A Dominance Variant Under the Multi-Unidimensional Pairwise-Preference Framework: Model Formulation and Markov Chain Monte Carlo Estimation

Applied Psychological Measurement
2016, Vol. 40(7) 500–516
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621616662226
apm.sagepub.com

Daniel Morillo¹, Iwin Leenen², Francisco J. Abad¹,
Pedro Hontangas³, Jimmy de la Torre⁴, and
Vicente Ponsoda¹

Abstract

Forced-choice questionnaires have been proposed as a way to control some response biases associated with traditional questionnaire formats (e.g., Likert-type scales). Whereas classical scoring methods have issues of ipsativity, item response theory (IRT) methods have been claimed to accurately account for the latent trait structure of these instruments. In this article, the authors propose the multi-unidimensional pairwise preference two-parameter logistic (MUPP-2PL) model, a variant within Stark, Chernyshenko, and Drasgow's MUPP framework for items that are assumed to fit a dominance model. They also introduce a Markov Chain Monte Carlo (MCMC) procedure for estimating the model's parameters. The authors present the results of a simulation study, which shows appropriate goodness of recovery in all studied conditions. A comparison of the newly proposed model with a Brown and Maydeu's Thurstonian IRT model led us to the conclusion that both models are theoretically very similar and that the Bayesian estimation procedure of the MUPP-2PL may provide a slightly better recovery of the latent space correlations and a more reliable assessment of the latent trait estimation errors. An application of the model to a real data set shows convergence between the two estimation procedures. However, there is also evidence that the MCMC may be advantageous regarding the item parameters and the latent trait correlations.

Keywords

forced-choice questionnaires, ipsative scores, Bayesian estimation, MCMC, multidimensional IRT

In the context of noncognitive trait measurement, several authors have proposed the use of forced-choice questionnaires (FCQs) as an alternative to traditional Likert-type scale response

¹Faculty of Psychology, Universidad Autónoma de Madrid, Spain

²Instituto Nacional para la Evaluación de la Educación, Mexico City, Mexico

³Universidad de Valencia, Spain

⁴University of Hong Kong

Corresponding Author:

Daniel Morillo, Facultad de Psicología, Universidad Autónoma de Madrid, C/Iván Pavlov, 6, 28049 Madrid, Spain.

Email: daniel.morillo@uam.es

formats (e.g., Christiansen, Burns, & Montgomery, 2005; Saville & Willson, 1991) as the latter are particularly sensitive to response styles such as conscious distortion (Baron, 1996). FCQs consist of blocks of two or more items, each one typically measuring a single, a priori specified, underlying trait or dimension. The respondent's task is to (partially) rank order the items in each block, according to how well they describe him or her, for example, by selecting the items that describe him or her best and/or worst.

FCQs have been criticized because traditional scores suffer from ipsativity. An individual's ipsative scores (Cattell, 1944) are dependent on each other and are useless for interindividual comparisons (Cornwell & Dunlap, 1994). Ipsativity also leads to problems with assessing reliability and validity (Closs, 1996; Hicks, 1970).

Recently, some authors have proposed scoring procedures within the framework of item response theory (IRT) which yield nonipsative, normative scores from FCQ data. The multi-unidimensional pairwise preference (MUPP; Stark, Chernyshenko, & Drasgow, 2005) model and the Thurstonian IRT (TIRT; Brown & Maydeu-Olivares, 2011) model are the most well-known examples. The MUPP is a model for forced-choice blocks consisting of two items and assumes for each item a latent response process that follows the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000). The TIRT models the probability of selecting an item along the lines of the Thurstone's (1927) law of comparative judgment and has the advantage of allowing blocks of two or more items. Although the TIRT is essentially a factor model, it can be expressed in IRT terms as well (like other item factor models, see Takane & de Leeuw, 1987).

Arguably, the essential difference between both models relates to the underlying process for item evaluation. Stark et al.'s (2005) MUPP, by relying on the GGUM, assumes an *unfolding* process (i.e., the probability of endorsing an item is a single-peaked function of the latent trait), whereas the TIRT assumes a *dominance* process (i.e., with a monotonic item response function). Whether unfolding or dominance models are more suited for the analysis of noncognitive items is an ongoing controversy in the literature (for a detailed discussion, see Drasgow, Chernyshenko, and Stark's [2010] focal article with commentaries in the December 2010 issue of *Industrial and Organizational Psychology*). Here, the authors mention some theoretical considerations as well as recent evidence in favor of dominance models. First, certain constructs (e.g., pathological aspects of personality) seem to better conform to a dominance model (Carvalho, de Oliveira, Pessotto, & Vincenzi, 2015; Cho, Drasgow, & Cao, 2015). Second, dominance models are usually more parsimonious than unfolding models. Third, scales formed by dominance items tend to have better psychometric properties, such as higher reliability and correlations with external criteria (Huang & Mead, 2014). Finally, items, rather than traits, are characterized by being dominance or unfolding, as a trait may actually be measured by both types of items. In fact, some authors argue that unfolding models only yield a better fit for items in the middle of the trait continuum, but these items are difficult to write, are not invariant to reverse scoring, and may be equally well fit by a higher dimensional dominance model (Brown & Maydeu-Olivares, 2010; Oswald & Schell, 2010). Given these arguments, one may consider replacing the GGUM in the original MUPP by a dominance model, as "there is nothing in the actual MUPP model that stops it from being populated with dominance items and, consequently, using a dominance model" (Brown & Maydeu-Olivares, 2010, p. 491).

The authors of both the TIRT and the MUPP have also presented their respective estimation procedures. The TIRT estimation, based on confirmatory factor analysis, estimates the item parameters and latent variance-covariance structure using a marginal bivariate-information method (Brown & Maydeu-Olivares, 2011, 2012). This procedure comes with some minor drawbacks: First, it disregards the correlation among component unities (in blocks that contain more than two items), which Brown and Maydeu-Olivares (2011) claimed to have a negligible

effect. Second, it ignores the estimation error associated with the structural parameters when the respondents' latent trait values are estimated. This is a common drawback of multistep serial procedures that use estimates from a previous step as fixed values in a subsequent step. Third, to ensure quality estimation results, the TIRT requires that some blocks combine items of opposite polarity (Brown & Maydeu-Olivares, 2011), that is, *direct* items (e.g., "Complete tasks successfully"; International Personality Item Pool, n.d.) and *inverse* items (e.g., "Yell at people"). However, opposite-polarity blocks are less robust against response biases, as the respondent would be prone to select the more desirable item, which are often considered the very reason to employ FCQs.

The MUPP estimation procedure only estimates the person parameters, assuming known values of the item parameters. The latter are typically obtained from a prior administration and calibration of the items in a graded-scale format (Stark et al., 2005). Apart from being less efficient, such a strategy disregards the uncertainty in the item parameters as well and relies on the assumption that the item parameters are equivalent across response formats. Stark et al. further suggest the inclusion of unidimensional blocks (of which both items address the same dimension) for metric identification. However, these blocks require items with distant locations on the latent scale. This property may make them prone to response biases.

The remainder of this article is organized as follows: First, the authors present the MUPP two-parameter logistic (MUPP-2PL) model, a MUPP variant for dominance items, and discuss its relation with other multidimensional IRT models. Second, they cast the model in a Bayesian framework and propose an estimation algorithm for joint estimation of structural and person parameters. Third, they evaluate the algorithm in a simulation study, with special attention to the above-mentioned limitations of the original MUPP and TIRT estimation procedures. Fourth, they present an empirical study to illustrate the practical use of the model. Finally, they conclude with a discussion. Throughout, whenever appropriate, the MUPP-2PL is compared with the TIRT model.

The MUPP-2PL Model

In the MUPP framework, the probability of person j choosing an item i_1 over item i_2 in block i is given by (Stark et al., 2005)

$$P(Y_{ij} = 1) = \frac{P(X_{i1j} = 1)P(X_{i2j} = 0)}{P(X_{i1j} = 1)P(X_{i2j} = 0) + P(X_{i1j} = 0)P(X_{i2j} = 1)}, \quad (1)$$

where Y_{ij} is a variable that denotes the selected item on the block (with a value of 1 if i_1 is the selected response, and 2 if it is i_2), and X_{i1j} and X_{i2j} are the latent responses on items i_1 and i_2 , respectively, being equal to 1 if respondent j endorses the item, and 0 otherwise.

In the original MUPP model, the probability functions at the right side of Equation 1 are item response functions described by the GGUM. To obtain the MUPP-2PL variant, the authors replace the GGUM by the 2PL (Birnbaum, 1968) model. The block characteristic function (BCF) can then be written as follows:

$$P_i(Y_{ij} = 1 | \theta_j) = \phi_L \left(a_{i1} \theta_{\tilde{i}1j} - a_{i2} \theta_{\tilde{i}2j} + d_i \right) = \frac{1}{1 + \exp \left[- \left(a_{i1} \theta_{\tilde{i}1j} - a_{i2} \theta_{\tilde{i}2j} + d_i \right) \right]}, \quad (2)$$

where ϕ_L is the logistic function; θ_j is a vector with a person's positions on each of the D latent traits addressed by the FCQ; $\theta_{\tilde{i}1j}$ and $\theta_{\tilde{i}2j}$ are the coordinates of θ_j in the dimensions addressed by items i_1 and i_2 , respectively (which are the same if the block is unidimensional); a_{i1} and a_{i2} are the scale (discrimination) parameters of items i_1 and i_2 , respectively; and d_i is the block

Figure 1. MUPP-2PL model BCFs of three blocks.

Note. Their parameters, expressed as $\{a_{i1}, a_{i2}, d_i\}$, are Block A = $\{1, 1, 0\}$, Block B = $\{1, 1, -2\}$, and Block C = $\{2, 1, 0\}$. MUPP-2PL = multi-unidimensional pairwise preference two-parameter logistic; BCF = block characteristic function.

intercept parameter, which combines the two item location parameters b_{i1} and b_{i2} involved in the 2PL; in particular, $d_i = a_{i2}b_{i2} - a_{i1}b_{i1}$. (Note that the two location parameters cannot be uniquely identified; the implications of this underdetermination will be considered further in the discussion.) For all parameters in Equation 2, the range of allowable values comprises the full set of real numbers. In this respect, note that the sign of the scale parameter defines the item's polarity; direct and inverse items have positive and negative polarity, respectively.

Figure 1 graphs the MUPP-2PL BCF for three bidimensional blocks with different item parameters. It illustrates how a change in the intercept translates the surface slope in the space. A change in the scale parameter rotates the slope (in addition to producing a net change in the gradient), making the block more discriminating in the corresponding dimension. The information matrix of a questionnaire made up of bidimensional blocks is presented in Online Appendix A.

Relationships of the MUPP-2PL to Other Models

Relationship with the multidimensional compensatory logistic model (MCLM). The MUPP-2PL model is algebraically equivalent to the MCLM (Reckase & McKinley, 1982), which is usually expressed as follows:

$$P_i(Y_{ij} = 1 | \theta_j) = \Phi_L(\mathbf{a}_i \theta_j + d_i), \quad (3)$$

where $\mathbf{a}_i$ is a D -dimensional vector with the scale parameters of the i th block, and d_i is the i th block intercept parameter. Comparing Equations 2 and 3 reveals the following differences with respect to the implied constraints: (a) In the MUPP-2PL, each block addresses either one or two a priori specified dimensions, which in terms of Equation 3 comes down to restricting all but one or two scale parameters to 0 and (b) the MCLM scale parameters are restricted to be positive, whereas in the MUPP-2PL, they can be negative (note that the sign of the scale parameter associated with the second item in the block is inverted in Equation 2).

Relationship with the TIRT model. Consider the IRT formulation of the TIRT (Brown & Maydeu-Olivares, 2011),

$$P(Y_{ij} = 1 | \eta_{i1j}, \eta_{i2j}) = \Phi_N(\alpha_i + \beta_{i1} \eta_{i1j} - \beta_{i2} \eta_{i2j}), \quad (4)$$

where Φ_N is the cumulative normal distribution function; η_{i_1j} and η_{i_2j} are the coordinates of a D -dimensional latent trait vector $\boldsymbol{\eta}$ in the dimensions addressed by items i_1 and i_2 , respectively; β_{i_1} and β_{i_2} are the slope parameters of items i_1 and i_2 , respectively; and α_i is the block intercept parameter.

It should be noted that, although the TIRT model is generally defined for blocks of two or more items, Equation 4 refers to the response probability of a *binary outcome* (i.e., the result of a latent comparison between two items within a block that possibly includes more than two items). By considering pairwise comparisons only, Equation 4 directly models the response probability on a block and turns out to be equivalent to Equation 2, except for the probit versus logit link functions (which are known to be very closely related; Haley, 1952).

Bayesian Estimation of the MUPP-2PL

Given the responses of N persons on a questionnaire of n item blocks collectively measuring D underlying dimensions, and assuming independence among subjects and local independence across responses within subjects, the likelihood function for the MUPP-2PL is given by

$$L(\mathbf{Y}|\boldsymbol{\theta}, \mathbf{a}, \mathbf{d}) = \prod_{j=1}^N \prod_{i=1}^n \left[P_i^{2-y_{ij}}(\boldsymbol{\theta}_j) Q_i^{y_{ij}-1}(\boldsymbol{\theta}_j) \right], \quad (5)$$

where $\mathbf{Y}$ is an $N \times n$ matrix of responses, $\boldsymbol{\theta}$ is an $N \times D$ array of person latent trait parameters, $\mathbf{a}$ is an $n \times 2$ array of item scale parameters, and $\mathbf{d}$ is an $n \times 1$ array of item intercept parameters.

To estimate the person and item parameters simultaneously, the authors formulate the model in a Bayesian framework. The prior distributions are specified as follows:

- a. $\boldsymbol{\theta}_j \stackrel{iid}{\sim} \text{MVN}(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$, $j = 1, \dots, N$, with $\boldsymbol{\mu}_\theta$ being a D -dimensional mean vector and $\boldsymbol{\Sigma}_\theta$ a $D \times D$ covariance matrix. For identification purposes (see next subsection), $\boldsymbol{\mu}_\theta$ will be restricted to 0 and $\boldsymbol{\Sigma}_\theta$ to a correlation matrix. The hyperprior distribution of $\boldsymbol{\Sigma}_\theta$ will be assumed uniform; that is, all positive-definite matrices with diagonal elements of 1 are considered equally likely a priori.
- b. $|a_{i_k}| \stackrel{iid}{\sim} \text{lognorm}(\mu_a, \sigma_a^2)$, $i = 1, \dots, n$; $k = 1, 2$; with μ_a and σ_a being prespecified constants, for which the authors suggest values of 0.25 and 0.5, respectively. The sign of the a_{i_k} is fixed a priori; in practical applications, it is typically derived from a content analysis of the item to reflect its polarity.
- c. $d_i \stackrel{iid}{\sim} N(\mu_d, \sigma_d^2)$, $i = 1, \dots, n$, with μ_d and σ_d being prespecified constants. The authors suggest values of 0 and 1 for these constants, respectively.

By Bayes's theorem, the posterior density of the parameters is proportional to

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{\Sigma}_\theta, \mathbf{a}, \mathbf{d}|\mathbf{Y}) &\propto L(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\Sigma}_\theta, \mathbf{a}, \mathbf{d}) f(\boldsymbol{\theta}, \boldsymbol{\Sigma}_\theta, \mathbf{a}, \mathbf{d}) \\ &= \prod_{j=1}^N \prod_{i=1}^n \left[P_i^{2-y_{ij}}(\boldsymbol{\theta}_j) Q_i^{y_{ij}-1}(\boldsymbol{\theta}_j) \right] \prod_{j=1}^N N(\boldsymbol{\theta}_j|\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta) \\ &\quad \prod_{i=1}^n \left[\text{lognorm}(a_{i_1}|\mu_a, \sigma_a^2) \text{lognorm}(a_{i_2}|\mu_a, \sigma_a^2) N(d_i|\mu_d, \sigma_d^2) \right]. \end{aligned}$$

The Markov Chain Monte Carlo (MCMC) algorithm, which the authors developed to sample from this posterior distribution, is a Metropolis–Hastings (or Metropolis-within-Gibbs) algorithm. For an introduction to the MCMC methodology in the context of IRT model estimation, see Patz and Junker (1999a, 1999b). The proposed algorithm runs multiple chains, where each chain starts with a distinct set of initial values for the parameters; in each iteration, all parameters are successively updated by drawing them one by one from their conditional distribution given the most recent values for the other parameters. The chains run until they have converged, according to Gelman and Rubin's (1992) statistic. A detailed description of the algorithm is provided in Online Appendix B.

Identification of Latent Trait and Item Parameters

Identifiability of the MUPP-2PL model is directly related to the MCLM.¹ The origin and unit of each dimension must be fixed to identify the metric (De Ayala, 2009), which explains the restrictions applied to $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ in the previous section. Rotational indeterminacy may be another source of unidentifiability, but only if $D = 2$ and each block is bidimensional. In other cases, the structural zeros in the rows (i.e., blocks) of the scale parameter matrix imply a *triangular* configuration (Thurstone, 1947), which solves this indeterminacy. For $D = 2$, the inclusion of unidimensional blocks in the FCQ would resolve the rotational indeterminacy. However, if block i is unidimensional, then Equation 2 reduces to the 2PL model equation with a scale parameter equal to $a_{i_2} - a_{i_1}$. Thus, the scale parameters cannot be uniquely identified for unidimensional blocks.

As an aside, note that the TIRT model also suffers from rotational indeterminacy when applied to pairwise blocks measuring two dimensions. To solve this problem, Brown and Maydeu-Olivares (2011) suggested “fix[ing] the two factor loadings of the first pair” (p. 473). However, this may have a drawback, in the sense that the final solution may strongly depend on the values assigned to those loadings.

Simulation Study

Design and Data Generation Process

The authors systematically manipulated the same three factors as in Brown and Maydeu-Olivares (2011), albeit at slightly different levels: (a) number of blocks that make up the questionnaire (questionnaire length [QL]), 18 or 36; (b) the proportion of these blocks that combine items of opposite polarity (opposite-polarity block proportion [OPBP]), 2/3, 1/3, or 0; and (c) the correlation between the latent traits (interdimensional correlation [IC]), .00, .25, or .50.

The three factors were completely crossed, yielding 18 different conditions; for each condition, they simulated 100 data sets. Data sets were independently generated by the following four-step procedure. First, for each of 1,000 simulees, a three-dimensional latent trait vector was independently drawn from a trivariate normal distribution with mean vector $\mathbf{0}$ and a covariance matrix $\boldsymbol{\Sigma}_0$, with all variances equal to 1 and covariances ρ equal to the level of IC for the condition. Second, the (18 or 36) blocks were equally divided into three groups; in each group, the items measured a pair of dimensions (Dimensions 1 and 2, 1 and 3, or 2 and 3). For each item, a scale parameter was independently drawn from a lognormal distribution, with both the log-mean and the log-SD parameters equal to .25. In each of the three groups, a proportion of the blocks was selected according to the level of OPBP, and one of their scale parameters was multiplied by -1 to obtain the inverse items. The number of inverse items for each dimension was kept constant across the three groups. Third, for each block, an intercept parameter was

independently drawn from a normal distribution, with mean and variance equal to 0 and 0.25, respectively. Fourth, a data matrix $\mathbf{Y}$ was generated by calculating for each cell the probability in Equation 2 based on the parameters drawn in the previous steps and converting this probability to a realized value of 1 or 2 by comparing it to a uniform random variate.

MCMC Analysis

Each data set was analyzed applying the MCMC algorithm introduced in the previous section. The authors specified four independent chains and ran 150,000 iterations for each data set. The first 50,000 draws were considered burn-in, and only every 25th draw was saved to the output file. Hence, the analysis of each data set yielded a total of $4 \text{ (chains)} \times 100,000 / 25 \text{ (saved draws/chain)} = 16,000$ draws.

The chains were initialized with the procedure explained in Online Appendix B (the random noise covariance matrix for initializing the latent trait parameters had all diagonal elements equal to .5, and all off-diagonal elements equal to .375). They were considered to have converged if and only if for all parameters the value on Gelman and Rubin's (1992) $\hat{R}$ statistic (calculated across the 16,000 draws) was below 1.2. Thirteen out of the 1,800 data sets did not satisfy the convergence criterion and were reanalyzed using different starting values. For each parameter, the Expected a Posteriori (EAP) estimate along with the 95% credibility interval (CrI; defined by the posterior sample .025 and .975 quantiles) and the standard error was computed from the 16,000 posterior draws.

Goodness-of-Recovery (GOR) Summary Statistics

For each data set, the authors analyzed the four types of parameters separately: the off-diagonal elements in Σ_θ (three correlation parameters in total), latent traits in θ (3,000 parameters), the scales in $\mathbf{a}$ ($2 \times \text{QL}$ parameters), and the intercepts in $\mathbf{d}$ (QL parameters). Let ξ_l and $\hat{\xi}_l$ be a generic notation of the true and the EAP estimate of a parameter, respectively, and L the number of parameters of the type under consideration. The following GOR summary statistics were calculated for each parameter type, in each replication:

1. Mean error, defined as

$$\text{ME}_{\hat{\xi}} = \frac{\sum_{l=1}^L (\hat{\xi}_l - \xi_l)}{L}, \quad (7)$$

2. Root mean squared error, given by

$$\text{RMSE}_{\hat{\xi}} = \sqrt{\frac{\sum_{l=1}^L (\hat{\xi}_l - \xi_l)^2}{L}}, \quad (8)$$

3. Proportion coverage by the 95% CrI, that is, the proportion of parameters ξ_l , across all L parameters, that are contained in the CrI derived for the parameter.

In addition, a mean reliability was computed as

$$\overline{\rho_{\theta}^2} = \frac{\sum_{d=1}^3 r_{\theta_d \theta_d}^2}{3} \quad (9)$$

from the latent trait estimates, and the correlations r_{aa} and r_{dd} , between the true values and the estimates of **a** and **d**, respectively.

For each of the GOR statistics, they calculated the means across the 100 data sets in each of the 18 conditions and examined the contributions of the main and interaction effects of the three factors manipulated in the study by ANOVA. They will focus on effects that are of moderate size at least ($\eta_p^2 > .06$; Cohen, 1988).

Results

The mean results for the GOR statistics at each level of the three factors are presented in Table 1. Three general results, which hold across the four parameter types, stand out. First, the mean errors are very close to 0 in all conditions. This result indicates there is no systematic distortion of the estimates for any type of parameter in a particular direction. Figure 2 plots the estimates against the true values for each parameter type in one particular condition. It illustrates that no systematic bias appears in the estimation, except for a slight relative bias toward the mean in the extreme values. This effect, typical of Bayesian analyses, is attributable to the prior distribution. Second, in all conditions, the proportion of true parameters contained in the corresponding CrI is very close to the nominal level of 95%. This suggests that the estimation method correctly accounts for the uncertainty in the parameter estimates. Third, the factor IC does not explain differences in GOR in any relevant way ($\eta_p^2 < .06$ for all GOR statistics, not only for the main effect of the IC factor but also for all interactions that involve this factor). The latter result is somewhat unexpected and contrary to Brown and Maydeu-Olivares's (2011), who found an inverse relationship between correlation and reliability. This difference might be due either to the estimation procedure or to the selected levels for each of the factors used in the study.

The authors now summarize the most important results for QL and OPBP on the precision of the estimates, as quantified by the RMSE and the correlation between true and estimated values. They differentiate among the following four parameter types.

Correlation parameters (Σ_{θ}). A moderate main effect on $\text{RMSE}_{\hat{\rho}}$ was found for both QL and OPBP: Longer questionnaires yielded more precise estimates of the latent trait correlations. Including blocks of items with opposite polarity (be it 2/3 or 1/3 of the items) caused the latent correlations to be estimated with smaller errors.

Latent trait parameters (θ). Large main effects of QL and OPBP were found for $\text{RMSE}_{\hat{\theta}}$ and $\overline{\rho_{\theta}^2}$: Longer tests and a higher proportion of opposite-polarity blocks resulted in more precise estimates. Moreover, a moderate interaction (with $\eta_p^2 = .12$) between QL and OPBP was found on $\overline{\rho_{\theta}^2}$ (see Figure 3). Note that in the worst condition (i.e., 18 blocks with direct items only), the reliability of the estimates was .63, somewhat below what is typically required in practical applications. However, for questionnaires of 36 direct-item blocks, the reliability was adequate with a value of .77.

Scale parameters (a**).** The precision of the estimates of the item scale parameters improved with the length of the questionnaire, although the improvement was relatively small. The presence of

Table 1. Mean Goodness-of-Recovery for each level of Questionnaire Length, Opposite-Polarity Block Proportion, and Interdimensional Correlation for the MCMC estimates.

	Questionnaire length			Opposite-polarity block proportion				Interdimensional correlation			
	18	36	η_p^2	2/3	1/3	0	η_p^2	.00	.25	.50	η_p^2
Correlation matrix (Σ_{θ})											
Mean error	0.001	0.000	0.000	−0.001	−0.001	0.003	0.002	−0.001	0.000	0.003	0.002
RMSE	0.052	0.038	0.090	0.040	0.040	0.055	0.096	0.049	0.047	0.041	0.025
95% CrI coverage	0.958	0.959	0.000	0.953	0.958	0.963	0.003	0.964	0.948	0.962	0.001
Latent traits (θ)											
Mean error	0.000	0.000	0.000	−0.001	−0.002	0.003	0.011	0.000	0.001	0.000	0.002
RMSE	0.529	0.414	0.872	0.433	0.438	0.544	0.842	0.469	0.474	0.471	0.001
95% CrI coverage	0.949	0.949	0.000	0.949	0.949	0.949	0.000	0.949	0.949	0.949	0.000
Mean reliability (ρ_{θ}^2)	0.717	0.827	0.835	0.810	0.806	0.700	0.812	0.775	0.770	0.772	0.008
Item scales (a)											
Mean error	0.010	0.005	0.004	0.005	0.004	0.015	0.012	0.004	0.006	0.014	0.010
RMSE	0.198	0.164	0.142	0.180	0.177	0.187	0.009	0.176	0.180	0.188	0.013
95% CrI coverage	0.950	0.950	0.000	0.950	0.950	0.950	0.000	0.952	0.949	0.950	0.002
True-estimate correlation	0.983	0.989	0.147	0.994	0.992	0.972	0.632	0.986	0.986	0.985	0.003
Block intercepts (d)											
Mean error	0.000	−0.001	0.000	0.000	0.000	−0.003	0.002	−0.002	0.000	0.000	0.001
RMSE	0.110	0.109	0.001	0.111	0.110	0.107	0.005	0.112	0.110	0.106	0.010
95% CrI coverage	0.952	0.952	0.000	0.956	0.951	0.951	0.002	0.953	0.953	0.951	0.000
True-estimate correlation	0.975	0.976	0.003	0.974	0.974	0.978	0.023	0.975	0.975	0.977	0.004

Note. The η_p^2 values are the partial eta-squared effect sizes associated with the main effect of the factor for the corresponding goodness-of-recovery statistic. The values in the other cells are the estimated marginal means of the goodness-of-recovery statistic across all replications for the corresponding factor level. MCMC = Markov Chain Monte Carlo; RMSE = root mean square error; CrI = credibility interval.

Figure 2. Plot of the estimates against the true values, across all replications, for the condition QL = 36, OPBP = 0, IC = .50.

Note. QL = questionnaire length; OPBP = opposite-polarity block proportion; IC = interdimensional correlation.

blocks that combine items of opposite polarity had an even smaller effect on the precision of the scale parameter estimates.

Intercept parameters (d). The GOR of the intercept parameters was highly accurate, and the factors considered in this study barely had any effect. Arguably, sample size (which was kept constant at 1,000 individuals in the present study) is a more important factor affecting the quality of estimation of the item parameters (both intercepts and scales).

Comparison With the TIRT Estimation

The TIRT estimation procedure was applied to the same simulated data, and the same GOR indices were computed (see Table 2; in the case of the structural parameters, the mean coverage of the 95% confidence interval [CI], rather than of the CrI, was computed). A comparison of the results from both procedures (through repeated-measures ANOVA) showed very little difference. The authors highlight here the two most relevant differences.

Figure 3. Interaction effect between OPBP and QL on the mean reliability ($\bar{\rho}_{\theta}^2$).
Note. OPBP = opposite-polarity block proportion; QL = questionnaire length.

First, the coverage by the 95% CrIs of latent trait parameters by the TIRT procedure was significantly less accurate than by the MCMC 95% CrIs (93.7% vs. 95.0%; $\eta_p^2 = .56$). Arguably, this probably relates to the joint estimation of item and person parameters by the MCMC procedure. Moreover, the coverage by the 95% TIRT CrIs was lower for QL = 18 (93.3%) than for QL = 36 (94.0%), whereas the MCMC CrIs maintained the coverage at the nominal level of 95.0% independently of test length.

Second, the latent trait correlations were more accurately estimated in the MCMC ($RMSE_{\hat{\rho}} = .045$) than in the TIRT ($RMSE_{\hat{\rho}} = .055$; $\eta_p^2 = .07$). However, this difference was exclusively found in the conditions with direct items only ($RMSE_{\hat{\rho}} = .055$ for MCMC vs. $RMSE_{\hat{\rho}} = .081$ for TIRT). In the opposite-polarity conditions, both procedures performed at the same level ($\eta_p^2 = .08$ for the interaction between OPBP and the estimation procedure).

Empirical Study

In this section, the authors briefly illustrate the application of the MUPP-2PL model to empirical data from a personality test. In particular, they applied an FCQ measuring the Big Five traits to a sample of 567 students from two Spanish universities. Sixteen cases were removed because of unresponded blocks, leaving 551 cases to analyze. The questionnaire, specifically assembled for this application, consisted of 30 blocks; its exact design is given in the first two columns of Table 3. They analyzed the responses with both the TIRT procedure and the MCMC algorithm. The latter was configured as in the simulation study. Convergence was found for both procedures.

The TIRT estimates obtained with Mplus showed an acceptable fit (root mean square error of approximation [RMSEA] = .035, $p(RMSEA < .05) = 1.000$; comparative fit index [CFI] = .906; Tucker-Lewis index [TLI] = .888). The MCMC and TIRT structural parameter estimates (see Table 3) strongly correlated (.88 and .89 for the scale parameters of the first and second

Table 2. Mean Goodness-of-Recovery for each level of Questionnaire Length, Opposite-Polarity Block Proportion, and Interdimensional Correlation for the TIRT estimates.

	Questionnaire length			Opposite-polarity block proportion				Interdimensional correlation			
	18	36	η_p^2	2/3	1/3	0	η_p^2	.00	.25	.50	η_p^2
Correlation matrix ($\hat{\Sigma}_\theta$)											
Mean error	-0.011	-0.007	0.001	-0.013	-0.010	-0.004	0.042	-0.001	-0.001	-0.025	0.006
RMSE	0.067	0.044	0.071	0.043	0.042	0.081	0.157	0.059	0.057	0.050	0.009
95% CI coverage	0.942	0.947	0.000	0.957	0.944	0.932	0.004	0.941	0.946	0.946	0.000
Latent traits (θ)											
Mean error	0.001	0.000	0.000	-0.001	-0.002	0.003	0.010	0.000	0.001	0.000	0.001
RMSE	0.532	0.416	0.851	0.434	0.439	0.548	0.824	0.472	0.476	0.473	0.005
95% CrI coverage	0.934	0.941	0.079	0.938	0.937	0.936	0.002	0.935	0.937	0.939	0.013
Mean reliability (ρ_θ^2)	0.714	0.826	0.810	0.809	0.805	0.695	0.791	0.772	0.767	0.770	0.006
Item scales (a)											
Mean error	0.010	-0.002	0.004	0.009	0.007	-0.004	0.004	-0.006	0.004	0.014	0.008
RMSE	0.281	0.192	0.008	0.249	0.214	0.246	0.001	0.245	0.238	0.227	0.000
95% CI coverage	0.945	0.946	0.000	0.947	0.947	0.943	0.002	0.945	0.945	0.946	0.000
True-estimate correlation	0.975	0.985	0.053	0.991	0.989	0.960	0.279	0.978	0.980	0.982	0.004
Block intercepts (d)											
Mean error	-0.002	-0.001	0.000	-0.002	0.001	-0.004	0.002	-0.005	0.000	0.000	0.003
RMSE	0.122	0.115	0.001	0.127	0.115	0.113	0.003	0.126	0.117	0.112	0.002
95%-CI coverage	0.951	0.948	0.000	0.953	0.948	0.947	0.002	0.951	0.948	0.949	0.001
True-estimate correlation	0.972	0.974	0.002	0.970	0.973	0.977	0.021	0.972	0.972	0.974	0.002

Note. The η_p^2 values are the partial eta-squared effect sizes associated with the main effect of the factor for the corresponding goodness-of-recovery statistic. The values in the other cells are the estimated marginal means of the goodness-of-recovery statistic across all replications for the corresponding factor level. TIRT = Thurstonian item response theory; RMSE = root mean square error; CI = confidence interval; CrI = credibility interval.

Table 3. Structure, Parameter Estimates, Correlations, and Empirical Reliabilities (as Variance of the Latent Trait Estimates) of the 30-Block Forced-Choice Questionnaire.

Block	Dimension/polarity		Item 1 scale		Item 2 scale		Intercept	
			MCMC	TIRT	MCMC	TIRT	MCMC	TIRT
1	OE+	Ag-	1.483	1.641	-1.373	-1.562	1.246	1.411
2	Ag-	ES+	-0.895	-0.757	0.680	0.584	-2.367	-2.310
3	Co-	Ex+	-0.463	-0.070	1.220	1.351	-1.042	-1.101
4	Ex+	ES+	1.630	2.270	1.346	2.017	0.509	0.568
5	OE+	ES-	1.003	0.895	-1.098	-1.028	0.626	0.665
6	OE+	Co+	1.254	1.181	1.189	0.936	0.104	0.075
7	Co+	OE-	1.446	1.516	-0.409	-0.174	1.369	1.457
8	Ex+	OE-	1.191	1.251	-1.076	-1.054	1.541	1.624
9	Ag-	Co+	-0.709	-0.579	0.772	0.888	-1.308	-1.389
10	Co+	ES+	2.405	6.544	2.820	7.506	-0.001	-0.046
11	Co-	Ag+	-1.048	-1.064	0.499	0.242	-0.354	-0.398
12	ES+	Co-	0.841	0.677	-0.836	-0.814	0.975	1.014
13	Ex-	Ag+	-0.579	-0.477	0.453	0.369	0.009	-0.026
14	OE+	Ex-	1.475	1.566	-1.374	-1.448	1.030	1.140
15	Ag+	Co+	1.391	1.471	0.618	0.541	0.305	0.346
16	Ag+	Ex+	0.535	0.097	0.488	-0.002	0.692	0.698
17	OE+	ES+	0.578	0.448	0.768	0.693	1.372	1.336
18	Ex+	OE+	0.964	0.924	1.074	1.065	0.462	0.497
19	Ex+	Ag-	0.913	0.851	-0.990	-0.970	1.707	1.765
20	Ag+	OE+	0.904	0.842	0.819	0.785	0.432	0.456
21	ES-	Ag+	-1.521	-1.294	0.850	0.597	1.028	0.865
22	ES+	OE-	0.696	0.609	-1.119	-1.048	1.960	1.968
23	Ag+	OE-	0.585	0.512	-0.474	-0.310	0.903	0.929
24	ES-	Ex+	-0.892	-0.720	1.020	1.050	-0.529	-0.606
25	OE+	Co-	0.541	0.366	-1.253	-1.091	1.770	1.733
26	Ex-	ES+	-0.515	-0.221	1.216	1.311	-1.120	-1.166
27	ES-	Co+	-0.681	-0.529	0.939	1.082	-0.864	-0.948
28	Co+	Ex-	0.638	0.300	-1.330	-1.408	0.773	0.841
29	Ex+	Co+	2.759	4.001	1.567	2.786	0.364	0.504
30	Ag+	ES+	1.155	1.162	1.354	1.299	0.588	0.592
			ES		Ex		OE	
			ES	.721	.654			
			Ex	.624	.758	.722	.645	
			OE	-.232	-.248	-.179	-.189	.579
			Ag	.363	.355	.521	.528	-.007
			Co	.668	.783	.384	.594	-.076
			Ag		Co			
			Ag	.592	.541			
			Co	.250	.280	.669	.656	

Note. The sign behind the dimension name indicates the item polarity. The values in bold are the empirical reliabilities. MCMC = Markov Chain Monte Carlo; TIRT = Thurstonian item response theory; OE = openness to experience; Ag = agreeableness; ES = emotional stability; Co = conscientiousness; Ex = extraversion.

items, respectively, and over .99 for the intercept). The estimates obtained by both procedures were highly similar, with a few exceptions: the TIRT estimate of the first scale parameter of Block 3 and for both scale parameters of Block 16 were close to zero, whereas the MCMC estimates had higher and more reasonable values. The two scale parameters of Block 10 and the

first one in Block 29 received extremely high estimates (with large associated estimation errors) from the TIRT procedure. The corresponding estimates (and their estimation errors) by the MCMC procedure, however, turned out to be more reasonable, which can be attributed to the prior distributions.

Both procedures yielded very similar results for the latent trait correlations (see bottom part of Table 3). The TIRT estimates generally were more extreme though. In contrast, in the simulation study, the TIRT correlation estimates tended to be more negatively biased. Thus, these results may be reflecting some phenomena not contemplated by the models. The pattern of correlations showed some differences as compared with the correlations among the NEO-PI-R traits in a representative Spanish sample of the general adult population (Costa & McCrae, 2008). The latter study reports positive correlations of openness to experience with extraversion and emotional stability, whereas the authors found negative correlations. They also found substantially higher correlations of extraversion with emotional stability and agreeableness. However, it is unknown whether these differences result from the particular sample of students in their study or are an artifact of the forced-choice response format.

Finally, the empirical reliabilities (taken as the variance of the latent trait estimates) were relatively low, especially for openness to experience and agreeableness. Interestingly, similar to the results of the simulation study, the MCMC yielded higher reliabilities than the TIRT procedure.

Discussion

In this article, the authors have proposed a new variant under the MUPP framework, which differs from Stark et al.'s (2005) original MUPP in two important ways. First, it assumes a dominance rather than an unfolding measurement model for the items. Apart from being more parsimonious, a dominance model may be more appropriate for certain types of items (as argued in the introduction). Second, the Bayesian estimation procedure allows for the item and person parameters to be jointly estimated, which obviates the need for a previous calibration of the items. The simulation study shows good recovery of both the structural and person parameters, even when only three dimensions underlie the FCQ. Note that a low number of latent dimensions generally implies more serious ipsativity issues (Clemans, 1966). Hence, the simulation results most probably generalize—or even turn out more favorable—with more than three dimensions.

An interesting possible extension to the new model (as well as to the original MUPP) consists in handling blocks of more than two items. Although Hontangas et al. (2015; Hontangas et al., 2016) made a theoretical proposal, a detailed exploration of the mathematical properties of their approach as well as the adaptation and testing of the estimation procedure are possible lines for further research.

The authors have discussed the near equivalence between the MUPP-2PL and Brown and Maydeu-Olivares's (2011) TIRT model when applied to paired items. The similarity between both models parallels the relation between Luce's (1959/2005) Choice Axiom and Thurstone's (1927) Case V. Indeed, the underlying assumption in the MUPP framework (see Equation 1) is a formalization of Luce's Choice Axiom (see also Andrich, 1989), whereas the TIRT is based on Thurstone's law of comparative judgment. Moreover, this theoretical equivalence translates empirically (as shown by the simulation study and application to real data). However, although both estimation procedures yield very similar results, one should consider that the MCMC algorithm (a) rates more accurately the estimation errors associated with the latent traits (as the results on the CRLs show), (b) is more precise at recovering the latent space correlational structure, and (c) yields more reasonable estimates when there is little information in the data. Also, for the empirical application, the reliability estimates were higher than with the TIRT. On the contrary, the TIRT procedure, as it relies on software for confirmatory factor analysis (e.g., in

Mplus; Brown & Maydeu-Olivares, 2012), immediately provides statistics to assess global model fit, whereas the Bayesian approach, although being more versatile with respect to model checking and allowing for tests of specific model assumptions (see Gelman, 2014, Chapters 7 and 8), generally requires more efforts from the user to implement the procedures.

Both the authors of the MUPP and the TIRT model discuss two related (although distinct) drawbacks: Stark et al. (2005; Chernyshenko et al., 2009) suggested including unidimensional blocks in the test to identify the latent metric; likewise, Brown and Maydeu-Olivares (2011) concluded, based on a theoretical analysis and simulation results, that opposite-polarity blocks should be included in the FCQ. These recommendations suggest that the quality of the MUPP and TIRT estimation results critically depends on responses given to such blocks; this being the case would cast doubts on the possible strengths of the forced-choice format to control response styles, as such blocks often imply a clearly distinct desirability between the items. The simulation study in this article showed that, for the MUPP-2PL model, the person parameters can be reliably estimated even if the test consists exclusively of bidimensional, direct-item blocks. For the latter to be true, the test should include a sufficient number of items from each latent trait (under the conditions in this simulation study, 24 items per trait yielded a reliability of more than .75). Nevertheless, unidimensional blocks could be included at the questionnaire designer's discretion (taking into account the additional underdetermination affecting the scale parameters).

In the MUPP-2PL, the location parameters of the items (similar to the latent utility means; Brown & Maydeu-Olivares, 2011) are not identified. If an estimate of these location parameters is desired, one may consider a precalibration of the items in a graded-scale format (as in the original MUPP procedure; Stark et al., 2005). However, there may be a risk of introducing biases due to the format. Alternatively, the individual item location parameters can be estimated with the Bayesian algorithm proposed in this article by using a more complex questionnaire design, where the same items are used across several blocks. In one of their ongoing research lines, the authors investigate how the blocks in the questionnaire should be composed to optimize certain aspects of the test (e.g., recovery of item locations, information optimization, or robustness against biases).

The application of a Bayesian joint estimation procedure to the original MUPP model should be quite straightforward (see also Wang, de la Torre, & Drasgow, 2015, which presents an MCMC algorithm to estimate the GGUM parameters). However, the approach would primarily require a prior investigation to find out under what conditions the model is identified. In this regard, it is possible that the underdeterminations and the identification constraints affecting the MUPP model are similar to those found in the MUPP-2PL version.

As a conclusion, the authors' extension of the MUPP framework offers an interesting generalization and applicability to a wider context, which includes dominance items. This extension also allows for a joint estimation of the item and person parameters by means of a Bayesian algorithm. The near equivalence with the TIRT model reveals properties that may also help to find a common framework and allow for model comparison and selection.

Acknowledgment

The authors thank Centro de Computación Científica–Universidad Autónoma de Madrid (CCC-UAM) for allocation of computer time for the simulations.

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: This research is funded by Spanish government's Ministerio de Economía y Competitividad, Projects PSI2012-33343 and PSI2013-44300-P.

Note

1. One may note that, in spite of the close relation between the multi-unidimensional pairwise preference two-parameter logistic (MUPP-2PL) and the Thurstonian item response theory (TIRT) models, results on identifiability cannot be interchanged, given that the Jacobian matrix differs in both models (A. Maydeu-Olivares, personal communication, July 31, 2013).

References

- Andrich, D. (1989). A probabilistic IRT model for unfolding preference data. *Applied Psychological Measurement*, 13, 193-216.
- Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, 69, 49-56.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 374-472). Reading, MA: Addison-Wesley.
- Brown, A., & Maydeu-Olivares, A. (2010). Issues that should not be overlooked in the dominance versus ideal point controversy. *Industrial and Organizational Psychology*, 3, 489-493. doi:10.1111/j.1754-9434.2010.01277.x
- Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, 71, 460-502. doi:10.1177/0013164410375112
- Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods*, 44, 1135-1147. doi:10.3758/s13428-012-0217-x
- Carvalho, L. d. F., de Oliveira, A. Q., Pessotto, F., & Vincenzi, S. L. (2015). Application of the unfolding model to the aggression dimension of the dimensional clinical personality inventory (IDCP). *Revista Colombiana de Psicología*, 23(2), 339-349. doi:10.15446/rcp.v23n2.41428
- Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review*, 51, 292-303. doi:10.1037/h0057299
- Chernyshenko, O. S., Stark, S., Prewett, M. S., Gray, A. A., Stilson, F. R., & Tuttle, M. D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Empirical comparisons with other formats. *Human Performance*, 22, 105-127. doi:10.1080/08959280902743303
- Cho, S., Drasgow, F., & Cao, M. (2015). An investigation of emotional intelligence measures using item response theory. *Psychological Assessment*, 27, 1241-1252. doi:10.1037/pas0000132
- Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance*, 18, 267-307. doi:10.1207/s15327043hup1803_4
- Clemans, W. V. (1966). *An analytical and empirical examination of some properties of ipsative measures* (Psychometric Monographs No. 14). Richmond, VA: Psychometric Society. Retrieved from <http://www.psychometrika.org/journal/online/MN14.pdf>
- Closs, S. J. (1996). On the factoring and interpretation of ipsative data. *Journal of Occupational and Organizational Psychology*, 69, 41-47.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cornwell, J. M., & Dunlap, W. P. (1994). On the questionable soundness of factoring ipsative data: A response to Saville & Willson (1991). *Journal of Occupational and Organizational Psychology*, 67, 89-100.
- Costa, P. T., & McCrae, R. R. (2008). *Inventario de personalidad neo revisado (NEO PI-R): Inventario neo reducido de cinco factores (NEO-FFI): Manual profesional* [Revised NEO personality inventory (NEO-PI-R): Reduced five-factor NEO inventory (NEO-FFI): Professional manual.] (3rd ed. rev. y ampl). Madrid, Spain: Técnicos Editores Asociados.

- De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.
- Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology*, 3, 465-476. doi:10.1111/j.1754-9434.2010.01273.x
- Gelman, A. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.
- Gelman, A. E., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472. doi:10.1214/ss/1177011136
- Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error* (Technical Report No. 15, Office of Naval Research Contract No. 25140, NR-342-02). Stanford, CA: Department of Statistics, Stanford University.
- Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, 74, 167-184. doi:10.1037/h0029780
- Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement*, 39, 598-612. doi:10.1177/0146621615585851
- Hontangas, P. M., Leenen, I., de la Torre, J., Ponsoda, V., Morillo, D., & Abad, F. J. (2016). Traditional scores versus IRT estimates on forced-choice tests based on a dominance model. *Psicothema*, 28, 76-82. doi:10.7334/psicothema2015.204
- Huang, J., & Mead, A. D. (2014). Effect of personality item writing on psychometric properties of ideal-point and Likert scales. *Psychological Assessment*, 26, 1162-1172. doi:10.1037/a0037273
- International Personality Item Pool: A scientific collaboratory for the development of advanced measures of personality and other individual differences. (n.d.). Available from <http://ipip.ori.org/>
- Luce, R. D. (2005). *Individual choice behavior: A theoretical analysis*. Mineola, NY: Dover Publications. (Original work published 1959)
- Oswald, F. L., & Schell, K. L. (2010). Developing and scaling personality measures: Thurstone was right—But so far, Likert was not wrong. *Industrial and Organizational Psychology*, 3, 481-484.
- Patz, R. J., & Junker, B. W. (1999a). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24, 342-366. doi:10.3102/10769986024004342
- Patz, R. J., & Junker, B. W. (1999b). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24, 146-178. doi:10.3102/10769986024002146
- Reckase, M. D., & McKinley, R. L. (1982). Some latent trait theory in a multidimensional latent space. In *Item response theory and computerized adaptive testing conference proceedings*. Retrieved from <http://eric.ed.gov/?id=ED264265>
- 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. doi:10.1177/01466216000241001
- Saville, P., & Willson, E. (1991). The reliability and validity of normative and ipsative approaches in the measurement of personality. *Journal of Occupational Psychology*, 64, 219-238.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement*, 29, 184-203. doi:10.1177/0146621604273988
- Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52, 393-408. doi:10.1007/BF02294363
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34, 273-286. doi:10.1037/h0070288
- Thurstone, L. L. (1947). *Multiple-factor analysis: A development and expansion of the vectors of mind*. Chicago, IL: The University of Chicago Press.
- Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22, 1701-1728.
- Wang, W., de la Torre, J., & Drasgow, F. (2015). MCMC GGUM: A new computer program for estimating unfolding IRT models. *Applied Psychological Measurement*, 39, 160-161. doi:10.1177/0146621614540514