstrate the advantages of the new model over the unidimensional model in obtaining a better model-data fit, a higher test reliability, and a stronger correlation between latent traits. Discussion on potential applications and suggestion for future studies are provided.

Keywords

item response theory, unfolding, multidimensional models, ideal-point, Bayesian, graded-response items

Graded-response items have been widely used in the social, behavioral, and health sciences to measure latent traits. In practice, testing time is typically limited, and one has to make a choice between measuring a specific latent trait with a high degree of precision or measuring a wide range of latent traits with much less precision. Test developers often have to sacrifice precision and develop several short tests to cover as many important latent traits as testing time allows. For example, the Job Satisfaction Survey (Spector, 1997) has nine subtests, each having four items with a 6-point Likert-type scale (disagree very much, disagree moderately, disagree slightly, agree slightly, agree moderately, agree very much). Given that scores of short tests can be terribly unreliable (low measurement precision), as suggested by the Spearman–Brown

¹The Hong Kong Institute of Education, Tai Po, Hong Kong

²National Chung Cheng University, Chiayi, Taiwan

Corresponding Author:

Wen-Chung Wang, Department of Psychological Studies, The Hong Kong Institute of Education, 10 Lo Ping Road, Tai Po, New Territories, Hong Kong.
Email: wcwang@ied.edu.hk

prophecy formula, it would be highly desirable if the reliability of scores for a short test could be increased to a more satisfactory level by adopting a more efficient statistical method.

When a test (it can be an educational test or a psychological inventory) consists of multiple subtests, and each subtest measures a distinct latent trait, multidimensionality occurs between items. In contrast, when a test consists of items that measure more than one latent trait simultaneously, multidimensionality occurs within items (Adams, Wilson, & Wang, 1997). In confirmatory factor analysis, between-item multidimensionality is referred to as simple factor structure, whereas within-item multidimensionality as complex factor structure. In this study, the authors focus on between-item multidimensionality.

In recent years, item response theory (IRT) models are commonly used. Most IRT models are unidimensional in that only a single latent trait is involved in the whole test. When a test is between-item multidimensional, fitting unidimensional IRT models is no longer appropriate because multiple latent traits are involved in the test. One may fit unidimensional IRT models consecutively to each of the subtests. Often, if not always, the latent traits measured by subtests within a test are correlated. The consecutive unidimensional approach is not statistically efficient because the inter-correlations between latent traits are ignored, and the measurement precision for each subtest cannot be improved by considering the inter-correlations.

Many multidimensional IRT models have been developed to fit between-item or within-item multidimensional tests (Reckase, 2009). Fitting multidimensional IRT models to a between-item multidimensional test not only preserves the subtest structure but also simultaneously calibrates all subtests, thus utilizing the correlations between subtests to increase the measurement precision of each subtest. Research shows that the greater the correlations between latent traits, the greater the number of subtests, thus resulting in a greater improvement in measurement precision, by fitting a multidimensional model to a between-item multidimensional test, compared with the consecutive unidimensional approach (Wang, Chen, & Cheng, 2004). When subtests are too short for the consecutive unidimensional approach to yield precise measures, the multidimensional approach will squeeze as much information as possible from all the data to yield measures that are more precise. In other words, the multidimensional approach enables test developers to develop several short subtests that measure a wide variety of latent traits without sacrificing too much measurement precision.

In many cases, the correlations between latent traits are also of great interest; for example, they are often used to evaluate the internal validity of an experiment or the concurrent validity or predictive validity of a test. In the consecutive unidimensional approach, the correlations can be estimated only indirectly. Specifically, one can first fit a unidimensional IRT model to each subtest separately, obtain person measures for each subtest, and then compute the Pearson correlation between person measures for every pair of subtests. Because the person measures contain a certain amount of measurement error (especially in the case of short subtests), the computed correlations underestimate the true correlations between latent traits, which is commonly known as “attenuation due to measurement error.” In the multidimensional approach, measurement error is taken into account directly and the correlations among latent traits are estimated jointly, and they are thus free from attenuation.

Many IRT models have been fit to Likert-type items, which can be classified into two approaches: the dominance approach and the ideal-point approach. It has been argued that responses to Likert-type items may be more consistent with the ideal-point approach than with the dominance approach (Roberts & Laughlin, 1996). This argument implies that attitude measures based on disagree–agree responses may be more appropriately developed with unfolding models of the ideal-point approach than with cumulative models of the dominance approach.

Recent research has witnessed a boom in theory and applications of IRT-based unfolding models (Carter et al., 2014; Chernyshenko, Stark, Drasgow, & Roberts, 2007; Roberts, 2008;

Stark, Chernyshenko, Drasgow, & Williams, 2006; Wang, Liu, & Wu, 2013). Major IRT-based unfolding models include the (generalized) hyperbolic cosine model (Andrich & Luo, 1993), generalized probabilistic unfolding models (Luo, 2001), and the (generalized) graded unfolding model (Roberts, Donoghue, & Laughlin, 2000, 2002; Roberts & Laughlin, 1996). Among these models, the generalized graded unfolding model (GGUM) appears to draw more attention from researchers and practitioners, partly because of the availability of free software (Roberts, 2001) and the flexibility of the generalized partial credit model (Muraki, 1992). The parameters in the GGUM can be estimated with a marginal maximum likelihood (MML) technique, a Markov chain Monte Carlo (MCMC) procedure, or a marginal maximum a posteriori method.

The GGUM is unidimensional. It is not statistically efficient for between-item multidimensional tests. Carter and Zickar (2011) investigated the influence of the proportion of items loading onto a second dimension and the degree of item bidimensionality on parameter estimation accuracy in the GGUM. The results suggested that the estimation error was more serious when the proportion and the bidimensionality were larger. To resolve these two problems, a few multidimensional unfolding models have been proposed. DeSarbo and Hoffman (1987) developed the multidimensional unfolding threshold model and created a joint space of persons and objects from binary choice data to investigate market structure. They defined a latent “disutility” variable to represent the weighted distance between the person’s ideal point and the coordinate for the object, the additive constant, and a stochastic error component. If the individual threshold value is greater than or equal to the value of the latent disutility variable, the person will choose the object, and vice versa. These researchers used a logit link function and the joint maximum likelihood for estimation. Takane (1996) proposed a maximum likelihood IRT model for multiple-choice data that postulated the probability of a person choosing an item as a decreasing function of the distance between the person point and the item category point in a multidimensional Euclidean space. The vector of the person point was assumed to follow a multivariate normal distribution with a mean zero vector and a diagonal variance–covariance matrix.

Maydeu-Olivares, Hernandez, and McDonald (2006) proposed the multidimensional ideal-point IRT model for binary data. This model imposed a linear function of latent traits and assumed that if a person point coincides with an item point in the space of the latent traits, the item will be endorsed. The item response function is an ideal hyperplane rather than a concentric circle. For example, when there are two dimensions, the model invokes an ideal line rather than an ideal point. Limited-information methods using only univariate and bivariate moments and the unweighted least squares estimator were used to estimate and evaluate the goodness of fit of the model.

Roberts and Shim (2010) extended the GGUM to an exploratory multidimensional unfolding model for polytomous response data. They adopted the MCMC estimation method implemented in the WinBUGS software (Spiegelhalter, Thomas, Best, & Lunn, 2003) and used a two-step procedure for parameter estimation. In the first step, the GGUM2004 software was adopted to obtain parameter estimates to serve as initial values for a WinBUGS estimation run. In the second step, the estimates from WinBUGS were resigned to serve as initial values for a second WinBUGS run. The estimation can take 15 days of computer time for 30 items and 2,000 persons. This model and the other multidimensional unfolding models mentioned above are “exploratory,” in that the dimensionality is explored from the data. The exploratory approach to dimensionality is often used at the initial stages of test development where the dimensionality is unclear, whereas the confirmatory approach is often used to validate or revise the theoretical structure of dimensionality. Furthermore, all models are limited to dichotomous items.

In this study, a confirmatory and multidimensional version of the GGUM for polytomous items, called the confirmatory multidimensional generalized graded unfolding model (the

CMGGUM), was developed. The model can be fit to between- and within-item multidimensional items. For demonstration, this study focuses on the between-item multidimensionality. In the following, after a brief introduction of the GGUM, the CMGGUM is specified. Parameter estimation methods and corresponding computer programs for the CMGGUM are described. A series of simulations was conducted to evaluate parameter recovery, and the results are summarized. A real data set about attitudes toward tattooing was analyzed to illustrate the implication and applications of the CMGGUM.

The GGUM

The GGUM is a unidimensional, symmetric, non-monotonic item response model (Roberts et al., 2000). When a person is asked to express agreement with an attitude item, he or she tends to agree with the item to the extent that it is located close to his or her position on a latent trait continuum. A person selects an observed response category for one of two reasons—He or she might disagree or agree with an item from below or above the item. In the dominance approach, where disagree responses are scored as a lower point and agree responses as a higher point, a person disagreeing with such a statement would be considered conservative. In contrast, in the ideal-point approach, the possibilities of being conservative and liberal are both considered.

In unfolding models, latent responses (i.e., the two reasons for each observed response) are assumed to follow a cumulative item response model. For example, the generalized partial credit model (Muraki, 1992) was used in the GGUM (Roberts et al., 2000). Of course, other cumulative item response models are also possible. Under the generalized partial credit model, the log-odds are defined as

$$\log \left(\frac{P(Y_i = k | \theta_j)}{P(Y_i = k - 1 | \theta_j)} \right) = \alpha_i (\theta_j - \delta_i - \tau_{ik}), \quad (1)$$

where Y_i is the latent response to item i , θ_j is the latent trait level of person j , α_i is the slope of item i , δ_i is the overall threshold of item i , and τ_{ik} is the k th threshold relative to the overall threshold of item i . For a 4-point Likert-type scale (e.g., strongly disagree, disagree, agree, strongly agree), there are eight latent responses—one pair per point.

Equation 1 defines an item response model at a latent response level. However, the model must ultimately be defined in terms of the observed response categories associated with the graded agreement scale. The two latent responses corresponding to a given observed response category are mutually exclusive. Therefore, the probability of a person selecting a particular category is the sum of the probabilities associated with the two corresponding latent responses:

$$\begin{aligned} P(Z_i = z | \theta_j) &= P(Y_i = z | \theta_j) + P(Y_i = (M - z) | \theta_j) \\ &= \frac{\exp \left\{ \alpha_i \left[z(\theta_j - \delta_i) - \sum_{k=0}^z \tau_{ik} \right] \right\} + \exp \left\{ \alpha_i \left[(M - z)(\theta_j - \delta_i) - \sum_{k=0}^z \tau_{ik} \right] \right\}}{\sum_{w=0}^C \left\{ \exp \left\{ \alpha_i \left[w(\theta_j - \delta_i) - \sum_{k=0}^w \tau_{ik} \right] \right\} + \exp \left\{ \alpha_i \left[(M - w)(\theta_j - \delta_i) - \sum_{k=0}^w \tau_{ik} \right] \right\} \right\}}, \quad (2) \end{aligned}$$

where Z_i is an observed response to item i ; $z = 0, 1, \dots, C$; C is the number of observed response categories minus 1; $M = 2C + 1$; τ_{ik} is symmetric about the point $\theta_j - \delta_i = 0$, which yields $\tau_{i(C+1)} = 0$, and $\tau_{iz} = -\tau_{i(M-z-1)}$ for $z \neq 0$. The premise of symmetric threshold parameters implies that persons are just as likely to agree with an item located either $-h$ or $+h$ units from their positions on the attitude continuum.

The CMGGUM

Let there be D dimensions. Under the CMGGUM, the probability of being in category z is

$$P(Z_i = z | \theta_j) = \frac{\exp\left\{\alpha_i \left[z(\mathbf{d}_i' \theta_j - \delta_i) - \sum_{k=0}^z \tau_{ik}\right]\right\} + \exp\left\{\alpha_i \left[(M - z)(\mathbf{d}_i' \theta_j - \delta_i) - \sum_{k=0}^z \tau_{ik}\right]\right\}}{\sum_{w=0}^C \left\{\exp\left\{\alpha_i \left[w(\mathbf{d}_i' \theta_j - \delta_i) - \sum_{k=0}^w \tau_{ik}\right]\right\} + \exp\left\{\alpha_i \left[(M - w)(\mathbf{d}_i' \theta_j - \delta_i) - \sum_{k=0}^w \tau_{ik}\right]\right\}\right\}}, \quad (3)$$

$$\theta_j \sim MVN(\mathbf{0}, \Sigma), \quad (4)$$

where θ_j is a D -dimensional vector representing the positions on D continuous latent traits for person j , Σ is a D by D variance-covariance matrix of the latent traits, α_i is a slope parameter for item i , $\mathbf{d}_i$ is a D -dimensional “design vector” specifying which dimensions are measured by item i , and the other parameters are defined as in Equation 2. When there is a single dimension, and $d_i = 1$ for every item, Equation 3 (CMGGUM) becomes Equation 2 (GGUM). Like in many multidimensional IRT models (Reckase, 2009), the location parameter of an item (δ_i) in the CMGGUM is a scalar rather than a vector. An item can have one discrimination parameter for each dimension but only a single location parameter across dimensions. Because $\mathbf{d}_i' \theta_j$ is a scalar, δ_i must be a scalar as well.

Specification of α_i and $\mathbf{d}_i$ as parameters or fixed values allows one to form customized models. For between-item multidimensional tests, where there are subtests and each subtest measures a distinct latent trait, $\mathbf{d}_i$ consists of a single 1 and $D - 1$ zeros, and α_i is a free parameter for a 2-parameter model, or a fixed value of unity for a 1-parameter model. For example, let there be three subtests, and each subtest measures a distinct latent trait with 10 items (i.e., Items 1-10 belong to Subtest 1, Items 11-20 belong to Subtest 2, and Items 21-30 belong to Subtest 3). Then, one can specify $\mathbf{d}_i = [1, 0, 0]'$ for Items 1 to 10, $\mathbf{d}_i = [0, 1, 0]'$ for Items 11 to 20, and $\mathbf{d}_i = [0, 0, 1]'$ for Items 21 to 30. Specifying α_i and $\mathbf{d}_i$ of Equation 3 this way essentially simplifies Equation 3 to Equation 2, because $\mathbf{d}_i' \theta_j$ equals θ_1 for Items 1 to 10, θ_2 for Items 11 to 20, and θ_3 for Items 21 to 30.

There are two ways to identify D -dimensional models. One way is to set the mean vector of latent traits at zero; the covariance matrix as an identity matrix; and slopes $\alpha_{id} = 0$ for $i = 1, \dots, D - 1$ and $d = i + 1, \dots, D$. The other way is to set $\delta_i = 0$ for $i = 1, \dots, D$; $\alpha_{id} = 1$ if $i = d$; and $\alpha_{id} = 0$ if $i \neq d$. Both ways set $D(D + 1)$ constraints. Following the suggestions, for the CMGGUM, one can constrain $\alpha_{id} = 1$ for $i = 1, \dots, D$ if $i = d$; and $\alpha_{id} = 0$ if $i \neq d$; and the mean vector of latent traits at zero, which also results in $D(D + 1)$ constraints. These constraints were applied in the following simulation study and empirical example.

Parameter Estimation

The GGUM is a highly parameterized model because it incorporates a large number of parameters. The CMGGUM is parameterized more. As with the GGUM, MML estimation methods can be developed to estimate the parameters in the CMGGUM. However, MML estimation methods become inefficient for the CMGGUM, especially when there are many latent traits (e.g., more than three), due to the complexity in integrating high dimensions, computing second derivatives, and inverting matrices. Roberts and Thompson (2011) developed a marginal maximum a posteriori estimation procedure for the GGUM to improve the accuracy and variability

of MML and the computational efficiency of the MCMC method, especially when the number of item response categories is small and the item locations are extreme. Similarly, the methods are inefficient when the dimensionality is high and/or the model is complicated. Therefore, in this study, the MCMC estimation method was adopted for the CMGGUM. WinBUGS was used to estimate the parameters with the default option of Gibbs sampler. The Gibbs sampler generates random values from a series of unidimensional posterior distributions, rather than a whole complex and multidimensional one. It accepts candidate point and does not need to specify proposal distribution.

Under the Bayesian framework, model comparison is often conducted by assessing the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002) and the pseudo-Bayes factor (PsBF; Gelfand & Dey, 1994). The DIC is comparable with Akaike's information criterion and Bayesian information criterion, and can be obtained directly from WinBUGS. The smaller the DIC, the better the model-data fit. The DIC should be used with caution. When the posterior distributions are highly skewed or bimodal, the DIC becomes inappropriate, because the posterior mean is no longer a good measure of central location.

The PsBF is defined as

$$\text{PsBF}_{12} = \prod_{i=1}^N \frac{f(\mathbf{y}_{\setminus i}, \text{model 1})}{f(\mathbf{y}_{\setminus i}, \text{model 2})} = \prod_{i=1}^N \frac{\text{CPO}_i(\text{model 1})}{\text{CPO}_i(\text{model 2})}, \quad (5)$$

where y_i is the i th observation, $\mathbf{y}_{\setminus i}$ is $\mathbf{y}$ after omitting y_i , and $f(\mathbf{y}_{\setminus i}, \text{model 1})$ is the leave-one-out cross-validation conditional predictive ordinate (CPO) for y_i under Model 1. It is very tedious to run the MCMC procedure N times by omitting one observation at a time. A more efficient method is to compute CPO by using the inverse of harmonic mean (Raftery, Newton, Satagopan, & Krivitsky, 2007):

$$\widehat{\text{CPO}}_i = f(y_i | \mathbf{y}_{\setminus i}) = \left(\frac{1}{T} \sum_{t=1}^T \frac{1}{f(y_i | \omega^{(t)})} \right)^{-1}, \quad (6)$$

where $t = 1, \dots, T$ is the i th iteration after burn-in. Substituting the value of Equation 6 for Equation 5 yields the PsBF. When the PsBF is greater than 1, Model 1 is favored.

In addition to model comparison, the posterior predictive check (PPC) approach (Meng, 1994) can be used to check absolute model-data fit. If the model fits the observed data, then the observed data should be very similar to the data generated from the model. The PPC computes the probability that the value of a given test statistic computed from the simulated data is as or more extreme than the value computed from the observed data. In this study, the Bayesian chi-square (Gelman, Meng, & Stern, 1996) was adopted as the test statistic. If the PPC p value is very close to 0 or 1, the model-data fit is considered poor.

The Simulation Study

Design

Two factors were manipulated: (a) sample size (500 and 1,500) and (b) correlations among latent traits (low: $r = .03$, median: $r = .60-.80$, and high: $r = .90$), resulting in a total of six situations. In the median condition, the correlation was .6 between θ_1 and θ_2 , .7 between θ_1 and θ_3 , and .8 between θ_2 and θ_3 . There were 45 four-point items in three subtests: Items 1 to 15 measured the first latent trait, Items 16 to 30 measured the second latent trait, and Items 31 to 45

measured the third latent trait. The item location parameters were generated from $U(-2, 2)$. The slope parameters were generated from a lognormal distribution with mean zero and variance 0.25, which has been used by Roberts and Thompson (2011). The first, second, and third thresholds were generated from $N(-2, 0.04)$, $N(-1, 0.04)$, and $N(-0.5, 0.04)$, respectively. The small variances in the thresholds made the inter-threshold distances large and the thresholds unlikely to be disordered. These settings were similar to those investigated by Roberts et al. (2002). A Matlab computer program was written to generate data. Thirty replications were performed.

Following the common practice in IRT that item parameters are often treated as fixed effects and person parameters as random effects, the item slope, location, and threshold parameters in the CMGGUM were treated as fixed effects, and the person parameters as random effects. Thus, the item parameters remained unchanged across replications, whereas the person parameters were regenerated in every replication. The model was identifiable because all items were unidimensional.

Analysis

The WinBUGS was used to analyze the generated data based on the true model. The prior distributions of the parameters were set as follows: $\delta_i \sim N(0, 4)$, $\alpha_i \sim \log N(0, 0.25)$, $\tau_{i1} \sim N(-2, 4)$, $\tau_{i2} \sim N(-1, 4)$, $\tau_{i3} \sim N(-0.5, 4)$, $\theta_n \sim MVN(\mathbf{0}, \Sigma)$, and Σ follows a Wishart distribution with a large variance. A brief simulation study was conducted to evaluate the convergence of the Markov chain. The Geweke (1992) diagnostic method yielded z scores between ± 1.96 for approximately 95% of parameters, indicating good evidence for convergence. The authors thus decided to use 4,000 iterations as burn-in, followed by an additional 2,000 iterations. The bias and root mean square error (RMSE) of an estimator were computed, where the expected a posteriori estimate of the parameter was used as a point estimate.

Results

Due to space constraints, the bias and RMSE for the slope, location, threshold, and variance–covariance parameters are not shown (but available upon request). Instead, their values were averaged. In calculating the mean absolute bias, the absolute bias values were taken and then averaged across parameters. Figure 1 shows the mean absolute bias values for the item and person parameters. As the persons were assumed to be randomly selected from a multivariate normal distribution, only the distributional parameters were estimated. For individual persons, one can use expected a posteriori or maximum a posteriori estimation. A larger sample size appeared to yield a smaller mean absolute bias; the correlation among latent traits appeared to have little impact on the slope, location, and threshold parameters, but a lower correlation yielded a smaller mean absolute bias on the variance–covariance parameters. The mean absolute bias values for the slope, location, and threshold parameters were not serious; they were between about 0.05 and 0.10 for $N = 1,500$, and between 0.10 and 0.17 for $N = 500$. The mean absolute bias values for the variance–covariance parameters were slightly larger than those for the item parameters. This finding was not surprising, because random-effect parameters are often more difficult to estimate than fixed-effect parameters.

Figure 2 shows the mean RMSE values. As expected, a larger sample size yielded a smaller mean RMSE. The correlation among latent traits appeared to have little impact on the slope, location, and threshold parameters, but a lower correlation yielded a smaller mean RMSE on the variance–covariance parameters. The mean RMSE values for the variance–covariance parameters were slightly larger than those for the slope, location, and threshold parameters. In

Figure 1. Mean absolute bias values of the slope, location, threshold, and variance–covariance parameters in the simulation study.

general, the magnitudes of bias and RMSE were comparable with those for the GGUM with similar settings (Roberts et al., 2002; Roberts & Thompson, 2011).

An Empirical Example

A real data set of attitude toward tattooing (Lin & Huang, 2005) was analyzed, which was retrieved from the Survey Research Data Archive supported by Academia Sinica in Taiwan, <https://srda.sinica.edu.tw/search/gensciitem/835>. A total of 1,172 juvenile delinquents were surveyed by cluster sampling. Most of them had dropped out of junior high school, were unemployed with low socioeconomic status, or had one or two offense records. Approximately half had tattoos. Their responses to the Tattoo Attitude Questionnaire were analyzed. The questionnaire consisted of three scales: Positive Evaluation for Self-Willingness Toward Tattooing (PESW), Positive Evaluation Toward Tattooing (PET), and Negative Evaluation Toward Tattooing (NET). Each scale had 9 to 10 four-point Likert-type items (0 = *strongly disagree*, 1 = *disagree*, 2 = *agree*, 3 = *strongly agree*). The questionnaire was between-item multidimensional, because items in the same scale measured the same latent traits; there were three latent traits. An example item was “Tattooing is brave.” For each observed response, there were two opposite latent responses. For example, a person might “disagree” because either tattooing is too negative to be brave (very negative attitude toward tattooing) or tattooing is too common to be brave (very positive attitude toward tattooing). In other words, the item was located at either (far) lower or (far) higher than the person’s point. Likewise, a person might “agree” because either the item was located at (slightly) lower or (slightly) higher than the person’s point.

Figure 2. Mean RMSE of the slope, location, threshold, and variance–covariance parameters in the simulation study.

Note. RMSE = root mean square error.

A three-dimensional CMGGUM was fit to the data, in which Items 1 to 10 measured PESW, Items 11 to 20 measured PET, and Items 21 to 29 measured NET. The WinBUGS codes are listed in the appendix. For comparison, the consecutive unidimensional approach was also adopted, in which the GGUM was fit to each scale consecutively. Both the multidimensional and unidimensional approaches were conducted with WinBUGS. The prior distributions of all parameters were set the same as those in the simulation study except the prior for the first location parameters. The prior distribution for the most negative item in each subtest was set as $U(-3, -0.5)$ because of its extremity. The Geweke diagnostic method was adopted to check convergence (6,000 iterations as burn-in, followed by an additional 4,000 iterations). The DIC and PsBF were used for model comparison. One single PsBF value was obtained in the unidimensional approach because the three scales were analyzed in one computer run with WinBUGS. The PPC with test quantity of Bayesian chi-square was performed to assess model-data fit. In addition, test reliability and correlations between latent traits were computed.

Results

Model–data fit. According to the Geweke diagnostic method, the Markov chains converged. The effective number of parameters (PD) was negative in the unidimensional approach, so the DIC obtained for this model was probably not to be trusted. The negative value of PD could have resulted because the posterior distributions were highly skewed or bimodal (Spiegelhalter et al., 2002). With respect to the PsBF, the logarithmic CPOs were $-34,306$ in the unidimensional approach and $-34,278.8$ in the multidimensional one. The logarithmic PsBF value for these

two approaches was thus 27.2, suggesting that the multidimensional approach had a better fit. Therefore, these three latent traits were not independent. The PPC p value for the Bayesian chi-square was .52 in the multidimensional approach, higher than the nominal value of .05. Hence, the multidimensional approach not only had a better fit than the unidimensional approach but also had a good fit to the data. Table 1 lists the parameter estimates for the three scales obtained with the multidimensional approach. The variation in the slope parameters in PESW (ranging from 0.96 to 1.98) was slightly larger than that in the other two scales.

In the multidimensional approach, each item measured one latent trait (but items in different scales measured different latent traits). Thus, the item characteristic curve, item information function, and test information function were identical to those in the GGUM (Roberts et al., 2000). For illustration, Figure 3 shows the item characteristic curves of the item—Tattooing is brave. The estimates for the slope, location, and three threshold parameters were 1.50, 1.16, -2.38 , -0.34 , and 0.40 , respectively. The most probable response was “strongly disagree” for persons with $\theta < -1.22$ or $\theta > 3.54$; “disagree” for persons with $\theta < 0.75$ or $\theta > 1.56$; “agree” for persons with $0.75 < \theta < 1.56$; “strongly agree” was never the most probable response.

Test reliability. As shown in Table 2, the reliability for the three scales was .915, .899, and .816, respectively, in the multidimensional approach, but was .865, .832, and .812, respectively, in the unidimensional approach. As expected, the multidimensional approach improved test reliability by considering the correlations among latent traits. According to the Spearman–Brown prophecy formula, the scales would have to be increased by 68%, 79%, and 3% in length, respectively, for the unidimensional approach to achieve test reliability as high as that of the multidimensional approach. Such a test length increment demonstrates the relative measurement efficiency of the multidimensional approach over the unidimensional approach.

Correlations among latent traits. The correlations among the three latent traits can be directly estimated only in the multidimensional approach. In the unidimensional approach, the Pearson correlation on the person measures of the three scales was computed. The multidimensional approach yielded a correlation of .80 (.56; the correlations in the parentheses were derived from the unidimensional approach) between PESW and PET, .39 (.19) between PESW and NET, and .30 (.05) between PET and NET. Compared with the multidimensional approach, the unidimensional approach underestimated the correlations among latent traits quite substantially, because measurement errors were not considered in the unidimensional approach (Wang et al., 2004). The high correlation between PESW and PET ($r = .80$) enabled the multidimensional approach to improve the test reliabilities.

Discussion and Conclusion

In this study, the CMGGUM was developed. Specifying α_i and $\mathbf{d}_i$ as fixed values or parameters, one can fit the CMGGUM to between-item multidimensional tests as 1-parameter or 2-parameter models. The CMGGUM is more efficient than the unidimensional GGUM for between-item multidimensional tests, because the former considers the correlations among latent traits, whereas the latter ignores them. Such a consideration enables the CMGGUM to improve test reliability and estimate correlations among latent traits accurately.

The parameters of the between-item CMGGUM can be recovered fairly well with WinBUGS, as demonstrated in the simulation study. The empirical example of attitudes toward tattoos illustrates the advantages of the multidimensional approach (the CMGGUM) over the consecutive unidimensional approach (the GGUM) in obtaining a better model-data fit, a higher reliability for the three scales, and a stronger correlation among latent traits. The computer time

Table 1. Parameter Estimates and Standard Errors From the Multidimensional Approach for the Three Scales of the Tattoo Attitude Questionnaire.

PESW			PET			NET		
Parameter	Estimate	SE	Parameter	Estimate	SE	Parameter	Estimate	SE
α_2	1.90	0.08	α_{12}	0.71	0.05	α_{22}	1.91	0.07
α_3	1.48	0.10	α_{13}	1.22	0.10	α_{23}	1.90	0.07
α_4	1.50	0.14	α_{14}	0.74	0.06	α_{24}	1.88	0.09
α_5	1.69	0.12	α_{15}	1.76	0.12	α_{25}	1.90	0.08
α_6	1.89	0.08	α_{16}	1.42	0.11	α_{26}	1.70	0.13
α_7	1.94	0.06	α_{17}	1.23	0.12	α_{27}	1.87	0.09
α_8	1.98	0.02	α_{18}	1.13	0.10	α_{28}	1.20	0.10
α_9	0.96	0.10	α_{19}	1.05	0.09	α_{29}	1.40	0.12
α_{10}	1.96	0.04	α_{20}	1.11	0.09			
δ_1	-2.51	0.15	δ_{11}	-2.01	0.01	δ_{21}	-2.09	0.08
δ_2	1.32	0.06	δ_{12}	1.55	0.16	δ_{22}	-1.49	0.15
δ_3	1.32	0.08	δ_{13}	0.81	0.10	δ_{23}	-1.67	0.19
δ_4	1.16	0.07	δ_{14}	1.49	0.11	δ_{24}	-1.66	0.14
δ_5	1.22	0.06	δ_{15}	1.11	0.09	δ_{25}	-1.62	0.13
δ_6	1.21	0.07	δ_{16}	1.22	0.09	δ_{26}	-1.43	0.23
δ_7	0.72	0.06	δ_{17}	0.74	0.11	δ_{27}	-1.52	0.13
δ_8	0.93	0.06	δ_{18}	0.75	0.08	δ_{28}	-1.78	0.16
δ_9	0.79	0.07	δ_{19}	0.89	0.12	δ_{29}	-1.70	0.17
δ_{10}	0.75	0.06	δ_{20}	0.73	0.10			
$\tau_{1,1}$	-5.15	0.19	$\tau_{11,1}$	-3.80	0.10	$\tau_{21,1}$	-5.44	0.27
$\tau_{1,2}$	-3.30	0.16	$\tau_{11,2}$	-1.19	0.09	$\tau_{21,2}$	-3.50	0.18
$\tau_{1,3}$	-0.73	0.17	$\tau_{11,3}$	0.68	0.14	$\tau_{21,3}$	0.58	0.20
$\tau_{2,1}$	-2.28	0.07	$\tau_{12,1}$	-3.85	0.19	$\tau_{22,1}$	-2.88	0.17
$\tau_{2,2}$	-0.24	0.08	$\tau_{12,2}$	-1.08	0.15	$\tau_{22,2}$	-2.17	0.16
$\tau_{2,3}$	0.22	0.10	$\tau_{12,3}$	1.21	0.20	$\tau_{22,3}$	-0.70	0.17
$\tau_{3,1}$	-1.94	0.08	$\tau_{13,1}$	-3.43	0.17	$\tau_{23,1}$	-2.97	0.20
$\tau_{3,2}$	0.55	0.12	$\tau_{13,2}$	-2.29	0.11	$\tau_{23,2}$	-2.01	0.19
$\tau_{3,3}$	0.30	0.15	$\tau_{13,3}$	0.03	0.11	$\tau_{23,3}$	-0.95	0.21
$\tau_{4,1}$	-2.38	0.09	$\tau_{14,1}$	-3.42	0.16	$\tau_{24,1}$	-2.97	0.15
$\tau_{4,2}$	-0.34	0.11	$\tau_{14,2}$	-0.67	0.14	$\tau_{24,2}$	-2.30	0.14
$\tau_{4,3}$	0.40	0.14	$\tau_{14,3}$	1.06	0.21	$\tau_{24,3}$	-0.94	0.15
$\tau_{5,1}$	-2.13	0.09	$\tau_{15,1}$	-2.96	0.11	$\tau_{25,1}$	-2.83	0.14
$\tau_{5,2}$	0.07	0.10	$\tau_{15,2}$	-1.61	0.07	$\tau_{25,2}$	-2.00	0.13
$\tau_{5,3}$	0.41	0.13	$\tau_{15,3}$	-0.04	0.08	$\tau_{25,3}$	-0.84	0.14
$\tau_{6,1}$	-2.01	0.07	$\tau_{16,1}$	-2.97	0.12	$\tau_{26,1}$	-2.21	0.25
$\tau_{6,2}$	0.15	0.09	$\tau_{16,2}$	-1.26	0.08	$\tau_{26,2}$	-2.13	0.25
$\tau_{6,3}$	0.40	0.13	$\tau_{16,3}$	0.32	0.12	$\tau_{26,3}$	-0.84	0.23
$\tau_{7,1}$	-1.71	0.06	$\tau_{17,1}$	-3.29	0.16	$\tau_{27,1}$	-2.82	0.16
$\tau_{7,2}$	-0.63	0.07	$\tau_{17,2}$	-2.59	0.13	$\tau_{27,2}$	-2.26	0.14
$\tau_{7,3}$	-0.09	0.07	$\tau_{17,3}$	-0.21	0.09	$\tau_{27,3}$	-0.95	0.14
$\tau_{8,1}$	-1.72	0.06	$\tau_{18,1}$	-3.09	0.16	$\tau_{28,1}$	-3.31	0.20
$\tau_{8,2}$	0.02	0.08	$\tau_{18,2}$	-2.47	0.12	$\tau_{28,2}$	-2.24	0.17
$\tau_{8,3}$	0.30	0.10	$\tau_{18,3}$	-0.15	0.09	$\tau_{28,3}$	-0.67	0.19
$\tau_{9,1}$	-1.96	0.10	$\tau_{19,1}$	-2.85	0.14	$\tau_{29,1}$	-2.93	0.19
$\tau_{9,2}$	-0.75	0.10	$\tau_{19,2}$	-1.32	0.09	$\tau_{29,2}$	-2.28	0.18
$\tau_{9,3}$	0.53	0.16	$\tau_{19,3}$	0.34	0.12	$\tau_{29,3}$	-0.91	0.17
$\tau_{10,1}$	-1.54	0.05	$\tau_{20,1}$	-2.43	0.13			
$\tau_{10,2}$	-0.35	0.07	$\tau_{20,2}$	-2.49	0.12			
$\tau_{10,3}$	-0.14	0.07	$\tau_{20,3}$	-0.19	0.09			
σ_{11}	1.65	0.10	σ_{22}	2.32	0.21	σ_{33}	0.35	0.03
σ_{12}	1.56	0.25	σ_{13}	0.30	0.03	σ_{23}	0.27	0.04
ρ_{12}	0.80	0.13	ρ_{13}	0.39	0.03	ρ_{23}	0.30	0.04

Note. The parameters are defined in Equation 3. PESW = Positive Evaluation for Self-Willingness Toward tattooing; PET = Positive Evaluation Toward Tattooing; NET = Negative Evaluation Toward Tattooing.

Figure 3. Item characteristic curves for “tattooing is brave” in the empirical example.

Table 2. Reliability (SE) and Test Length Increment for the Three Scales of the Tattoo Attitude Questionnaire Obtained From the Multidimensional and Unidimensional Approaches.

Dimension	PESW	PET	NET
Number of Items	10	10	9
Reliability (MD)	0.915 (0.020)	0.899 (0.024)	0.816 (0.029)
Reliability (UD)	0.865 (0.020)	0.832 (0.022)	0.812 (0.026)
Test Length Increment	68%	79%	3%

Note. PESW = Positive Evaluation for Self-Willingness toward tattooing; PET = Positive Evaluation toward Tattooing; NET = Negative Evaluation toward Tattooing; MD = multidimensional, UD = unidimensional.

to perform the CMGGUM is feasible. It depends on sample size, test length, and model complexity. For example, in the simulation where $N = 500$, 3 dimensions, 15 items per dimension, and between-item multidimensionality, each replication took approximately 3 hr to converge; when $N = 1,500$, each replication took approximately 10 hr to converge. The computer time for the real data was approximately 15 hr.

The authors focused on between-item multidimensionality in this study to be in line with the definition of unfolding process. As a mathematical equation, the CMGGUM (Equation 3) is identifiable even when $\mathbf{d}_i$ (the design vector for item i) contains multiple 1s; that is, mathematically speaking, item i can measure more than one dimension. However, this resulting within-item compensatory multidimensional model would lose its status as a traditional unfolding model because item responses would no longer be a function of the proximity between a person and item in the latent space. A potential application of such a model would be dimensional decomposition. In a unidimensional unfolding model, let θ_j be the target latent trait of person j , and δ_i be the location of item i . It is generally the distance between θ_j and δ_i that determines

the item responses. If there exists a strong rationale that θ_j can be decomposed into K components: $\theta_j = \sum_{k=1}^K a_k \theta_{kj}$, where a_k is the weight for component k and θ_{kj} is the k th component for person j , then, item i can be viewed as measuring the K components. It is feasible to specify α_i and $\mathbf{d}_i$ to accommodate such an application. More investigation into the issue of within-item multidimensionality is welcome.

Future studies can also be conducted to address the following issues. First, the CMGGUM was evaluated with a few conditions and replications. More conditions (e.g., test lengths, number of dimensions) and replications are recommended. Second, although WinBUGS took approximately 10 hr per replication in the simulation study, it may take a much longer time for a higher dimensionality, a longer test, or a larger sample size. To make the CMGGUM more feasible, a tailor-made computer program is preferred. Third, applications of the CMGGUM, including computerized adaptive testing, test linking, and differential item functioning, need further investigation. Fourth, the CMGGUM can become exploratory when the design vectors of all items consist of parameters. Its parameter recovery in high dimensions can be very challenging. Finally, recent research has extended item response models to account for multilevel structure, longitudinal design, many-faceted data, and others (Hung, 2010; Wang & Liu, 2007). It would be of great value to extend the CMGGUM to account for such testing situations. In this study, the unidimensional GGUM was generalized to a multidimensional one. It also would be of great value for future studies to generalize other unidimensional IRT unfolding models to multidimensional ones, such as those included in the work of Luo (2001).

Appendix

Winbugs Codes for the Tattoo Data

```

Model {
  for ( j in 1 : N) {
    theta[j,1:3] ~ dmnorm(mu[],tau[],)
#model the first dimension for unidimensional items 1-10
    for ( k in 1 : 10) {
      Q[ j , k,1] <- 1+exp(a[k]*(7*theta[j,1]-7*b[k]))
      Q[ j , k,2]<-exp(a[k]*(theta[j,1]-b[k]-step[k,1])) +
      exp(a[k]*(6*theta[j,1]-6*b[k]-step[k,1]))
      Q[ j , k,3]<-exp(a[k]*(2*theta[j,1]-2*b[k]-step[k,1]-step[k,2])) +
      exp(a[k]*(5*theta[j,1]-5*b[k]-step[k,1]-step[k,2]))
      Q[ j , k,4]<-exp(a[k]*(3*theta[j,1]-3*b[k]-step[k,1]-step[k,2]-
      step[k,3])) + exp(a[k]*(4*theta[j,1]-4*b[k]-step[k,1]-step[k,2]-
      step[k,3]))
      sum[j,k]<- Q[ j , k, 1] + Q[ j , k, 2] + Q[ j , k, 3] + Q[ j , k, 4]
      p[ j , k,1] <- Q[ j , k, 1]/sum[j,k]
      p[ j , k,2] <- Q[ j , k, 2]/sum[j,k]
      p[ j , k,3] <- Q[ j , k, 3]/sum[j,k]
      p[ j , k,4] <- Q[ j , k, 4]/sum[j,k]
      r[ j , k] ~ dcat(p[j,k,])
    }
#model the second dimension for unidimensional items 11-19
    for ( k in 11 : 19) {
      Q[ j , k,1] <- 1+exp(a[k]*(7*theta[j,2]-7*b[k]))

```

```

Q[j, k, 2] <- exp(a[k]*(theta[j, 2] - b[k] - step[k, 1])) +
exp(a[k]*(6*theta[j, 2] - 6*b[k] - step[k, 1]))
Q[j, k, 3] <- exp(a[k]*(2*theta[j, 2] - 2*b[k] - step[k, 1] - step[k, 2]))
+ exp(a[k]*(5*theta[j, 2] - 5*b[k] - step[k, 1] - step[k, 2]))
Q[j, k, 4] <- exp(a[k]*(3*theta[j, 2] - 3*b[k] - step[k, 1] - step[k, 2] -
step[k, 3])) + exp(a[k]*(4*theta[j, 2] - 4*b[k] - step[k, 1] - step[k, 2] -
step[k, 3]))
sum[j, k] <- Q[j, k, 1] + Q[j, k, 2] + Q[j, k, 3] + Q[j, k, 4]
p[j, k, 1] <- Q[j, k, 1]/sum[j, k]
p[j, k, 2] <- Q[j, k, 2]/sum[j, k]
p[j, k, 3] <- Q[j, k, 3]/sum[j, k]
p[j, k, 4] <- Q[j, k, 4]/sum[j, k]
r[j, k] ~ dcat(p[j, k,])
}

#model the third dimension for unidimensional items 20-29
for (k in 20 : T) {
Q[j, k, 1] <- 1+exp(a[k]*(7*theta[j, 3] - 7*b[k]))
Q[j, k, 2] <- exp(a[k]*(theta[j, 3] - b[k] - step[k, 1])) +
exp(a[k]*(6*theta[j, 3] - 6*b[k] - step[k, 1]))
Q[j, k, 3] <- exp(a[k]*(2*theta[j, 3] - 2*b[k] - step[k, 1] - step[k, 2]))
+ exp(a[k]*(5*theta[j, 3] - 5*b[k] - step[k, 1] - step[k, 2]))
Q[j, k, 4] <- exp(a[k]*(3*theta[j, 3] - 3*b[k] - step[k, 1] - step[k, 2] -
step[k, 3])) + exp(a[k]*(4*theta[j, 3] - 4*b[k] - step[k, 1] - step[k, 2] -
step[k, 3]))
sum[j, k] <- Q[j, k, 1] + Q[j, k, 2] + Q[j, k, 3] + Q[j, k, 4]
p[j, k, 1] <- Q[j, k, 1]/sum[j, k]
p[j, k, 2] <- Q[j, k, 2]/sum[j, k]
p[j, k, 3] <- Q[j, k, 3]/sum[j, k]
p[j, k, 4] <- Q[j, k, 4]/sum[j, k]
r[j, k] ~ dcat(p[j, k,])
}

#Posterior predictive checks
for (k in 1: T) {
E[j, k] <- 1*p[j, k, 1] + 2*p[j, k, 2] + 3*p[j, k, 3] + 4*p[j, k, 4]
Var[j, k] <- 1*p[j, k, 1] + 4*p[j, k, 2] + 9*p[j, k, 3] + 16*p[j, k, 4] - E[j, k]*
E[j, k]

resid[j, k] <- r[j, k] - E[j, k]
teststat[j, k] <- pow(resid[j, k], 2)/Var[j, k]

r.rep[j, k] ~ dcat(p[j, k,])
resid.rep[j, k] <- r.rep[j, k] - E[j, k]
teststat.rep[j, k] <- pow(resid.rep[j, k], 2)/Var[j, k]
}

teststatsum <- sum(teststat[,])
teststatsum.rep <- sum(teststat.rep[,])
pvalue <- step(teststatsum.rep - teststatsum)

```



```

# Specify the priors
for ( k in 1:T) {
  step[k,1]~dnorm(-2,0.25]
  step[k,2]~dnorm(-1.5,0.25]
  step[k,3]~dnorm(-1,0.25]  }
for ( k in 2:10) {
  a[k] ~ dlnorm(0,4)
  b[k] ~ dunif(-3,3) }
for ( k in 12:19) {
  a[k] ~ dlnorm(0,4)
  b[k] ~ dunif(-3,3)  }
for ( k in 21:T) {
  a[k] ~ dlnorm(0,4)
  b[k] ~ dunif(-3,3)  }

a[1] <- 1.0
a[11] <- 0.40
a[20] <- 0.99
b[1]~dunif(-3,-0.5]
b[11] ~dunif(-3,-0.5]
b[20] ~dunif(0.5,3)

mu[1] <- 0
mu[2] <- 0
mu[3] <- 0

tau[1:3,1:3]~dwish(R[ , ],3)
R[1,1]<- 0.001
R[1,2]<- 0
R[2,1]<- 0
R[2,2]<- 0.001
R[1,3]<- 0
R[2,3]<- 0
R[3,1]<- 0
R[3,2]<- 0
R[3,3]<- 0.001

sigma2[1:3,1:3]<-inverse(tau[,])
rho12<-sigma2[1,2]/sqrt(sigma2[1,1]*sigma2[2,2])
rho13<-sigma2[1,3]/sqrt(sigma2[1,1]*sigma2[3,3])
rho23<-sigma2[2,3]/sqrt(sigma2[3,3]*sigma2[2,2])
}

```

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The first author was sponsored by General Research Fund (842709), University Grants Committee, Hong Kong.

References

- Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.
- Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement, 17*, 253-276.
- Carter, N. T., Dalal, D. K., Boyce, A. S., O'Connell, M. S., Kung, M.-C., & Delgado, K. (2014). Uncovering curvilinear relationships between conscientiousness and job performance: How theoretically appropriate measurement makes an empirical difference. *Journal of Applied Psychology, 99*, 564-586.
- Carter, N. T., & Zickar, M. J. (2011). The influence of dimensionality on parameter estimation accuracy in the generalized graded unfolding model. *Educational and Psychological Measurement, 71*, 765-788.
- Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19*, 88-106.
- DeSarbo, W. S., & Hoffman, D. L. (1987). Constructing MDS joint spaces from binary choice data: A multidimensional unfolding threshold model for marketing research. *Journal of Marketing Research, 24*, 40-54.
- Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B, 56*, 501-514.
- Gelman, A., Meng, X.-L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733-807.
- Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. Berger, A. P. Dawid, & J. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 169-193). Oxford, UK: Oxford University Press.
- Hung, L.-F. (2010). The multigroup multilevel categorical latent growth curve models. *Multivariate Behavioral Research, 45*, 359-392.
- Lin, R.-C., & Huang, B.-F. (2005). Tattooing and juvenile delinquency in Taiwan. *Indigenous Psychological Research in Chinese Societies, 23*, 321-372.
- Luo, G. (2001). A class of probabilistic unfolding models for polytomous responses. *Journal of Mathematical Psychology, 45*, 224-248.
- Maydeu-Olivares, A., Hernandez, A., & McDonald, R. P. (2006). A multidimensional ideal point item response theory model for binary data. *Multivariate Behavioral Research, 41*, 445-471.
- Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics, 22*, 1142-1160.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.
- Raftery, A. E., Newton, M. A., Satagopan, J., & Krivitsky, P. (2007). Estimating the integrated likelihood via posterior simulation using harmonic mean identity. In J. Bernardo, J. Bayarri, & J. Berger (Eds.), *Bayesian statistics 8* (pp. 1-45). London, England: Oxford University Press.
- Reckase, M. (2009). *Multidimensional item response theory*. New York, NY: Springer.
- Roberts, J. S. (2001). Computer program exchange: GGUM2000: Estimation of parameters in the generalized graded unfolding model. *Applied Psychological Measurement, 25*, 38.
- Roberts, J. S. (2008). Modified likelihood-based item fit statistics for the generalized graded unfolding model. *Applied Psychological Measurement, 32*, 407-423.
- Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3-32.

- Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2002). Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model. *Applied Psychological Measurement, 26*, 192-207.
- Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20*, 231-255.
- Roberts, J. S., & Shim, H. S. (2010, July). *Multidimensional unfolding with item response theory: The multidimensional generalized graded unfolding model*. Paper presented at the IMPS 2010 Annual Meeting, Athens, GA.
- Roberts, J. S., & Thompson, V. M. (2011). Marginal maximum a posteriori item parameter estimation for the generalized graded unfolding model. *Applied Psychological Measurement, 35*, 259-279.
- Spector, P. E. (1997). *Job satisfaction: Application, assessment, causes, and consequences*. Thousand Oaks, CA: Sage.
- Spiegelhalter, D., Best, N., Carlin, B., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, 64*, 583-639.
- Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS (version 1.4) [Computer program]. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.
- Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*, 25-39.
- Takane, Y. (1996). An item response model for multidimensional analysis of multiple-choice data. *Behaviormetrika, 23*, 153-167.
- 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.
- Wang, W.-C., Liu, C.-W., & Wu, S.-L. (2013). The random-threshold generalized unfolding model and its application of computerized adaptive testing. *Applied Psychological Measurement, 37*, 179-200.
- Wang, W.-C., & Liu, C.-Y. (2007). Formulation and application of the generalized multilevel facets model. *Educational and Psychological Measurement, 67*, 583-605.